mathematics

MDPI

Article

Comparative Analysis of Machine Learning Models for Predicting Student Success in Online Programming Courses: A Study Based on LMS Data and External Factors

Felipe Emiliano Arévalo-Cordovilla [1,*] and Marta Peña [2]

[1] Faculty of Science and Engineering, Universidad Estatal de Milagro, Milagro 091706, Ecuador
[2] Department of Mathematics, Universitat Politècnica de Catalunya—BarcelonaTech, 08019 Barcelona, Spain; marta.penya@upc.edu
* Correspondence: farevaloc@unemi.edu.ec

Abstract: Early prediction of student performance in online programming courses is essential for implementing timely interventions to enhance academic outcomes. This study aimed to predict academic success by comparing four machine learning models: Logistic Regression, Random Forest, Support Vector Machine (SVM), and Neural Network (Multilayer Perceptron, MLP). We analyzed data from the Moodle Learning Management System (LMS) and external factors of 591 students enrolled in online object-oriented programming courses at the Universidad Estatal de Milagro (UNEMI) between 2022 and 2023. The data were preprocessed to address class imbalance using the synthetic minority oversampling technique (SMOTE), and relevant features were selected based on Random Forest importance rankings. The models were trained and optimized using Grid Search with cross-validation. Logistic Regression achieved the highest Area Under the Receiver Operating Characteristic Curve (AUC-ROC) on the test set (0.9354), indicating strong generalization capability. SVM and Neural Network models performed adequately but were slightly outperformed by the simpler models. These findings suggest that integrating LMS data with external factors enhances early prediction of student success. Logistic Regression is a practical and interpretable tool for educational institutions to identify at-risk students, and to implement personalized interventions.

Keywords: academic performance prediction; educational data mining; machine learning models; student retention

MSC: 97U10; 62J12

check for updates

Citation: Arévalo-Cordovilla, F.E.; Peña, M. Comparative Analysis of Machine Learning Models for Predicting Student Success in Online Programming Courses: A Study Based on LMS Data and External Factors. *Mathematics* **2024**, *12*, 3272. https://doi.org/10.3390/math12203272

Academic Editors: Victor S. Sheng, Shih-Wei Lin and Wei Fang

Received: 28 August 2024
Revised: 12 October 2024
Accepted: 17 October 2024
Published: 18 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The ability to predict early academic performance is crucial in the field of higher education, particularly in online object-oriented programming courses. Early predictions help identify students at risk of failure and enable the implementation of personalized pedagogical interventions, thereby improving educational outcomes. Research on predicting academic performance in online courses has addressed this issue using various methodological and theoretical approaches, machine learning techniques, and data analysis from Learning Management Systems (LMS) [1,2].

Machine learning techniques have gained significant relevance in recent years because of their ability to analyze large volumes of data and identify patterns that can predict academic success or failure, thus facilitating early educational interventions [3,4]. Various studies have demonstrated the effectiveness of these methodologies in different educational contexts using both academic and non-academic data to improve the accuracy of predictions [5,6].

Interaction data in Virtual Learning Environments (VLE) have proven to be a valuable source of information for predicting academic performance [1,2]. Students' actions in

LMS, such as viewing courses and resources, submitting and evaluating assignments, and interacting in quizzes and forums, can be used to predict their course performance with reasonable accuracy through the application of machine learning algorithms [1,7]. Logistic regression models were employed to determine the influence of the different variables on student performance. For instance, ref. [1] used a logistic regression model, dividing the main response variable (Grade) into two categories, Fail and Not Fail, to analyze the impact of the "Event Context" on students who fail.

However, most studies have focused on collecting variables and predicting student performance at the end of a course. While these results are useful for identifying significant variables that influence student performance, they often do not provide solutions to prevent dropouts and failures [3,8]. Online learning platforms generate vast amounts of data associated with student interactions from the beginning of a course. Ref. [3] suggested that a comprehensive predictive model could be developed by analyzing variable data from the start of the course, which would be effective at preventing dropouts and failures, and would allow instructors to conduct effective interventions at the optimal time.

Additionally, ref. [4] examined a wide range of factors that influence academic performance and dropout rates in higher education. Using Educational Data Mining (EDM) methods and Structural Equation Modeling (SEM), they identified intrinsic motivation, self-regulated learning strategies, and family background as critical factors affecting academic performance. They also concluded that earning the first 20 credits was crucial for predicting dropout rates. These findings highlight the importance of including demographic and psychological variables in predictive models, supporting our hypothesis that external factors combined with Moodle's data can improve prediction accuracy.

The prediction of learning outcomes in blended courses was investigated by [5] using a machine-learning algorithm based on online learning behavior data. They found that student engagement in various activities was crucial for improving prediction accuracy. In a systematic review, ref. [6] identified that the most effective algorithms for predicting university graduation include Support Vector Machines (SVM), Random Forests (RF), and Logistic Regression (LR), highlighting the importance of integrating new data sources and addressing ethical considerations. This theoretical framework supports our methodology, which combines Moodle's data with external factors.

Furthermore, ref. [7] employed a probabilistic logistic regression model to identify at-risk students at different stages of the academic calendar, underscoring the importance of quantifying uncertainty in model predictions for more reliable identification of at-risk students. Similarly, ref. [8] conducted a comprehensive study on predicting academic performance in programming courses using data mining techniques. Using data extracted from Moodle, they implemented various classification algorithms, such as Logistic Regression and Random Forests, to predict students' final grades and found that these models were particularly effective.

The early prediction of student performance in higher education has also been explored through clustering and classification techniques. Ref. [9] highlighted the usefulness of factors such as admission grades and first-level courses in predicting students' GPA. Ref. [10] focused on identifying university students at risk of dropping out using predictive analytics and found that the Random Forest algorithm was the most effective. Additionally, ref. [11] conducted a systematic literature review on the use of EDM and Learning Analytics (LA) techniques to improve student retention.

