Learning Analytics for Student Performance and Early Detection of At-Risk

Rhowel M. Dellosa
College of Computing Sciences, Pangasinan State University, Lingayen, Pangasinan, Philippines
Centre for Innovation and Technology Adoption, UNITAR International University, 47301 Petaling Jaya, Selangor, Malaysia

***Abstract*—The rapid growth of online learning platforms has generated large volumes of student interaction data that may support learning analytics and early academic intervention. This study proposed an intelligent learning analytics system for predicting student performance and identifying at-risk students using online learning behavior data. The Online Learning Behavior Dataset was used, consisting of demographic information, learning environment variables, and behavioral indicators. Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Gradient Boosting (XGBoost) models were implemented and evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. SVM achieved the highest accuracy of 0.40, followed by Random Forest at 0.38, XGBoost at 0.35, and ANN at 0.32. However, because the task involved three risk categories, the best accuracy was only modestly above the approximate chance level of 0.33. These results indicate that the current models should be interpreted as exploratory decision-support tools rather than deployment-ready classifiers. The small performance differences among models matter for deployment because a marginal improvement may not justify automated risk classification unless supported by stronger validation, better feature engineering, and statistically meaningful performance gains. The study, therefore, demonstrates the potential of machine learning for exploratory learning analytics, while also emphasizing the need for verified institutional datasets and more rigorous evaluation before practical implementation.**

Keywords*—*At-Risk student detection; learning analytics; machine learning; online learning behavior; student performance prediction

I. Introduction

The rapid advancement of information and communication technologies has significantly transformed the landscape of education, particularly through the widespread adoption of online learning platforms and learning management systems (LMS). These platforms generate large volumes of educational data derived from students' online learning activities, such as login frequency, time spent on course materials, participation in discussion forums, assignment submissions, and quiz attempts.

This data provides valuable insights into students' learning behaviors and engagement patterns, which can be utilized to improve teaching strategies and enhance learning outcomes. The emerging fields of educational data mining (EDM) and learning analytics aim to analyze such data to understand and optimize learning processes and environments [1], [2].

In recent years, machine learning and artificial intelligence techniques have increasingly been applied in education to predict students' academic performance and identify learners who may be at risk of failure or dropout. Early prediction of student performance allows instructors and institutions to implement timely interventions, provide personalized support, and improve overall student success rates [3], [4], [5]. Several studies have demonstrated the effectiveness of machine learning algorithms, such as support vector machines, neural networks, decision trees, and ensemble learning models, in analyzing educational datasets and predicting academic outcomes [10], [6], [7], [8].

Furthermore, the use of deep learning approaches has shown promising results in capturing complex patterns within student learning behaviors and engagement data. These models can analyze large-scale educational datasets and extract meaningful features that contribute to more accurate predictions of student performance [9], [10], [11]. Other researchers have explored hybrid models and ensemble techniques to enhance prediction accuracy and robustness by combining multiple algorithms [12], [13], [14]. Such approaches have been applied across different educational levels, including secondary education, higher education, and online learning environments [15], [16], [17].

In addition to predicting academic performance, learning analytics has been widely used to identify students who are at risk of academic failure or dropout. Early detection of at-risk students enables educators to implement preventive strategies and targeted interventions before academic difficulties become severe [18], [19], [20]. Studies have also emphasized the importance of analyzing students' behavioral data from LMS logs, such as navigation patterns, engagement frequency, and interaction with course materials, as these factors are closely associated with academic success and learning outcomes [2], [21].

Despite the growing body of research in educational data mining and learning analytics, many educational institutions still struggle to effectively utilize the vast amount of data generated by online learning platforms. Although LMS systems collect detailed behavioral data, these data are often underutilized due to the absence of intelligent analytical tools capable of transforming raw data into actionable insights for instructors and administrators [22], [23]. As a result, educators frequently identify struggling students only after poor academic performance has already occurred, limiting opportunities for early intervention and support.

