Iranian Distance Education Journal *Open Access*

winter-Spring (2025) 7 (1): 77-87
DOI: 10.30473/IDEJ.2025.70959.1188

ORIGINAL ARTICLE

Early Prediction of Students' Academic Performance Using Interaction Data from Virtual Learning Environments

Seyedeh Mahboobeh Hosseyni[1], Marjan Kaedi[2], Fakhroddin Noorbehbahani[3]

1. M.Sc., Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran
2. Associate Prof., Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran
3. Assistant Prof., Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran

Correspondence:
Marjan Kaedi
Email:
kaedi@eng.ui.ac.ir

Received: 06/April/2024
Accepted: 23/April/2025

How to cite:
Hosseyni, M; Kaedi, M; Noorbehbahani, F (2025)., Early Prediction of Students' Academic Performance Using Interaction Data from Virtual Learning Environments, **Iranian Distance Education Journal**, 7 (1), 77-87.
DOI:10.30473/IDEJ.2025.70959.1188

ABSTRACT

Online learning programs have gained significant popularity in recent years. However, despite their widespread adoption, completion and success rates for online courses are notably lower than those for traditional in-person education. If students' final academic performance could be predicted early by analyzing their behavior within the virtual learning environment, timely alerts could be issued, and targeted interventions could be recommended to prevent underperformance and course abandonment. Previous studies have predicted academic performance using various features, such as demographic data, academic history, in-term exam results, and assignment assessments. However, many online learning platforms do not provide access to such data, rendering these methods ineffective. This study focuses on the early prediction of students' academic performance by extracting novel behavioral features based on their interactions with the online learning platform. To develop robust predictive models, we utilize an integrated approach combining multiple feature selection methods to extract the most informative interaction patterns, followed by application of advanced machine learning algorithms including ensemble learning techniques and artificial neural networks (ANNs). The evaluation results demonstrate that our proposed approach can predict students' final academic performance with an accuracy of 90.62%, using only data collected during the first third of the online course.

KEYWORDS

Early prediction, student performance, e-learning, virtual learning environment, interaction data, machine learning

Extended Abstract

Introduction

E-learning has become a highly convenient, efficient, and effective approach to education. Despite its numerous advantages, however, this mode of learning faces several limitations and challenges when compared to traditional in-person education, which can impact its overall effectiveness. To address these challenges, a variety of studies have been conducted with the aim of gaining deeper insights into students and their learning processes by analyzing data collected from virtual learning environments during the course. One key objective of this data analysis is to predict students' future academic performance. Early prediction of students' final outcomes allows for timely alerts to be issued to both students and relevant stakeholders, enabling effective planning and intervention strategies to prevent student failure.

Previous studies have typically relied on demographic data, academic history, in-term exam results, and assignment grades to predict students' performance early. However, in many virtual and open learning courses, such assessments are either absent or minimal, making these traditional prediction methods ineffective. In contrast, our study proposes a novel approach that predicts students' academic outcomes by analyzing behavioral features derived from their interactions with the virtual learning environment, without depending on midterm scores, assignments scores, or other similar assessments. The structure of the paper is as follows: The related work is first reviewed, followed by the presentation of the proposed method. The evaluation of this method is then provided, and the paper concludes with a summary of the findings.

2. Related Work

In this section, previous studies are classified into two categories. The first part reviews studies based on the characteristics utilized for predicting academic performance, while the second part focuses on studies that specifically address the early prediction of academic outcomes.

2.1. Types of Features in Predicting Academic Performance

In a certain category of studies, students' academic performance has been predicted using demographic information, without considering data derived from their interactions with the virtual learning environment. These studies typically gathered demographic data from students during enrollment via paper or digital forms. For example, a study by Ram et al. (2021) utilized demographic characteristics and academic backgrounds, including gender, marital status, urban or rural residence, type of admission, income, family size, parental qualifications, parental occupation, in-term assessment scores, final exam scores, and the previous year's academic status. These features were collected from 831 samples, and machine learning techniques were employed to predict academic performance in three classes (good, average, and poor). Although studies in this category varies in terms of feature selection and modeling methods, none of these studies utilized valuable data from the virtual learning environment to predict students' final academic performance. These methods often suffer from limited predictive accuracy, and furthermore, in many virtual learning courses—such as free, open learning courses—demographic and background data are not available. In these contexts, the only data that can be leveraged is the student's behavior and interaction with the virtual learning environment.