Complementing these findings, ref. [12] explored the estimation of academic performance in distance education using multiple data-mining methods. Their study demonstrated that Deep Learning algorithms, Random Forest, and SVM achieved the highest accuracy, with success rates exceeding 96%. Ref. [13] applied logistic regression models to predict student failure in exact science disciplines, showing that factors such as age and previous academic history are significant predictors of student failure. The developed models achieved an accuracy of over 70%, highlighting the effectiveness of logistic regression in identifying at-risk students.

Ref. [14] demonstrated that integrating explainable machine learning with automated prescriptive analytics can significantly improve academic performance. Using a combination of algorithms, such as Logistic Regression, k-nearest Neighbors, Random Forest, Naïve Bayes, and CatBoost, their study achieved notable accuracy, especially with the CatBoost algorithm, which achieved an accuracy of 75%.

Another significant study [15] focused on predicting the academic performance of master's students in Germany. Using a dataset that included demographic and academic characteristics, as well as data mining algorithms such as Random Forest and Artificial Neural Networks, they demonstrated that it is possible to accurately predict academic performance. The Random Forest classifier showed superior performance, with an accuracy ranging from 77% to 94%. The study also highlighted the importance of attributes such as semester grades and the distance from housing to university, which were critical for predictions.

Ref. [16] addressed the portability of predictive models of academic performance and found that the predictive accuracy of the models decreased as the similarity between the new group and the base group decreased. Ref. [17] explored how information on submitted assignments can improve the prediction of student performance in online courses, indicating that using Multiple Instance Learning (MIL) can increase predictive accuracy by more than 20%, highlighting the relevance of assignments as a predictive factor.

The prediction of academic performance using assessment grades and online activity data within an LMS was the focus of research in [18]. Their work demonstrated that assessment grades are the most significant predictors of academic performance and that models combining these data with online activity data offer superior predictive performance. Ref. [19] explored the use of clickstream data in a virtual learning environment to predict at-risk students. Using a Long Short-Term Memory (LSTM) model, they achieved significant improvements in prediction accuracy compared to traditional models, such as logistic regression and Artificial Neural Networks (ANN). Ref. [20] investigated the effectiveness of machine learning models, specifically Random Forest and Logistic Regression, in predicting student outcomes in introductory physics courses. They found that integrating institutional and class data provided the most accurate predictions of early student success, allowing for the timely identification of at-risk students to implement appropriate educational interventions.

In summary, although machine-learning techniques have shown great potential for predicting academic performance in online learning environments, significant challenges related to the complexity and accessibility of these models persist. This study aims to address these gaps by developing and comparing multiple machine learning models—logistic regression, random forest, SVM, and neural networks—based on Moodle data and external factors. By providing a practical and efficient solution for the early prediction of academic performance in online object-oriented programming courses, this approach not only promises to offer an accessible tool for educational institutions, but also enhances the understanding of how various factors contribute to academic performance. It is expected that this methodology will allow educators at the Universidad Estatal de Milagro (UNEMI) to identify students who require additional support in a timely manner, thereby improving their academic performance and retention, as well as the quality of virtual education at the institution.

2. Materials and Methods

The study was conducted at the Universidad Estatal de Milagro (UNEMI) with the objective of predicting the early academic performance of second-semester students in the Information Technologies Engineering program who were enrolled in the online Object-Oriented Programming course during the academic periods of 2022 and 2023. A quantitative, descriptive, and correlational approach was designed involving the integration and analysis of academic, demographic, and online interaction data. Machine learning models were applied to identify at-risk students and propose timely interventions.

2.1. Data Collection

2.1.1. Academic and Demographic Data

Academic and demographic information were collected from 591 students through the university's administrative system. The collected data included the following.

- First Partial Grades: Grade 1, Grade 2, and Exam 1, each with a specific weight of 50 points.
- Demographic Characteristics: Current age, sex, nationality, ethnicity, presence of any disability, province and canton of birth, and province and canton of residence.

Inclusion Criteria:

- Students enrolled in an Object-Oriented Programming course.
- Complete availability of academic and interaction data in Moodle.

Exclusion Criteria:

- Students with incomplete or significantly missing data.
- Students who dropped the course before the first midterm.

2.1.2. Moodle Interaction Data

Detailed interaction records from the Moodle platform were extracted as follows:

- Course Logins: Number of times students logged into the course.
- Reviewed Resources: Number of times students accessed the study materials.
- Assignment Submissions: Quantity and punctuality of submissions.
- Evaluation of Participation: Completion and review of quizzes.
- Grade Access and Review: Accessing and reviewing grades.

All data were anonymized to ensure student privacy in accordance with current ethical and legal standards.

2.2. Definition of the Dependent Variable

The dependent variable, early academic performance, was defined as the sum of the first partial grades (Grades 1, 2, and Exam 1). Because the academic period consists of two partials worth 50 points each, the maximum grade is 100 points. A threshold of 35 out of 50 points in the first partial was established to classify the students:

- At Risk (Class 0): Students with grades below 35.
- Not at Risk (Class 1): Students with grades of 35 points or higher.

The 35 points threshold was selected considering that if a student achieved the same score in the second partial, they would reach the minimum required grade of 70 points to pass the course.

2.3. Data Preprocessing

2.3.1. Cleaning and Preparation

- Data Type Conversion: Categorical variables were transformed into the 'category' type and numerical variables into 'float' or 'int' as appropriate.
- Categorical variables encoding One-Hot Encoding were used for nominal categorical variables.

2.3.2. Feature Selection

Feature selection was performed using the Random Forest algorithm to identify the most influential variables for early academic performance. The selected features were as follows:

- Academic Grades: Grade 1, Grade 2, Exam 1.
- Moodle Interactions: Course logins, submitted assignments, completed quizzes, reviewed resources.
- Demographic Data: Current age, sex, and presence of disability.

2.4. Machine Learning Models

Four supervised models were implemented for the prediction:

1. Logistic Regression: A Linear model used for binary classification.
2. Random Forest Classifier: An ensemble of decision trees that enhances accuracy and reduces overfitting.
3. Support Vector Machine (SVM): An algorithm that seeks the hyperplane that best separates classes.
4. Artificial Neural Network (MLP): A model capable of capturing complex nonlinear relationships in the data.