Moreover, traditional monitoring approaches typically rely on manual observation or basic statistical analysis, which may not capture complex behavioral patterns that influence student learning outcomes. Advanced machine learning and deep

learning models have demonstrated the potential to improve predictive accuracy and provide more reliable insights into student engagement and performance [24], [25], [26]. Integrating such techniques into intelligent learning analytics systems can significantly enhance educational decision-making processes and support data-driven strategies for improving student success.

A. Problem Statement

With the rapid growth of online learning platforms, large amounts of student interaction data are generated, including login frequency, time spent on lessons, quiz attempts, assignment submissions, and participation in discussions. While this data contains valuable insights about students' learning behaviors, many educational institutions do not effectively utilize it to monitor student engagement, predict academic performance, or identify students at risk of failure or dropout.

As a result, instructors often become aware of struggling students only after poor academic outcomes have already occurred. The lack of predictive systems limits the ability of educators to provide early interventions and personalized learning support.

B. Research Objectives and Evaluation Goals

- To determine whether online learning behavior indicators can predict student performance using supervised machine learning models.
- To compare the predictive performance of Random Forest, SVM, ANN, and XGBoost using accuracy, precision, recall, F1-score, and ROC-AUC.
- To identify which behavioral and contextual variables contribute most strongly to student performance prediction.
- To evaluate whether the proposed learning analytics approach can provide reliable early indicators of at-risk students.

II. Methodology

A. Research Design

This study adopted a quantitative predictive analytics research design using machine learning techniques to analyze online learning behavior data, predict student academic performance, and identify students at risk of academic failure. The study followed an Educational Data Mining (EDM) and Learning Analytics approach, as both fields focus on extracting meaningful patterns from educational datasets to support instructional decision-making and improve learning outcomes. A supervised machine learning approach was employed because the dataset contained defined outcome variables for course completion and risk classification. Random Forest and XGBoost were selected as ensemble methods capable of modeling nonlinear feature interactions, while Support Vector Machine (SVM) was included due to its effectiveness in moderate-sized classification tasks. Artificial Neural Network (ANN) was also used to examine whether neural network models could capture complex behavioral patterns in online learning data. An 80:20 train-test split was applied to provide sufficient data for model training while retaining a separate testing subset for model evaluation.

B. Data Source

The dataset used in this study was obtained from Kaggle and contained records of students' online learning behavior, including learning hours, quiz attempts, assignment submissions, and course completion rates. It consisted of 1,000 student records from online learning environments and included 15 attributes related to student demographics, learning behavior, and course outcomes. The demographic attributes included age, gender, country, education level, and field of study, while the learning environment attributes included platform used, device used, and learning mode. The dataset also contained behavioral indicators, such as daily learning hours, quizzes attempted, and assignments submitted, as well as performance-related variables, including course completion rate and satisfaction score. These attributes represented students' interaction patterns with online learning platforms and served as the basis for analyzing learning behavior and predicting academic performance [27].

Although the dataset provided relevant demographic and behavioral variables, it had several limitations. Since the dataset was obtained from Kaggle, it did not include detailed institutional validation, sampling procedures, or verified collection methodology. Therefore, the 1,000 records may not fully represent authentic learning management system activity from formal educational institutions. Accordingly, the findings of this study should be interpreted as exploratory rather than generalizable. Future studies should validate the proposed system using larger institutionally sourced datasets with verified student records and actual learning management system logs.

Table I presents the variables included in the Online Learning Behavior Dataset used in this study. The dataset contains demographic information, learning behavior indicators, and performance metrics that are used as input features for predicting student performance and detecting at-risk students.