The subsequent studies highlight studies that leverages valuable data from virtual learning environments to predict academic performance. This approach is particularly useful and effective for improving the outcomes of educational programs that lack demographic data on students. Our proposed method in the present study also falls into this category. One example of research in this category is a study conducted at one of the largest open universities in England, which has 170,000 students in social sciences and engineering disciplines (Brooks, Thompson, and Teasley, 2015), utilized only two features—evaluation results and online interaction reports, which included a daily summary of student clicks. Time series were generated based on daily interaction data

(clickstream) between students and resources. The study found that it was possible to predict course dropout with 90% accuracy using the entire dataset and 84% accuracy using just 5% of the data (Brooks, Thompson, and Teasley, 2015).

Other studies have incorporated additional features. For example, in 2019, a study conducted at a higher educational institution in Kerala, India, aimed to predict students' academic performance using demographic, academic, and behavioral features from virtual learning environments, as well as additional factors such as parents' education levels and students' absenteeism. The study employed four machine learning methods: Support Vector Machine, Naive Bayes, Decision Tree, and Neural Network, alongside the k-means clustering method. The findings demonstrated that behavioral and additional features significantly enhanced prediction accuracy (Francis and Babu, 2019).

In a more recent study by Dang and Nguyen (2022), the focus was on predicting graduation likelihood and student GPA using three primary categories of features: data from a student information system (SIS), a learning management system (LMS), and a video interactions platform, all sourced from the higher educational institution (HEI) in the Sultanate of Oman. The study employed Decision Trees and Multiple Linear Regression to classify students into two categories: In classification, the authors divide students into potential or not. Students who are not in potential class will be labeled in the system. Counselors and lecturers will keep attention to these students. The results help educators and counselors focus their attention on students identified as at risk of failure. The Decision Tree model achieved an accuracy of 47%, while the Multiple Linear Regression model had an accuracy of 52% (Dang and Nguyen, 2022).

In another related study, the focus was on discovering the best machine learning algorithm for the early prediction of students' academic performance. The researchers stated that the goal of their research was to identify the most effective boosting algorithms: AdaBoost, HistGradientBoosting, and Ultimately, they concluded that the standout winner was CatBoost. One of the most impressive achievements of CatBoost was its ability to identify students who may be at risk while reducing false positive predictions. This allows educators to concentrate on areas that require attention without being overwhelmed by unnecessary alerts (Tirumanadham et al., 2024).

This study highlights the critical importance of selecting the right algorithm for classification tasks.

2.2. Early Prediction of Academic Performance

Most studies in this area focus on identifying behavioral patterns and characteristics of students to predict their performance in subsequent courses. Research in this category typically involves modeling a group of students, with the model later applied to similar students in future cohorts (Tsiakmaki et al., 2019). In some studies, a small subset of the data (e.g., 5% of the students) is used to form clusters, and the results are then generalized to the larger student population (He et al., 2015).

On the other hand, another category of studies aims to make early predictions at the beginning of the course using historical data, including demographic and academic information, to estimate students' performance by the end of the course. However, the limitation of this approach lies in the fact that for open-access courses, demographic and historical data about students are often not available (Kovacic, 2010).

In some studies, features related to students' performance during the course have been utilized to predict their success or failure early on. For example, factors such as a specific number of absences, failure to submit assignments, or low scores on midterm assessments have been used as early indicators of potential failure (Luo et al., 2018). However, many learning programs do not include assignments or midterm assessments, leaving students' behaviors and interactions with the e-learning system as the only available data.

To our knowledge, previous studies have not fully explored the use of characteristics related to students' interactions with the system for the early prediction of their academic performance.

Our proposed method aims to address this gap by utilizing behavioral data from students' interactions with the virtual learning environment to predict their success or failure in the course at an early stage.

To highlight the significance of employing machine learning techniques for improving the accuracy of educational outcome predictions, thereby enabling targeted support and resource allocation, one can refer to the study by Kumar (2025). In his research, Kumar evaluates and compares the predictive performance of various machine learning models—namely decision trees, random forests, support vector machines, and neural networks—in forecasting student academic outcomes. The study is based on a dataset from the UCI Machine Learning Repository, which includes student performance data from Portuguese secondary schools, taking into account both academic and demographic factors. The findings reveal that neural networks and random forests demonstrated the highest accuracy rates, achieving 87.4% and 85.6%, respectively. These results underscore the potential of these models for effective educational analytics and the development of early intervention strategies, emphasizing their value in enhancing predictive accuracy within educational contexts (Kumar, 2025).