2.5. Experimental Procedure

2.5.1. Dataset Division

The dataset was divided into the following categories.

- Training Set: 70% of the data were used to train the models.
- Test Set: The remaining 30% was used to evaluate the performance of the models.

A balanced class distribution was ensured in both partitions using stratified sampling.

2.5.2. Handling Class Imbalance

The SMOTE (synthetic minority oversampling technique) was applied to generate synthetic examples of the minority class and balance the training set.

2.5.3. Normalization and Scaling

The numerical variables were standardized using Z-Score Standardization to ensure that all features contributed equally to the model.

2.5.4. Hyperparameter Optimization

Grid Search with 5-fold stratified cross-validation was used to find the best combination of hyperparameters for each model:

- Logistic Regression: The regularization parameter was adjusted *C*.
- Random Forest: Various tree depths, estimators, and splitting criteria were explored.
- SVM: Different kernels (linear and RBF) and parameter values were tested *C*.
- Neural Network: The number of neurons, activation functions, learning rate, and number of epochs were adjusted.

2.5.5. Training and Validation

Each model was trained using the optimized training set and validated using the following metrics.

- Area Under the ROC Curve (AUC-ROC).
- Precision, Recall, and Specificity.
- Confusion Matrix.

2.6. Model Evaluation

2.6.1. Performance Metrics

The AUC-ROC was selected as the primary metric because of its ability to evaluate binary classification performance without relying on a specific threshold. Additionally, the following were analyzed:

- ROC Curves: To Visualize the balance between true positive and false positive rates.
- Classification reports: Precision, recall, and F1-score.

2.6.2. Interpretation of Results

The models were compared based on their predictive capabilities, and the most influential features were analyzed. The feature importance provided by the Random Forest

model and the coefficients from Logistic Regression were used to understand the impact of each variable.

2.7. Tools and Technologies Used

- Programming Language: Python 3.11 (Python Software Foundation, Beaverton, OR, USA).
- Libraries.
- Pandas 2.2.3 (NumFOCUS, Austin, TX, USA) and NumPy 1.26.4 (NumFOCUS, Austin, TX, USA): data manipulation and processing.
- Scikit-learn 1.5.2 (Scikit-learn Developers, licencia BSD): Implementing machine-learning models and preprocessing.
- Imbalanced-learn 0.12.3 (Imbalanced-learn Developers, Europe): Handling class imbalance with SMOTE.
- TensorFlow 2.17.0 (Google LLC, Mountain View, CA, USA) and Keras 3.5.0 (Google LLC, Mountain View, CA, USA) (via SciKeras): Build and train the neural network.
- Matplotlib 3.9.2 (Matplotlib Development Team) and Seaborn 0.13.2 (Michael L. Waskom): For data and result visualization.

2.8. Ethical Considerations

Confidence and anonymity of student data were ensured at all stages of the study. The study complied with institutional and legal regulations regarding the use of personal and academic data.

3. Results

In this section, we present the results obtained from the analysis conducted on the four machine learning models to predict the early academic performance of students in UNEMI's online Object-Oriented Programming courses.

3.1. Descriptive Analysis

Table 1 summarizes the descriptive statistics of the features selected for this study. These features include note1, note2, exam1, and course_accesses, which were identified as relevant for predicting early academic performance. The table includes key statistics, such as the mean, standard deviation, minimum and maximum values, and percentiles (25%, 50%, and 75%) for each variable.

Table 1. Descriptive statistics of selected features.

Feature	Count	Mean	Std Dev	Min	25th Percentile	50th Percentile (Median)	75th Percentile	Max
note1	591	12.55	2.46	0	11.83	13.17	14.17	15.00
note2	591	11.28	3.67	0	10.00	12.00	14.00	15.00
exam1	591	16.18	4.29	0	14.00	18.00	19.50	20.00
course_accesses	591	119.34	93.65	0	59.50	99.00	151.00	1203.00
grades_reviewed	591	4.49	9.81	0	0.00	1.00	5.00	114.00
quizzes_completed	591	2.62	0.63	0	2.00	3.00	3.00	4.00
quizzes_reviewed	591	54.82	25.89	0	37.50	50.00	69.00	214.00
resources_reviewed	591	29.25	26.39	0	8.00	24.00	43.00	276.00
updated_assignments_submitted	591	1.19	2.14	0	0.00	0.00	2.00	18.00
assignments_reviewed	591	39.53	32.76	0	18.00	32.00	53.00	301.00
current_age	591	28.02	7.66	17	22.00	27.00	33.00	60.00

The mean provides the average value of each feature, helping to understand the overall behavior of the data, whereas the standard deviation indicates the dispersion or variability of the values relative to the mean. The minimum and maximum values show the

range of the data, whereas the percentiles help understand the distribution of the values and detect the presence of outliers or extreme values.

In Table 1, it can be observed that some variables, such as course_accesses and grades_reviewed, showed high variability, indicating significant differences in system usage among students. For example, course_accesses has a mean of 119.34 but a maximum of 1203, suggesting highly active students.

The variables note1 and note2 had a reduced range, with means of 12.5 and 11.3%, respectively, indicating similar scores among students. The maximum value of assignments_reviewed, 301, also suggests considerable dispersion in student engagement with course activities. These analyses help understand student behavior and provide key information for improving academic performance predictions.

3.2. Linear Regression Model Using the Ordinary Least Squares (OLS)

Table 2 presents the results of the multiple linear regression using the Ordinary Least Squares (OLS) method to predict the variable final_note. This table provides key metrics such as R-squared, F-statistic, and several diagnostic statistics that assess the quality of the model fit, as well as the presence of issues such as autocorrelation and normality of residuals. The R-squared and Adjusted R-squared values indicate that the model has strong explanatory power regarding the variability in the final score. The table includes tests for normality and multicollinearity, which are crucial for evaluating the validity of model assumptions.