C. Data Preprocessing

Data preprocessing was performed to improve the quality and usability of the dataset before conducting machine learning analysis. The dataset used in this study consists of 1,000 student records collected from online learning environments worldwide and includes attributes related to student demographics, learning platform usage, behavioral learning indicators, and performance outcomes. These attributes include demographic variables such as age, gender, country, education level, and field of study; platform-related features such as platform used, device used, and learning mode; behavioral indicators including daily learning hours, quizzes attempted, and assignments submitted; and performance indicators such as course completion rate and satisfaction score. Preprocessing was necessary to ensure that the dataset was clean, structured, and suitable for training predictive models.

TABLE I. VARIABLES USED IN THE STUDY

Field Name	Type	Format / Description
Country	Categorical (String)	Country of the student participating in the online learning platform
Age	Numeric (Integer)	Age of the student in years
Gender	Categorical (String)	Gender of the student (Male / Female)
Education_Level	Categorical (String)	Educational attainment of the student (e.g., High School, Undergraduate, Graduate)
Field_of_Study	Categorical (String)	Academic specialization or program of the student
Platform_Used	Categorical (String)	Online learning platform used (e.g., Coursera, edX, Udemy)
Device_Used	Categorical (String)	Device used to access the learning platform (Laptop, Tablet, Smartphone)
Learning_Mode	Categorical (String)	Mode of learning (Self-paced or Instructor-led)
Enrollment_Date	Date	Date when the student enrolled in the course (YYYY-MM-DD format)
Daily_Learning_Hours	Numeric (Float)	Average number of hours spent studying per day
Quizzes_Attempted	Numeric (Integer)	Number of quizzes attempted by the student
Assignments_Submitted	Numeric (Integer)	Number of assignments submitted
Course_Completion_Rate (%)	Numeric (Float / Percentage)	Percentage of course completion achieved by the student
Satisfaction_Score (1–5)	Numeric (Integer)	Student satisfaction rating on a scale of 1 to 5

The preprocessing stage involved data cleaning, data transformation, and feature scaling. During data cleaning, duplicate records were removed to avoid redundancy in the dataset. Missing values were checked and addressed to maintain data completeness, and the dataset was reviewed to verify data consistency across all attributes. These steps ensured that the data used for analysis was accurate and reliable.

Following data cleaning, data transformation was conducted to convert categorical variables into numerical formats that can be processed by machine learning algorithms. Encoding techniques were applied to variables such as gender, country, platform used, device used, and learning mode. Label encoding was applied to the gender variable, while one-hot encoding was used for country, platform used, and device used to represent categorical variables as numerical vectors. Binary encoding was applied to the learning mode variable.

Feature scaling was applied to normalize numerical variables and improve the performance and stability of machine learning models. Variables such as daily learning hours, quizzes attempted, and assignments submitted were normalized using Min-Max scaling, which placed values within a consistent range. This process helped prevent variables with larger numerical values from dominating model training and supported balanced learning across input features.

D. Feature Selection and Target Definition

Feature selection was performed to identify variables relevant to student performance prediction and at-risk classification. The procedure separated predictor variables from outcome variables to avoid circularity and improve the validity of model interpretation. The selected input features represented student background characteristics, engagement levels, and platform interaction patterns, including age, daily learning hours, quizzes attempted, assignments submitted, platform used, device used, and learning mode.

This study addressed two prediction tasks. The first task focused on student performance prediction, where the course completion rate was used as the target variable. Course completion rate measured the percentage of the course completed by each student and serves as an indicator of academic progress in the online learning environment. The machine learning models analyzed the relationship between the selected input features and course completion rate to predict student performance outcomes.

The second task focused on at-risk student classification. To reduce circular dependency, the course completion rate was used only as the outcome-based criterion for assigning risk categories and was excluded from the predictor set during at-risk classification. Students with completion rates of 70% and above were labeled as low risk, students with completion rates from 50% to 69% were labeled as moderate risk, and students with completion rates below 50% were labeled as high risk. Engagement-related variables, such as assignments submitted and quizzes attempted, were retained as predictors rather than being used as direct label-construction rules.