3. Proposed Method

As previously mentioned, the extraction and selection of relevant features are critical in determining the accuracy of early predictions of students' academic performance. This study focuses on utilizing students' interactions with the virtual learning environment for this purpose.

Our methodology adopts a data-driven approach for early prediction of student academic performance using machine learning techniques. The dataset, sourced from a higher educational institution in Oman, comprises 207 Computer Science students' records across five courses, including academic history, video-watching behaviors, and virtual learning environment (VLE) interactions (58,340 activity logs). We partition the semester into three intervals (first third, two-thirds, and full duration) to identify the earliest viable prediction point. Key steps include: (a) Feature Extraction from VLE logs (e.g., session views, file uploads) for each interval; (b) Data Preprocessing (merging datasets, handling missing values, discretization, and resampling); (c) Feature Selection via four techniques (Forward Selection, Mutual Information, etc.); and (d) Modeling with seven ML algorithms (k-NN, SVM, Random Forest, etc.), optimized via hyperparameter tuning. The goal is to determine the optimal time window and feature set for accurate early prediction.

First, we describe the dataset in detail in the following subsections, then present the four key methodological stages with their respective implementation details. These steps are illustrated in Figure 1

3.1. Data Description

The dataset used in this study is publicly available and sourced from a higher educational institution (HEI) in the Sultanate of Oman contains, the capital of Oman (Hasan et al., 2021). It includes data from Computer Science students enrolled in five courses: Object-Oriented Programming, .Net Programming, E-Commerce Technology, E-Commerce, and Business Technology Management, across ten classes from Spring 2017 to Spring 2021. The dataset contains 207 student records, with the following features: - Academic information, including students' academic history, instances of plagiarism, and the number of times a course was retaken. These features were obtained from the student information system.

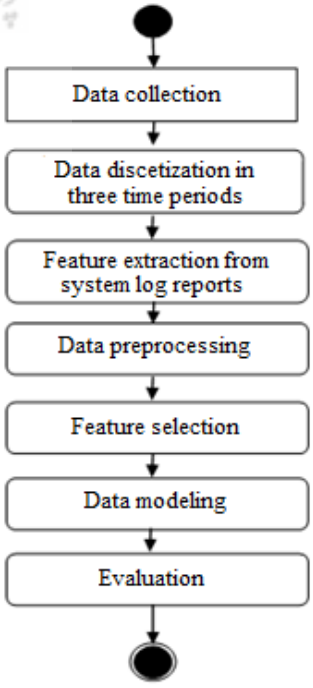

Figure 1- Steps of the proposed method

- Students' interactions with educational video content, including metrics such as the number of times they played, paused, or liked the video, among others. These features were gathered from the video-watching application.

- Student activity within the virtual learning environment, which includes reports on students' usage of the e-learning system. This dataset contains 58,340 records with two key fields: 'username' and 'activity name.' These features were collected from the virtual learning environment.

In this study, the educational period (semester) is divided into three distinct time intervals:

- From the start of the semester until the time of receiving the 'first class grade' (first third of the semester).

- From the start of the semester until the time of receiving the 'second class grade' (two-thirds of the semester).

- The final period, which covers the entire semester, including both class grades, the end-of-semester grade, and the cumulative student GPA ('Grade Point Average').

To perform early prediction of students' success or failure, the entire student dataset is divided into three distinct subsets based on time: data corresponding to the first third of the course, data corresponding to the first two-thirds, and data representing the full course duration. Next, the features of students' activities within the virtual learning environment are extracted for each of these time intervals. The goal is to identify the earliest possible point at which predictions can be made with an acceptable level of accuracy.

This study adopts a data analysis process for this purpose which involves data preprocessing, key feature selection, and modeling with seven machine learning algorithms to determine the optimal timing and most influential predictive factors (Figure 1). All the steps outlined in Figure 1 are applied to each of the three datasets, and the optimal prediction time point and corresponding accuracy are determined. The full details of this data analysis process will be thoroughly explained in the subsequent sections of the paper.