Table 2. OLS regression results.

Dep. Variable	final_note	R-squared	0.794
Model	OLS	Adj. R-squared	0.790
Method	Least Squares	F-statistic	202.9
Prob (F-statistic)	1.96×10^{-190}	Log-Likelihood	−2027.3
No. Observations	591	AIC	4079
Df Residuals	579	BIC	4131
Df Model	11	Covariance Type	Non-robust
Omnibus	100.335	Durbin–Watson	1.800
Prob (Omnibus)	0.000	Jarque–Bera (JB)	198.003
Skew	−0.961	Prob (JB)	1.01×10^{-43}
Kurtosis	5.084	Cond. No.	1.27×10^{3}

The following metrics provide insights into the quality of the OLS regression model used to predict final_note. Below is a detailed breakdown of these metrics:

R-squared and Adjusted R-squared:

- R-squared (R^2 = 0.794): This value indicates the proportion of variability in the dependent variable (final_note) is explained by the model. A value of 0.794 indicates that approximately 79.4% of the variation in final_note is explained by the predictor variables (note1, note2, exam1, etc.). This suggests that the model has a good explanatory power.
- Adjusted R-squared (0.790): The Adjusted R-squared is similar to R^2, but adjusts for the number of predictors in the model. This adjustment helps prevent R^2 from simply increasing with more predictors without truly improving the model. In this case, the value of 0.790 was very close to R^2, indicating that the number of variables was appropriate for explaining the outcome without overfitting.

F-statistic and Prob (F-statistic):

- F-statistic (202.9) and Prob (F-statistic) (1.96×10^{-190}): The F-statistic measures the overall quality of the model fit by comparing it with a model without predictors (only the mean). A high F-statistic value and a very low *p*-value (1.96×10^{-190}) indicate that the model with predictor variables is significantly better than that without them, suggesting that the model has a good overall fit.

Diagnostic Statistics:

- Omnibus, Prob (Omnibus), Jarque–Bera (JB), Prob (JB): These tests evaluate whether the model residuals are normally distributed. A low *p*-value (0.000) in both cases suggests that the residuals do not follow a normal distribution, which may indicate that the model does not fit well in all the cases.
- Skew (−0.961) and kurtosis (5.084): A negative skew indicates a distribution with a longer tail to the left. A Kurtosis greater than 3 indicates a distribution that is more "peaked" than normal.
- Durbin–Watson (1.800): This value is relatively close to 2, which indicates no strong evidence of autocorrelation in the residuals.

3.3. Model Optimization and Hyperparameter Selection

Four machine-learning models were implemented and tuned to predict students' early academic performance. Hyperparameter optimization was performed using a Grid Search with five-fold cross-validation, obtaining the following optimal parameters for each model, as detailed in Table 3.

Table 3. Optimal hyperparameters for each model after cross-validation.

Model	Optimum Hyperparameters
Logistic Regression	C = 0.1
Random Forest	max_depth: None, min_samples_split: 2, n_estimators: 200
SVM	C: 10, kernel: 'rbf'
Artificial Neural Network (MLP)	activation: 'tanh', batch_size: 32, dropout_rate: 0.0, epochs: 50, neurons: 32

3.4. Cross-Validation

The performance of the models was evaluated using the Area Under the ROC Curve (AUC-ROC) during cross validation. Table 4 presents the results.

Table 4. AUC-ROC scores obtained in five-fold cross-validation.

Model	Cross-Validation AUC Scores	Mean AUC
Logistic Regression	[0.9458, 0.9603, 0.9588, 0.9705, 0.9569]	0.9584
Random Forest	[0.9911, 0.9899, 0.9892, 0.9861, 0.9921]	0.9897
SVM	[0.9802, 0.9928, 0.9747, 0.9792, 0.9880]	0.9830
Artificial Neural Network (MLP)	[0.9901, 0.9932, 0.9922, 0.9871, 0.9918]	0.9909

The results indicated that the Neural Network (MLP) achieved the highest mean AUC-ROC during cross-validation, closely followed by Random Forest.

3.5. Evaluation on the Test Set

The trained models were then evaluated using an independent test set. The key performance metrics, including precision, recall, f1-score, and AUC-ROC, are summarized in Table 5.

Logistic Regression achieved the highest AUC-ROC value on the test set, indicating an excellent balance between sensitivity and specificity. The Random Forest had the highest overall accuracy.

Logistic Regression and Random Forest stood out in terms of different metrics, suggesting that both models are effective for this type of prediction. The choice between them may depend on their preference for interpretability (Logistic Regression) or the ability to capture complex relationships (Random Forest).

Table 5. Model performance on the test set.

Model	Accuracy	Precision	Recall	f1-Score	AUC-ROC
Logistic Regression	87%	90%	87%	88%	0.9354
Random Forest	89%	89%	89%	89%	0.9103
SVM	85%	88%	85%	86%	0.8558
Artificial Neural Network (MLP)	86%	89%	86%	87%	0.9016

3.6. Classification Reports

Detailed classification reports for each model are presented in Table 6.

Table 6. Comparative classification metrics for each model.

Model	Class	Precision	Recall	f1-Score
Logistic Regression	Class 0 (At Risk)	57%	83%	68%
	Class 1 (Not at Risk)	96%	88%	92%
Random Forest	Class 0 (At Risk)	66%	66%	66%
	Class 1 (Not at Risk)	93%	93%	93%
SVM	Class 0 (At Risk)	54%	76%	63%
	Class 1 (Not at Risk)	95%	87%	91%
Artificial Neural Network (MLP)	Class 0 (At Risk)	55%	79%	65%
	Class 1 (Not at Risk)	96%	87%	91%

Key Observations:

- Logistic Regression has high precision for Class 1 (96%) but moderate precision for Class 0 (57%), indicating that it is very effective at identifying students not at risk, but less precise in identifying those at risk.
- Random Forest shows balanced precision and recall for both classes, especially excelling in Class 1 with 93% across all metrics.
- SVM has high precision for Class 1 (95%), similar to Logistic Regression, but lower precision for Class 0 (54%) compared to Random Forest.
- The neural Network (MLP) maintains high precision for Class 1 (96%) and reasonable performance for Class 0 (55% precision), similar to the SVM.