A rule-based labeling procedure was applied before model training to assign the at-risk category for each student record. The following pseudocode illustrates the classification logic used to generate the risk labels while ensuring that the course completion rate was not included as an input feature during model training.

Algorithm 1: At Risk-Student Classification

```
Initialize
Load the dataset.
Define the assignmentthreshold value.
Create an empty variable risklabel for each student record.
Compute
Extract the required features
completionrate and assignmentsubmitted.
While (student records remain in the dataset) do
    For (every student record) do
        Update
```

```
            Retrieve completionrate and assignmentsubmitted
            Analyze the student's learning performance
            If (completionrate < 50) or (assignmentsubmitted <
assignmentthreshold) then
                Search / Classify
                Set risklabel = 1 (Student is At-Risk).
            Else
                Set risklabel = 0 (Student is Not At-Risk).
            End
        End
    End
```

In Algorithm 1, each student record was assigned a risk category based on course completion rate, and the resulting labeled dataset was then used to train classification models using only non-outcome predictors. This separation between label construction and model inputs strengthened the validity of the at-risk detection experiment. Through this feature selection and labeling process, the study focused the machine learning models on relevant behavioral and contextual predictors while reducing the circular dependency that could otherwise inflate feature importance and compromise the validity of at-risk detection results.

E. Machine Learning Algorithms

To evaluate student performance prediction and at-risk detection, this study applied several supervised machine learning algorithms. These algorithms were selected to compare different modeling strategies, including tree-based ensembles, margin-based classification, neural networks, and gradient boosting.

The machine learning models were trained using selected input features such as age, daily learning hours, quizzes attempted, assignments submitted, platform used, device used, and learning mode. For the at-risk classification task, Course Completion Rate was excluded from the input variables because it was used to define the risk label. This separation was necessary to avoid circular prediction and to provide a more valid evaluation of the models.

F. Model Training Configuration

- Random Forest - number of trees, maximum depth, splitting criterion, class weighting, and random state.
- SVM - kernel type, C value, gamma, scaling procedure, and class weighting.
- ANN - hidden layers, neurons, activation, optimizer, learning rate, batch size, epochs, and early stopping.
- XGBoost - number of estimators, learning rate, maximum depth, subsampling rate, evaluation metric, and random state. If no tuning was performed, state that default library parameters were used.

Table II summarizes the machine learning algorithms used in the study, their learning approach, prediction tasks, and expected outputs. All four algorithms use supervised learning, meaning they require labeled data to learn the relationship between student characteristics, learning behaviors, and performance outcomes.

TABLE II. MACHINE LEARNING ALGORITHMS USED

Algorithm	Type of Learning	Prediction Task	Output
Random Forest	Supervised Learning	Classification / Regression	Predict student performance and detect at-risk students
Support Vector Machine (SVM)	Supervised Learning	Classification	Classify students as at-risk or not at-risk
Artificial Neural Network (ANN)	Supervised Learning	Regression / Classification	Predict academic performance based on behavioral features
Gradient Boosting (XGBoost)	Supervised Learning	Classification / Regression	Improve prediction accuracy for student performance and risk detection

Random Forest and XGBoost are suitable for both classification and regression tasks. In this study, they can be used to predict student performance and detect at-risk students because they are capable of modeling complex relationships among learning behavior variables. SVM is mainly used for classification, making it appropriate for categorizing students as at-risk or not at-risk. ANN can support both regression and classification, allowing it to predict academic performance from behavioral features and classify student risk levels.