3.2. Feature Extraction

Students engage in a variety of activities while interacting with the virtual learning environment. All their actions are automatically logged in their personal accounts. The log files, which document the activities of students across 10 classes in the virtual learning environment, contain a total of 58,340 records with two key attributes: 'Username' and 'User Activity Type.' These logs provide valuable information that can be utilized to monitor and track students' progress throughout the course. For each time interval, a set of features has been extracted from the total number of records, with the details provided in Table 1.

Features such as 'Number of sessions viewed,' 'Number of files uploaded,' 'Comment creation,' 'Comment viewing,' 'User list views,' and other user behaviors, along with the frequency of each activity, were extracted. These behaviors and activities are expected to be useful and effective in predicting students' final performance at the end of the training period.

Features such as 'number of sessions viewed,' 'number of files uploaded,' 'comment creation,' 'comment viewing,' 'user list access,' and other behaviors, along with the frequency of each activity performed by each student, were extracted. These behaviors and activities are expected to be valuable and influential in predicting the students' final performance at the end of the course.

Table 1- The number of extracted features in each time interval

Time frame	The number of features extracted from students' activity
The first third of the semester	29
The first two-thirds of the semester	33
The entire semester	33

3.3. Data Preprocessing

After extracting features and collecting the dataset, the data must be preprocessed to achieve

the best performance during modeling. The main stages of data preprocessing are displayed in Figure 2.

Figure 2. The stages of data preprocessing in our study

Initially, all feature tables from three different datasets were merged and integrated. This stage involved identifying and removing redundant features that had little to no significant impact on modeling. For instance, features with zero variance or those providing minimal information were eliminated to create a more optimized data space.

Next, various techniques were employed to manage missing values. Missing values were imputed using the mean, median, or mode of the respective features, depending on the type of data and its distribution. This approach helped maintain data integrity and prevent negative impacts on model results.

In the subsequent step, continuous features were converted into discrete features. This was accomplished using binarization techniques, which allowed us to analyze more complex features in a simpler categorical format.

To ensure balance within the dataset, resampling techniques were applied. Specifically, synthetic data was generated to create balance between the two classes—successful and unsuccessful students. This not only helped prevent model bias but also facilitated the training of more accurate models.

Finally, after completing all data cleaning steps, a clean and suitable dataset for building more precise models was created. These preprocessing stages enabled us to confidently analyze and predict students' academic performance.

3.4. Feature Selection

From the sorted features in the balanced dataset, those with the highest correlation to the target variable and the lowest correlation with other features were selected to enhance model performance and accuracy through dimensionality reduction. In this study, four feature selection techniques were employed: 'Forward Selection,' 'Backward Elimination,' 'Mutual Information,' and 'Correlation-based' feature selection.

3.5. Modeling

During the data training phase, seven machine learning methods were employed for modeling the student data, which include:

- K-nearest neighbors
- Support vector machine
- Logistic regression
- Multilayer neural network
- Decision tree
- Random forest
- Adaptive Boosting (AdaBoost)

In this study, the k-Nearest Neighbors (k-NN) method was implemented using 5-fold cross-validation with a neighbor size of k=2, where k represents the number of nearest neighbors considered for classification.

Another method used for training and modeling student performance is the Support Vector Machine (SVM). To optimize the SVM model's performance, its parameters were carefully fine-tuned to align with the specific problem. In particular, the model was configured with a cost parameter set to 1 and a polynomial kernel of degree 3.

The Logistic Regression method was employed for training labeled categorical data. The optimization of this model is highly contingent upon the fine-tuning of its hyperparameters. By selecting the optimal parameters for each dataset and problem, the model minimizes the error between the predicted values and the actual outcomes of the dependent variable.

Another machine learning method explored in our study is the Multilayer Perceptron (MLP) neural network, which can be structured with multiple hidden layers and a variable number of neurons per layer. Key parameters, such as the learning rate, batch size, and maximum number of iterations, were adjusted during implementation. In this study, the MLP model was configured with one hidden layer consisting of 100 neurons and a maximum of 800 iterations. To ensure optimal performance, three critical parameters—alpha (learning rate), batch size,

and maximum iterations—were fine-tuned. After optimizing these parameters, the model was applied to the training dataset, and predictions were generated for the test set.