3.7. ROC Curves

The ROC curves of the models are illustrated in Figure 1, showing the relationship between the true positive rate and false positive rate for different classification thresholds.

Performance Analysis:

- Logistic Regression: Although model was initially used alone, its performance on the test set was outstanding, especially in detecting at-risk students (high sensitivity). Its high AUC-ROC value (0.9354) indicates a strong discriminative ability.
- Random Forest: This model showed the highest overall accuracy (89%) and a good balance between precision and recall in both classes. Its ability to handle non-linear relationships and capture complex interactions between variables may have contributed to this performance.
- SVM: Although it had an acceptable performance, its precision and AUC-ROC were lower than those of the other models. This suggests that, for this dataset, SVMs with an RBF kernel did not sufficiently capture the present complexities.
- Neural Network (MLP): Despite its excellent performance in cross-validation, its performance on the test set was slightly lower, indicating possible overfitting. However, it maintained a good overall predictive ability.

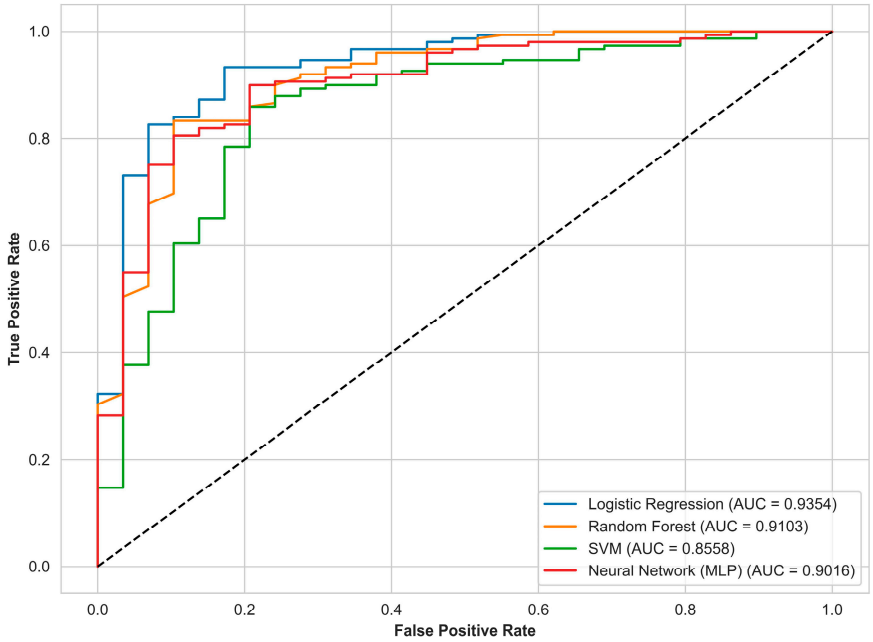

Figure 1. Comparative ROC Curves of the Four Models.

3.8. Confusion Matrices

The confusion matrices of the models are shown in Figure 2. These matrices allow for the visualization of each model's performance in terms of true positives, true negatives, false positives, and false negatives.

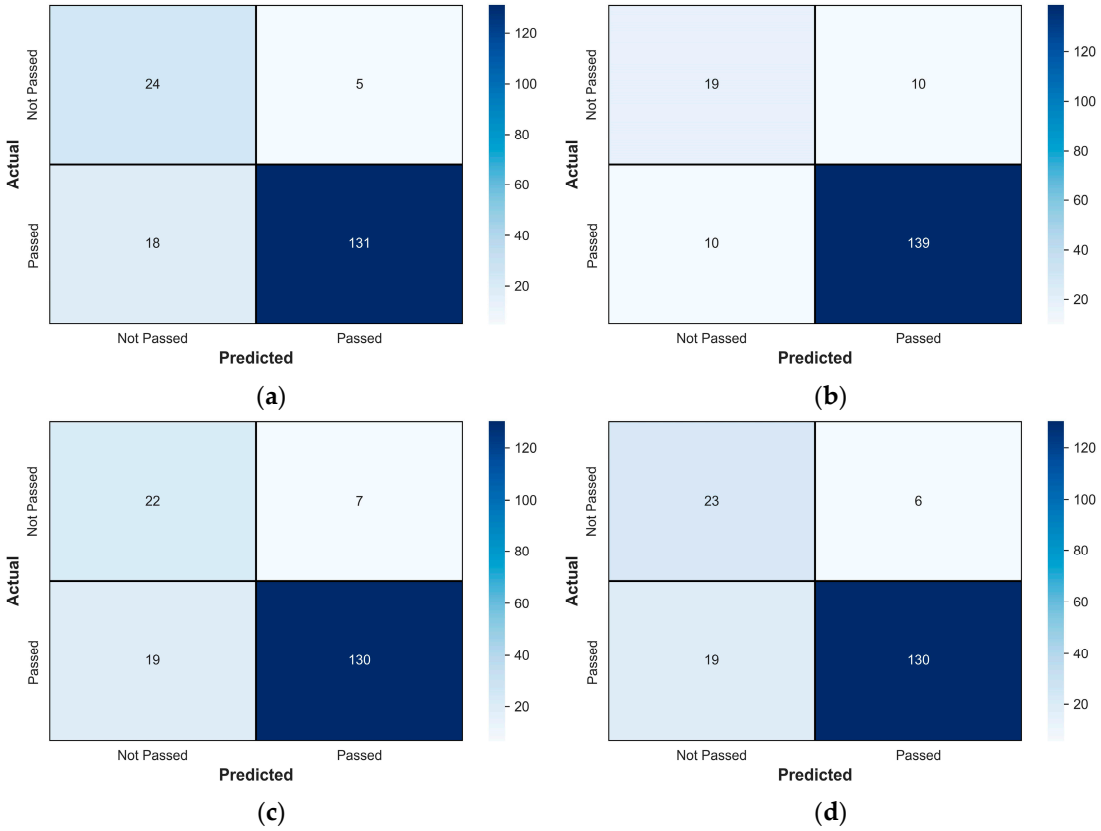

Figure 2. Confusion matrices for each model on the test set. (**a**) Logistic Regression; (**b**) Random Forest; (**c**) SVM; (**d**) Neural Network (MLP).