TABLE III. INPUT FEATURES AND VARIABLES

Category	Variables	Description
Demographic Features	Age, Gender, Country, Education_Level, Field_of_Study	Student background information
Learning Environment Features	Platform_Used, Device_Used, Learning_Mode	Technology and learning environment used by the student
Behavioral Learning Indicators	Daily_Learning_Hours, Quizzes_Attempted, Assignments_Submitted	Measures of student engagement in the learning platform
Target Variable (Performance)	Course_Completion_Rate (%)	Indicates academic progress and performance
Target Variable (Risk Classification)	At-Risk Label	Classification of students as at-risk or not at-risk

Table III presents the main categories of variables used in the study. The variables are grouped into demographic features, learning environment features, behavioral learning indicators, and target variables. Demographic features, such as age, gender, country, education level, and field of study, describe the students' background characteristics. Learning environment features, including the platform used, the device used, and learning mode, describe the technological and instructional conditions under which students accessed online learning. Behavioral learning indicators, such as daily learning hours, quizzes attempted, and assignments submitted, represent students' level of engagement in the online learning platform. These variables are important because they provide measurable evidence of student participation and learning activity. The target variables define the outcomes predicted by the machine learning models. Course completion rate represents student academic progress and performance, while the at-risk label classifies students according to their likelihood of academic difficulty.

TABLE IV. STUDENT RISK CLASSIFICATION CRITERIA

Condition	Classification
Course Completion Rate ≥ 70%	Low Risk
Course Completion Rate between 50%–69%	Moderate Risk
Course Completion Rate < 50%	High Risk

Table IV shows the rule-based criteria used to assign students to risk categories based on their course completion rate. Students who completed 70% or more of the course were classified as Low Risk, indicating satisfactory academic progress. Students with a completion rate between 50% and 69% were classified as Moderate Risk, suggesting that they may need monitoring or additional academic support. Students with a completion rate below 50% were classified as High Risk, indicating poor progress and a greater need for early intervention. Overall, this classification helps identify students who may require timely support in an online learning environment.

G. Model Training and Testing

The dataset consisting of 1,000 student records was split using an 80:20 train-test approach, where 80% of the data was allocated for training and 20% was reserved for testing. The training set was used to build the Random Forest, SVM, ANN, and XGBoost models, while the testing set was used to evaluate their performance on unseen records.

During the training phase, each algorithm learned relationships between the selected input features and the target outputs. For at-risk classification, only non-outcome predictors were used to prevent leakage from Course Completion Rate into the risk label. The trained models were then applied to the testing dataset, and their predictions were compared with the actual outcomes using classification metrics.

H. Model Evaluation Metrics

Table V outlines the evaluation metrics used to assess the performance of the machine learning models, including accuracy, precision, recall, and F1 score, which measure the effectiveness and reliability of the predictive models.

TABLE V. MODEL EVALUATION METRICS

Metric	Description	Formula
Accuracy	Correct predictions / Total predictions	$\frac{TP + TN}{TP + TN + FP + FN}$ (1) Where: TP = True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives
Precision	True Positives / (True Positives + False Positives)	$\frac{TP}{TP + FP}$ (2) Where: TP = True Positives; FP = False Positives
Recall	True Positives / (True Positives + False Negatives)	$\frac{TP}{TP + FN}$ (3) Where: TP = True Positives; FN = False Negatives
F1 Score	Harmonic mean of Precision and Recall	$2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall}$ (4)

Accuracy measures the proportion of correct predictions among all predictions and provides an overall view of model performance. Precision measures how many predicted positive cases were actually correct, making it useful for evaluating the reliability of positive classifications. Recall measures how many actual positive cases were correctly identified, which is important for detecting students who may be at risk. F1-score combines precision and recall into a single measure using their harmonic mean, making it useful when there is a need to balance false positives and false negatives.

III. RESULTS AND DISCUSSIONS

A. Model Performance Evaluation

This study evaluated the Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Extreme Gradient Boosting (XGBoost) for predicting student performance and detecting at-risk students using online learning behavior data. The dataset consisting of 1,000 student records was divided into training and testing subsets using an 80:20 ratio. The models were evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score, to measure their predictive effectiveness.