The Decision Tree method was also implemented, not only for training and prediction but also for extracting relationships between features and deriving rules from significant attributes. In this study, a decision tree with a maximum depth of 9 was constructed using the entropy criterion, which measures the impurity of each node. At this depth, the entropy level reached zero, indicating optimal node purity.

Ensemble learning methods, which combine multiple models to enhance prediction performance, were also employed in this study. One such method is the Random Forest algorithm. For implementation, 50 decision trees were constructed, with the Gini index used as the criterion for measuring node impurity. The Gini index quantifies impurity by summing the squared proportions of samples belonging to each class within a node.

Another ensemble learning technique utilized in this study is Adaptive Boosting (AdaBoost). A decision tree classifier was chosen as the base model for AdaBoost. The base model acts as a weak learner, which is progressively enhanced through iterative boosting.

4. Results of Evaluation

In this study, we predict students' academic performance using four different feature selection methods, as well as a model without feature selection, in combination with seven distinct machine learning algorithms. The predictive performance of each model is evaluated using four standard metrics: accuracy, precision, recall, and F1-Score

To evaluate our proposed method's early prediction capability, we designed three distinct scenarios, each corresponding to performance prediction at a specific time point during the semester.

In the first scenario, we simulated having complete semester-long data for each student to predict end-of-semester performance. The accuracy results, obtained by combining various machine learning algorithms with different feature selection techniques, are presented in Table 2. This table also facilitates a comparison between the performance achieved using the behavioral features extracted from student interactions in the virtual learning environment (as presented in this study) and the performance when these features are excluded. As shown in Table 2, the highest accuracy among the models was achieved by the Decision Tree model utilizing backward feature selection, based on the features extracted in this study related to student activities within the virtual education environment. This model achieved an accuracy of 96.77%. It is worth noting that while this scenario achieved high accuracy, its practical utility remains limited, as predictions made at the semester's end leave insufficient time for meaningful educational interventions.

In the second scenario, predictions were made using data from the first two-thirds of the semester. After 35 iterations of data modeling, with various combinations of seven different modeling techniques and with five distinct categories of features derived from five feature selection methods, the highest accuracy was obtained using the Decision Tree model with forward feature selection, achieving an accuracy of 90.62%. While there was a slight decrease in accuracy compared to the previous dataset, the results still fall within an acceptable range.

In the last scenario, the time window was further reduced, and student success or failure was predicted using only the first one-third of the semester's data. The results of this analysis are presented in Table 3. As shown in Table 3, the highest accuracy was achieved with the K-Nearest Neighbors (KNN) model in combination with the Mutual Information feature selection method, yielding an accuracy of 90.62%. Notably, the accuracy obtained from the first one-third of the semester matches that achieved with the first two-thirds dataset, which is a significant finding. The other evaluation metrics—precision, recall, and F1-Score—obtained during this phase are detailed in Tables 4, 5, and 6, respectively.

Table 2- Accuracy of models with various feature selection methods when predictions are made at the end of semester

The prediction accuracy at the end of semester										
Feature selection methods	**Without feature selection**		**Correlation**		**Backward Selection**		**Forward Selection**		**Mutual Information**	
Machine learning Models	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features
Adaptive Boosting	88.46	89.28	88.57	87.87	87.87	88.57	90	88.88	93.93	91.17
Decision tree	86.36	94.44	89.18	84.61	90	85.18	**96.77**	92	87.87	90.32
K-nearest neighbors	93.75	90.62	93.75	87.09	90.9	86.66	87.5	86.2	89.65	90
Logistic regression	91.17	90.32	88	91.66	91.42	86.95	87.09	88.46	96.15	96.15
Multilayer neural network	91.17	91.17	90.9	90.47	91.66	87.5	88.57	88.57	93.54	90.9
Random forest	91.42	91.17	91.66	88.88	91.66	87.87	91.66	91.42	90.62	91.42
Support vector machine	90.62	89.65	89.65	93.33	93.33	84	90	85.18	92	96.29

Table 3- Accuracy of models with various feature selection methods when predictions are made after the first third of the semester