3.9. Feature Importance

The feature importance was analyzed using a Random Forest model, as shown in Figure 3.

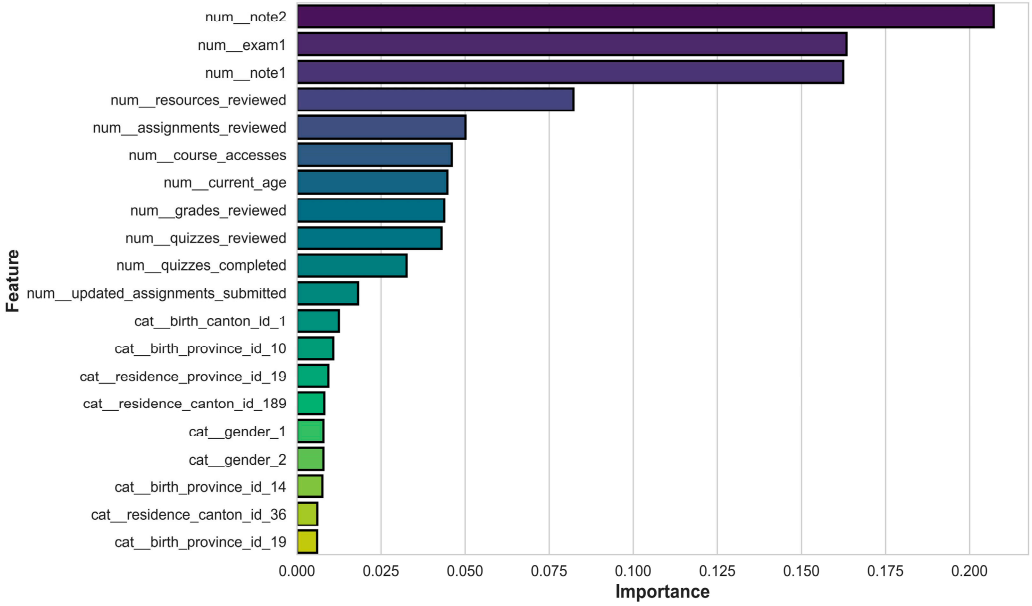

Figure 3. Feature importance based on Random Forest.

The five most influential features are as follows.

- note2: 20.72%
- exam1: 16.34%
- note1: 16.24%
- resources_reviewed: 8.22%
- assignments_reviewed: 5.01%

Interpretation of Key Variables

- First Partial Grades: The variables note1, note2, and exam1 were consistently the most important, highlighting the relevance of early academic performance in predicting final success. This is consistent with academic logic, in which initial grades reflect students' understanding of and adaptation to the course.
- Interactions in Moodle: Variables related to online activities, such as resources_reviewed and assignments_reviewed, also had a significant influence. This finding indicates that greater engagement with online resources and assessments is associated with better academic outcomes.
- Demographic Variables: Features such as age, gender, ethnicity, disability, and nationality had less influence on the prediction, suggesting that, while they are relevant, their impact is less pronounced compared to grades and participation in Moodle.

4. Discussion

This study aimed to predict early academic performance in online object-oriented programming courses by comparing four machine learning models: Logistic Regression, Random Forest, Support Vector Machine (SVM), and Neural Network (Multilayer Perceptron, MLP). By integrating Moodle's LMS interaction data with academic and demographic factors, we sought to identify the most effective model for early prediction, enabling timely interventions for at-risk students. The results demonstrated that the Logistic Regression achieved the highest AUC-ROC score on the test set, indicating a strong generalization capability. This suggests that Logistic Regression is a practical and interpretable tool for ed-

ucational institutions to implement early warning systems, facilitating their understanding and adoption by educators who may not have advanced technical expertise.

The strong performance of Logistic Regression aligns with previous studies that emphasize its effectiveness in educational settings. For instance, refs. [1,13] successfully employed logistic regression to predict student performance, highlighting its applicability in early intervention. The simplicity and interpretability of the model make it a valuable asset for institutions that aim to implement predictive analytics without extensive computational resources.

Additionally, studies by [21,22] provide valuable insights into the performance of regression estimation methods, specifically Maximum Likelihood (ML) and Ordinary Least Squares (OLS). Ref. [21] found that ML offers more accurate estimates and higher efficiency in cases involving polynomial regression with small sample sizes, suggesting its advantage in reducing bias and improving model fit. By contrast, ref. [22] concluded that OLS is more effective in recovering weak common factors, especially under conditions of model error or small sample sizes. These findings highlight that, while ML may be advantageous for accurate parameter estimation and dealing with outliers, OLS shows consistency in situations involving weak influencing factors or specific types of errors. In the context of predicting early academic performance, these considerations emphasize the need to carefully select an appropriate estimation method based on the specific characteristics of the data.

While the Random Forest model achieved the highest overall accuracy, it was slightly less generalizable than Logistic Regression based on the AUC-ROC metric. Random Forest's ability to handle complex interactions between variables likely contributed to its strong performance, which is consistent with findings by [23], who reported high accuracy using Random Forest in predicting student grades. The use of SMOTE to address class imbalance in our study mirrors the approach in [23], underscoring the importance of handling imbalanced datasets to enhance model performance.

The Neural Network (MLP) and SVM models performed adequately but were outperformed by simpler models. This contrasts with some studies in which complex models, such as Artificial Neural Networks (ANNs), showed superior performance [24]. The potential overfitting observed in our Neural Network model suggests that, with the available data, simpler models may generalize better. Ref. [6] emphasized the importance of model interpretability and the careful integration of advanced techniques, suggesting that the choice of model should consider both performance and practical applicability.