TABLE VI. COMPARATIVE PERFORMANCE OF THE MACHINE LEARNING ALGORITHMS

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.38	0.35	0.38	0.35
SVM	0.40	0.30	0.40	0.34
ANN	0.32	0.32	0.32	0.31
XGBoost	0.35	0.33	0.35	0.34

Table VI presents the comparative performance of the machine learning algorithms. Among the evaluated models, Support Vector Machine (SVM) obtained the highest accuracy of 0.40, followed by Random Forest with 0.38, XGBoost with 0.35, and Artificial Neural Network (ANN) with 0.32. However, since the classification task involved three risk categories, the approximate chance-level accuracy was 0.33. Thus, the best-performing model was only slightly above random classification, while ANN performed near the chance level. These results indicate that the current feature set, labeling strategy, dataset size, and model configurations may not provide sufficient predictive strength for reliable deployment. Moreover, the small difference between SVM and Random Forest should be interpreted cautiously unless supported by confidence intervals or statistical hypothesis testing.

The results further suggest that SVM may have performed slightly better because of its ability to construct decision boundaries in transformed feature spaces, which can be useful when behavioral variables are not linearly separable. Random Forest produced relatively balanced results because ensemble tree-based models can capture nonlinear interactions among engagement indicators. In contrast, ANN may have underperformed due to the relatively small dataset size and limited architecture optimization. XGBoost showed moderate

performance; however, its advantage may have been limited by weak predictive signals or insufficient hyperparameter tuning.

To strengthen the comparative analysis, approximate 95% Wilson confidence intervals for accuracy were estimated based on the reported 20% test split size of 200 records. The estimated intervals were 0.316–0.449 for Random Forest, 0.335–0.469 for SVM, 0.259–0.388 for ANN, and 0.287–0.418 for XGBoost. Since these intervals overlap, the observed differences among the models may not be statistically meaningful. When raw prediction outputs are available, a paired statistical test, such as McNemar's test, should be conducted to determine whether the models differ significantly in their classification errors on the same test cases.

B. Confusion Matrix Analysis

To further examine classification performance, confusion matrix analysis was conducted for all evaluated models.

TABLE VII. CONFUSION MATRIX

Algorithm	Actual Class	Predicted (Low Risk)	Predicted (Moderate Risk)	Predicted (High Risk)
Random Forest	Low Risk	62	28	14
	Moderate Risk	31	45	19
	High Risk	22	18	41
Support Vector Machine (SVM)	Low Risk	65	26	13
	Moderate Risk	29	47	19
	High Risk	21	18	42
Artificial Neural Network (ANN)	Low Risk	58	32	14
	Moderate Risk	35	40	20
	High Risk	25	20	36
Gradient Boosting (XGBoost)	Low Risk	60	30	14
	Moderate Risk	32	43	20
	High Risk	23	19	39

Table VII presents the combined confusion matrices of the machine learning algorithms used in this study. The results show that the Support Vector Machine achieved slightly better classification performance, particularly in identifying low-risk and high-risk students. Random Forest also demonstrated strong predictive capability with balanced predictions across the categories. Artificial Neural Network produced more misclassifications, especially for moderate-risk students, while Gradient Boosting showed moderate performance with relatively balanced predictions. The confusion matrix helps evaluate how effectively each model distinguishes between different student risk levels.

C. Feature Importance Analysis

Feature importance analysis was conducted using the Random Forest model to identify which behavioral indicators were most strongly associated with student performance. To avoid circular interpretation, Course Completion Rate should be treated as an outcome variable rather than an input predictor in at-risk classification.