Accuracy										
Feature selection methods	**Without feature selection**		**Correlation**		**Backward Selection**		**Forward Selection**		**Mutual Information**	
Machine learning Models	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features
Adaptive Boosting	82.75	85.71	86.11	85.71	86.11	85.29	86.48	85.71	85.29	85.29
Decision tree	88.23	87.09	85.29	87.09	85.71	85.29	90	82.14	87.5	85.71
K-nearest neighbors	85.29	86.11	85.29	86.11	86.11	85.29	85.29	86.11	**90.62**	90
Logistic regression	83.87	82.75	83.33	82.75	82.75	84	86.2	84	85.18	84.61
Multilayer neural network	86.11	85.29	84.84	85.29	85.71	85.29	85.71	85.71	86.11	88.57
Random forest	86.48	86.48	86.48	86.48	88.57	86.48	86.11	86.11	86.48	86.48
Support vector machine	85.29	85.29	85.71	85.29	84.37	84.84	87.5	85.29	90.32	87.09

Table 4- Precision of models with various feature selection methods when predictions are made after the first third of the semester

Precision										
Feature selection methods	**Without feature selection**		**Correlation**		**Backward Selection**		**Forward Selection**		**Mutual Information**	
Machine learning Models	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features
Adaptive Boosting	64.86	81.08	83.78	81.08	83.78	78.37	86.48	81.08	78.37	78.37
Decision tree	83.78	75.67	78.37	75.67	81.08	78.37	56.75	62.16	78.37	81.08
K-nearest neighbors	78.37	83.78	78.37	83.78	83.78	78.37	78.37	83.78	83.78	78.37
Logistic regression	70.27	64.86	67.56	64.86	64.86	59.45	70.27	59.45	64.86	62.16
Multilayer neural network	83.78	78.37	75.67	78.37	81.08	78.37	81.08	81.08	883.78	86.64
Random forest	86.48	86.48	86.48	86.48	86.48	86.48	83.78	83.78	86.48	86.48
Support vector machine	78.37	78.37	81.08	78.37	72.97	75.67	78.37	78.37	81.08	75.67

Table 5- Recall of models with various feature selection methods when predictions are made after the first third of the semester

F1-Score										
Feature selection methods	**Without feature selection**		**Correlation**		**Backward Selection**		**Forward Selection**		**Mutual Information**	
Machine learning Models	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features
Adaptive Boosting	78.68	89.55	91.17	89.55	91.17	87.87	92.75	89.55	87.87	87.87
Decision tree	90.9	85.71	87.87	85.71	89.55	87.87	69.23	76.66	87.5	89.55
K-nearest neighbors	87.87	91.17	87.87	91.18	91.17	87.87	87.87	91.17	90.62	87.09
Logistic regression	82.53	78.68	80.64	78.68	78.68	73.68	81.96	83.68	77.96	75.86
Multilayer neural network	91.17	87.87	86.15	87.87	89.55	87.87	89.55	89.55	91.17	92.53
Random forest	92.75	92.75	92.75	92.75	92.53	92.75	91.17	91.17	92.75	92.75
Support vector machine	87.87	87.87	89.55	87.87	84.37	86.15	87.5	87.87	88.88	85.17

Table 6- F1-Score of models with various feature selection methods when predictions are made after the first third of the semester

Recall										
Feature selection methods	**Without feature selection**		**Correlation**		**Backward Selection**		**Forward Selection**		**Mutual Information**	
Machine learning Models	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features	With activity features
Adaptive Boosting	75	93.75	96.87	93.75	96.78	90.62	100	93.75	90.62	90.62
Decision tree	93.75	84.37	90.62	84.37	93.75	90.62	56.25	71.87	87.5	93.75
K-nearest neighbors	90.62	96.87	90.62	96.87	96.78	90.62	90.62	96.87	90.62	84.37
Logistic regression	81.25	75	78.12	75	75	65.62	78.12	65.62	71.87	86.75
Multilayer neural network	96.87	90.62	87.5	90.62	93.75	90.62	93.75	93.75	98.75	96.87
Random forest	100	100	100	100	96.87	100	96.87	96.87	100	100
Support vector machine	90.62	90.62	93.75	90.62	84.37	87.5	87.5	90.62	87.5	84.37

The results of predicting students' final performance, conducted after the first third of the semester, demonstrate that student behaviors and interactions within the virtual learning environment are valuable and contain significant information. This is evident as the final performance of students was predicted with satisfactory accuracy using only data related to user interactions with the virtual learning environment and a single grade od students (without utilizing the two other students' grades, their final exam grade, or the students' GPA). It is important to note that this study did not incorporate demographic data, lifestyle factors, or other additional features in the student modeling.