Feature importance analysis revealed that early academic grades, specifically grades 1 and 2, and exam 1, were the most significant predictors of the final performance. This underscores the critical role of initial assessments in forecasting student success, corroborating the findings of [18], which identified assessment grades as key predictors. Additionally, Moodle interaction features, such as reviewed resources and assignments, significantly contributed to the predictions. This indicates that higher engagement with online resources correlates with better outcomes, aligning with [2], which demonstrated the predictive value of LMS log data.

Demographic variables had a lesser impact on the predictions. While factors such as age, gender, and disability status are relevant, their influence is minimal compared to academic performance and engagement metrics. This observation suggests that, in the context of online programming courses, student behavior and performance are more critical indicators, a conclusion also reached by [4]. The minimal impact of demographic factors may be due to the specific nature of programming courses, where engagement with course materials plays a more significant role in determining success.

The practical implications of these findings are significant for educational institutions. Implementing Logistic Regression models can help identify at-risk students early in the course, allowing for timely support and resources to improve academic outcomes. Understanding that engagement metrics are significant predictors enables institutions to encourage behaviors that enhance learning, such as regular access to course materials and

the timely submission of assignments. The simplicity of Logistic Regression facilitates its adoption and interpretation by educators, making it a practical choice for institutions that lack extensive technical resources.

Despite these positive outcomes, this study had limitations that should be acknowledged. The data were collected from a single institution and focused on a specific course, which may have limited the applicability of the findings to other contexts. Ref. [16] noted that the predictive accuracy can decrease when models are applied to different groups, indicating the need for caution when generalizing these results. Additionally, while our model included academic and engagement data, incorporating other factors, such as psychological assessments, socioeconomic status, or reading behaviors, could enhance predictive accuracy. Ref. [25] demonstrated that e-book reading behaviors are effective predictors of academic performance, suggesting that integrating such data could improve model robustness.

Future research should explore the use of ensemble models to determine whether combining multiple algorithms improves the predictive performance in educational settings. Ref. [26] proposed an ensemble model that showed improved accuracy over individual models, indicating the potential benefits of educational predictions. In addition, incorporating Big Data technologies and diverse data sources, as discussed in [27], may enhance model scalability and accuracy, allowing the processing of larger datasets and integrating more complex features.

Conducting research across multiple institutions and courses would help assess the generalizability of the models and refine them for broader application. Such cross-institutional studies could identify common predictors of academic success and more effectively tailor interventions. Moreover, investigating the integration of behavioral data, as suggested by [25], and advanced data processing techniques, as employed by [27], could further improve the predictive capabilities of the models.

In conclusion, this study contributes to the growing body of literature demonstrating the effectiveness of machine learning models in predicting early academic performance. By comparing multiple models and integrating various data sources, we provide insights that can help educational institutions implement practical and interpretable tools for early interventions. The findings support the notion that while complex models and ensemble methods have their place, simpler models such as Logistic Regression offer substantial benefits in terms of interpretability and ease of implementation, especially when early academic indicators and engagement metrics are considered. Future studies should aim to validate these findings in diverse educational contexts and explore the integration of additional predictive factors to enhance model performance.

5. Conclusions

This study demonstrates the effectiveness of integrating Learning Management System (LMS) interaction data with academic and demographic factors in predicting early academic performance in online object-oriented programming courses. By comparing four machine learning models—Logistic Regression, Random Forest, Support Vector Machine (SVM), and Neural Network (Multilayer Perceptron, MLP)—we found that Logistic Regression achieved the highest Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for the test set, indicating its superior generalization capability and practical applicability.

The prominence of Logistic Regression in our findings underscores its value as a practical and interpretable tool for educational institutions seeking to implement early warning systems. The simplicity and effectiveness of the model facilitate the timely identification of at-risk students, allowing the deployment of personalized interventions that can enhance academic outcomes and retention rates.

Our analysis revealed that early academic assessment (grades 1 and 2, and exam 1) was the most significant predictor of student success, highlighting the critical importance of initial performance indicators. Additionally, engagement metrics from the LMS, such as the number of resources reviewed and assignments submitted, were influential factors that emphasized the role of active participation in online learning environments.

Although Random Forest also demonstrated strong predictive performance, its complexity and reduced interpretability compared with Logistic Regression may pose challenges for practical implementation in educational settings. Although adequate, the neural Network and SVM models did not outperform the simpler models, suggesting that increased model complexity does not necessarily translate into better predictive accuracy.

This study contributes to the existing body of knowledge by providing empirical evidence that supports the integration of LMS data with external factors for early prediction of academic performance. It offers a scalable and efficient approach that can be readily adopted by educational institutions to enhance student support mechanisms.

However, this study had some limitations that warrant consideration. The data were collected from a single institution and focused on a specific course, which may have limited the generalizability of the findings. Future research should explore the application of this methodology to diverse educational contexts and disciplines. Additionally, incorporating a broader range of external factors such as psychological or socioeconomic variables, may further enhance the predictive capabilities of the models.

In conclusion, the findings of this study affirm that simple, interpretable machine learning models such as Logistic Regression can effectively predict early academic performance when integrating LMS interaction data with academic and demographic factors. Implementing such predictive models can empower educators to proactively support students, ultimately improving educational outcomes in online learning environments.

Author Contributions: Methodology, F.E.A.-C.; software, F.E.A.-C.; formal analysis, F.E.A.-C.; investigation, F.E.A.-C. and M.P.; supervision, M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the Universidad Estatal de Milagro (UNEMI) Scholarship.

Data Availability Statement: The datasets presented in this article are not readily available because they must first be approved by the ethics committee. Access to the data will be granted only upon completion of the necessary procedures and receipt of approval from the ethics committee.