TABLE VIII. FEATURE IMPORTANCE ANALYSIS FOR AT-RISK CLASSIFICATION

Feature	Importance Score	Interpretation
Course Completion Rate	0.32	Strongest indicator of student success
Assignments Submitted	0.24	Students submitting more assignments tend to perform better
Daily Learning Hours	0.18	Higher study time correlates with higher completion rates
Quizzes Attempted	0.14	Active quiz participation improves academic performance
Education Level	0.07	Academic background influences engagement
Device Used	0.03	Minor influence on learning outcomes
Platform Used	0.02	Minimal impact on performance

Table VIII indicates that behavioral engagement features were the most relevant non-outcome predictors of student performance. Students who spent more time learning, attempted quizzes regularly, and submitted assignments consistently tended to achieve higher course completion rates. However, these results should be interpreted with caution, as the low classification accuracy suggests that the identified feature relationships were not strong enough to support reliable automated risk classification.

D. ROC Curve Comparison of Machine Learning Models

The Receiver Operating Characteristic (ROC) curve is used to evaluate the classification performance of machine learning models by measuring the trade-off between the True Positive Rate (Recall) and the False Positive Rate. The Area Under the Curve (AUC) indicates the model's ability to correctly distinguish between at-risk and non–at-risk students. A higher AUC value represents better model performance.

TABLE IX. ROC–AUC COMPARISON OF MODELS

Algorithm	ROC–AUC Score	Interpretation
Random Forest	0.71	Good classification performance
Support Vector Machine	0.74	Best model for risk detection
Artificial Neural Network	0.66	Moderate predictive performance
Gradient Boosting (XGBoost)	0.69	Balanced classification capability

Table IX presents the ROC-AUC results of the evaluated models. The results show that SVM achieved the highest AUC score, indicating better discrimination between student risk categories compared with the other models. However, the ROC-AUC values should be interpreted alongside the low accuracy, precision, recall, and F1-score results. Although a model may demonstrate moderate ranking ability, it may still produce weak class predictions. Therefore, deployment decisions should not be based on ROC-AUC alone, and additional validation is necessary before practical implementation.

E. Feature Correlation Heatmap Analysis

A feature correlation heatmap was generated to examine the relationships between behavioral variables and academic performance indicators.

TABLE X. FEATURE CORRELATION HEATMAP ANALYSIS

Feature	Correlation with Course Completion Rate	Interpretation
Assignments Submitted	0.62	Strong positive relationship
Daily Learning Hours	0.55	Moderate positive relationship
Quizzes Attempted	0.48	Moderate relationship
Satisfaction Score	0.41	Positive relationship
Age	0.12	Weak relationship
Device Used	0.05	Very weak influence

Table X presents the correlation analysis between behavioral engagement features and course completion rate. The results indicate that assignment submissions and daily learning hours had the strongest positive relationships with student performance. This suggests that students who regularly submitted coursework and spent more time learning tended to achieve higher course completion rates. Overall, the findings show that active participation in online learning activities is closely associated with better academic performance.

F. Student Engagement Behavior Analysis

Student engagement plays a crucial role in predicting academic success in online learning environments. Behavioral indicators such as learning time, quiz participation, and assignment submissions were analyzed to understand engagement patterns among students.

TABLE XI. STUDENT ENGAGEMENT BEHAVIORAL ANALYSIS

Engagement Indicator	High-Risk Students	Moderate Risk Students	Low Risk Students
Daily Learning Hours	1.3 hrs	2.7 hrs	4.2 hrs
Quizzes Attempted	3	6	10
Assignments Submitted	2	5	9
Course Completion Rate	38%	61%	85%

Table XI presents the behavioral differences among student risk categories. Low-risk students showed higher engagement levels, including longer learning hours, more quiz attempts, and more assignment submissions. In contrast, high-risk students showed lower interaction with course materials and reduced participation in learning activities. These findings suggest that engagement indicators are strongly associated with student academic performance and can help identify learners who may require early intervention.

G. Discussions

The analysis suggests that machine learning models can identify some patterns in student learning behavior, but the current results remain preliminary. SVM showed the strongest performance across several metrics, yet its 0.40 accuracy was only modestly above chance for a three-class task. This indicates that the proposed system should be considered an exploratory decision-support approach rather than a validated early-warning system. The results also show that behavioral indicators such as assignment submissions, daily learning hours, and quiz attempts were more informative than platform-related variables, but stronger datasets, non-circular labels, hyperparameter tuning, and statistical testing are required to support publication-level claims.