The results of this study demonstrate a significant improvement over similar studies conducted by Dang and Nguyen on the same dataset (Dang and Nguyen, 2022), with an approximate 43% enhancement in predictive model performance.

5. Conclusion

This study presents a method for the early prediction of students' academic status in e-learning

environments. The approach focuses on

analyzing patterns of student behavior and activities within the virtual education environment, extracting relevant features from their interactions. After applying feature selection techniques and machine learning algorithms, the resulting models achieved an accuracy of 90.62% in predicting students' success or failure using data collected from only the first one-third of the academic period.

In addition to the features used in this study, the timing of each activity—such as the time of day, day of the week, and so on—performed by students may also play a significant role in determining their future status and performance. It is recommended that future studies incorporate the timing of students' activities to predict their success or failure. Furthermore, demographic data (Negaresh et al., 2023) of students could also be utilized to enhance the prediction results. Considering that certain personality traits of students, such as determination, self-confidence, and adaptability, can influence success in a course, a set of inputs that could enhance the prediction accuracy of the method includes the estimated personality traits, emotions, and preferences of the students (SadighZadeh and Kaedi, 2022; PourMohammadBagher et al, 2009, Sadeghian and Kaedi, 2021).

References

Brooks, C., Thompson, C., & Teasley, S. (2015, March). A time series interaction analysis method for building predictive models of learners using log data. In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 126-135).

Dang, T. K., & Nguyen, H. H. X. (2022). A Hybrid Approach Using Decision Tree and Multiple Linear Regression for Predicting Students' Performance Based on Learning Progress and Behavior. *SN Computer Science*, *3*(5), 393.

Francis, B. K., & Babu, S. S. (2019). Predicting academic performance of students using a hybrid data mining approach. *Journal of medical systems*, *43*(6), 162.

Tirumanadham, N. K. M. K., Thaiyalnayaki, S., & SriRam, M. (2024, January). Evaluating Boosting Algorithms for Academic Performance Prediction in E-Learning Environments. In *2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)* (pp. 1-8). IEEE.

Kumar, P. (2025). Evaluating Machine Learning Algorithms for Enhanced Prediction of Student Academic Performance.

Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., & Sarker, K. U. (2021). Dataset of students' performance using student information system, moodle and the mobile application "eDify". *Data*, *6*(11), 110.

He, J., Bailey, J., Rubinstein, B., & Zhang, R. (2015, February). Identifying at-risk students in massive open online courses. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 29, No. 1).

Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data. *In Informing Science + Information Technology Education Joint Conference, Cassino, Italy*.

Lu, O. H., Huang, A. Y., Huang, J. C., Lin, A. J., Ogata, H., & Yang, S. J. (2018). Applying learning analytics for the early prediction of Students' academic performance in blended learning. *Journal of Educational Technology & Society*, *21*(2), 220-232.

Negaresh, F., Kaedi, M., & Zojaji, Z. (2023). Gender identification of mobile phone users based on Internet usage pattern, *International Journal of Engineering*, 36 (2), 335-347.

PourMohammadBagher, L., Kaedi, M., Ghasem-Aghaee, N., & Ören, T. I. (2009). Anger evaluation for fuzzy agents with dynamic personality. *Mathematical and Computer Modelling of Dynamical Systems*, 15(6), 535–553.

Ram, M. S., Srija, V., Bhargav, V., Madhavi, A., & Kumar, G. S. (2021, September). Machine learning based student academic performance prediction. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 683-688). IEEE.

Sadeghian, A., Kaedi, M. (2021). Happiness recognition from smartphone usage data considering users' estimated personality traits, *Pervasive and Mobile Computing*, 73 ,101389.

SadighZadeh, S. and Kaedi, M. (2022). Modeling user preferences in online stores based on user mouse behavior on page elements, *Journal of Systems and Information Technology*, 24(2), 112-130.

Tsiakmaki, M., Kostopoulos, G., Koutsonikos, G., Pierrakeas, C., Kotsiantis, S., & Ragos, O. (2018, July). Predicting university students' grades based on previous academic achievements. In *2018 9th international conference on information, Intelligence, Systems and Applications (IISA)* (pp. 1-6). IEEE.