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

References

1. Gaftandzhieva, S.; Talukder, A.; Gohain, N.; Hussain, S.; Theodorou, P.; Salal, Y.K.; Doneva, R. Exploring Online Activities to Predict the Final Grade of Student. *Mathematics* **2022**, *10*, 3758. [CrossRef]
2. Riestra-González, M.; del Puerto Paule-Ruíz, M.; Ortin, F. Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Comput. Educ.* **2021**, *163*, 104108. [CrossRef]
3. Adnan, M.; Habib, A.; Ashraf, J.; Mussadiq, S.; Raza, A.A.; Abid, M.; Bashir, M.; Khan, S.U. Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. *IEEE Access* **2021**, *9*, 7519–7539. [CrossRef]
4. Kocsis, A.; Molnár, G. Factors influencing academic performance and dropout rates in higher education. *Oxf. Rev. Educ.* **2024**, 1–19. [CrossRef]
5. Luo, Y.; Han, X.; Zhang, C. Prediction of learning outcomes with a machine learning algorithm based on online learning behavior data in blended courses. *Asia Pac. Educ. Rev.* **2022**, *25*, 267–285. [CrossRef]
6. Pelima, L.R.; Sukmana, Y.; Rosmansyah, Y. Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review. *IEEE Access* **2024**, *12*, 23451–23465. [CrossRef]
7. Nimy, E.; Mosia, M.; Chibaya, C. Identifying At-Risk Students for Early Intervention—A Probabilistic Machine Learning Approach. *Appl. Sci.* **2023**, *13*, 3869. [CrossRef]
8. Peraic, I.; Grubisic, A. Predicting Academic Performance of Students in a Computer Programming Course using Data Mining. *Int. J. Eng. Educ.* **2023**, *39*, 836–844.
9. Alhazmi, E.; Sheneamer, A. Early Predicting of Students Performance in Higher Education. *IEEE Access* **2023**, *11*, 27579–27589. [CrossRef]
10. Gonzalez-Nucamendi, A.; Noguez, J.; Neri, L.; Robledo-Rella, V.; García-Castelán, R.M.G. Predictive analytics study to determine undergraduate students at risk of dropout. *Front. Educ.* **2023**, *8*, 1244686. [CrossRef]
11. Shafiq, D.A.; Marjani, M.; Habeeb, R.A.A.; Asirvatham, D. Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review. *IEEE Access* **2022**, *10*, 72480–72503. [CrossRef]
12. Bütüner, R.; Calp, M.H. Estimation of the Academic Performance of Students in Distance Education Using Data Mining Methods. *Int. J. Assess. Tools Educ.* **2022**, *9*, 410–429. [CrossRef]

13. Costa, S.F.; Diniz, M.M. Application of logistic regression to predict the failure of students in subjects of a mathematics undergraduate course. *Educ. Inf. Technol.* **2022**, *27*, 12381–12397. [CrossRef]
14. Ramaswami, G.; Susnjak, T.; Mathrani, A. Supporting Students' Academic Performance Using Explainable Machine Learning with Automated Prescriptive Analytics. *Big Data Cogn. Comput.* **2022**, *6*, 105. [CrossRef]
15. Alturki, S.; Cohausz, L.; Stuckenschmidt, H. Predicting Master's students' academic performance: An empirical study in Germany. *Smart Learn. Environ.* **2022**, *9*, 38. [CrossRef]
16. Arroyo-Barrigüete, J.L.; Carabias-López, S.; Curto-González, T.; Hernández, A. Portability of Predictive Academic Performance Models: An Empirical Sensitivity Analysis. *Mathematics* **2021**, *9*, 870. [CrossRef]
17. Esteban, A.; Romero, C.; Zafra, A. Assignments as Influential Factor to Improve the Prediction of Student Performance in Online Courses. *Appl. Sci.* **2021**, *11*, 10145. [CrossRef]
18. Alhassan, A.; Zafar, B.; Mueen, A. Predict Students' Academic Performance based on their Assessment Grades and Online Activity Data. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 185–194. [CrossRef]
19. Aljohani, N.R.; Fayoumi, A.; Hassan, S.-U. Predicting At-Risk Students Using Clickstream Data in the Virtual Learning Environment. *Sustainability* **2019**, *11*, 7238. [CrossRef]
20. Zabriskie, C.; Yang, J.; DeVore, S.; Stewart, J. Using machine learning to predict physics course outcomes. *Phys. Rev. Phys. Educ. Res.* **2019**, *15*, 020120. [CrossRef]
21. Peter, A.K.; David, A.A. Application of the Maximum Likelihood Approach to Estimation of Polynomial Regression Model. *Int. J. Math. Comput. Res.* **2022**, *10*, 2693–2700. [CrossRef]
22. Briggs, N.E.; MacCallum, R.C. Recovery of Weak Common Factors by Maximum Likelihood and Ordinary Least Squares Estimation. *Multivar. Behav. Res.* **2003**, *38*, 25–56. [CrossRef] [PubMed]
23. Bujang, S.D.A.; Selamat, A.; Ibrahim, R.; Krejcar, O.; Herrera-Viedma, E.; Fujita, H.; Ghani, N.A.M. Multiclass Prediction Model for Student Grade Prediction Using Machine Learning. *IEEE Access* **2021**, *9*, 95608–95621. [CrossRef]
24. Sandra, L.; Lumbangaol, F.; Matsuo, T. Machine Learning Algorithm to Predict Student's Performance: A Systematic Literature Review. *TEM J.* **2021**, *10*, 1919–1927. [CrossRef]
25. Chen, C.-H.; Yang, S.J.H.; Weng, J.-X.; Ogata, H.; Su, C.-Y. Predicting at-risk university students based on their e-book reading behaviours by using machine learning classifiers. *Australas. J. Educ. Technol.* **2021**, *37*, 130–144. [CrossRef]
26. Yan, L.; Liu, Y. An Ensemble Prediction Model for Potential Student Recommendation Using Machine Learning. *Symmetry* **2020**, *12*, 728. [CrossRef]
27. Ouatik, F.; Erritali, M.; Ouatik, F.; Jourhmane, M. Predicting Student Success Using Big Data and Machine Learning Algorithms. *Int. J. Emerg. Technol. Learn. (iJET)* **2022**, *17*, 236–251. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.