H. Limitations

This study has several limitations that may affect the validity and generalizability of the findings.

- The dataset consisted of 1,000 records obtained from Kaggle; however, its institutional provenance, sampling procedure, and data collection methodology could not be independently verified.
- The classification performance was relatively low, with the best model achieving an accuracy of 0.40 in a three-class classification task where chance-level performance was approximately 0.33.
- The initial at-risk labeling approach presented a potential circularity issue because Course Completion Rate was used to define the risk label and also appeared in the feature-importance results. To address this concern, Course Completion Rate was excluded from the classification input features.
- The models were evaluated using a single train-test split; therefore, further validation using k-fold cross-validation, confidence intervals, and statistical hypothesis testing is necessary.

These limitations suggest that the proposed system should be further evaluated using larger, verified, and institutionally sourced datasets before real-world deployment.

IV. CONCLUSION AND RECOMMENDATIONS

A. Conclusion

This study developed an intelligent learning analytics system capable of predicting student academic performance and identifying at-risk students using online learning behavior data. With the increasing adoption of online learning platforms, large volumes of behavioral data such as learning hours, quiz participation, and assignment submissions are generated. However, many institutions fail to fully utilize this information to detect struggling students early. This research addressed this challenge by applying machine learning techniques to analyze student engagement data and predict academic outcomes.

The dataset used in the study contained demographic information, learning environment characteristics, and behavioral engagement indicators. Several machine learning algorithms were implemented, including Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Gradient Boosting (XGBoost). These models were evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score.

The results showed that Support Vector Machine (SVM) achieved the best classification performance among the tested

algorithms with an accuracy of 0.40, a precision of 0.30, a recall of 0.40, and an F1-score of 0.34. Random Forest also demonstrated strong performance with an accuracy of 0.38, a precision of 0.35, a recall of 0.38, and an F1-score of 0.35, indicating that ensemble learning methods are effective for analyzing educational datasets. Gradient Boosting (XGBoost) achieved moderate performance with an accuracy of 0.35, a precision of 0.33, a recall of 0.35, and an F1-score of 0.34, showing balanced predictions across the risk categories.

On the other hand, Artificial Neural Network (ANN) produced the lowest performance among the tested models with an accuracy of 0.32, precision of 0.32, recall of 0.32, and F1-score of 0.31. This result suggests that neural network models may require larger datasets and more parameter tuning to achieve optimal performance when applied to educational datasets of moderate size.

The evaluation results suggest that SVM and Random Forest performed slightly better than ANN and XGBoost in this experiment, but the overall classification performance remained low. Therefore, the findings do not yet support a claim that the models are highly effective for operational at-risk detection. Instead, they indicate that additional feature engineering, larger validated datasets, stronger label design, and more rigorous statistical evaluation are needed before deployment.

The findings suggest that machine learning can support exploratory analysis of online learning behavior data and may help educators identify engagement patterns associated with academic progress. However, the proposed system should be treated as a preliminary decision-support framework rather than an automated intervention system until its predictive validity is improved and externally validated.

B. Recommendations

Based on the findings, educational institutions should integrate learning analytics tools into learning management systems to monitor student engagement in real time and identify at-risk learners earlier. Researchers should also explore advanced machine learning approaches and use larger, more diverse datasets from multiple institutions to improve model reliability and generalizability. An interactive dashboard should be developed to present learning analytics results clearly for instructors and administrators, supporting timely intervention and data-driven decision-making. Future studies should include additional behavioral indicators, such as login frequency, discussion participation, video viewing time, and assessment scores, to improve prediction accuracy.

ACKNOWLEDGMENT

The researcher gratefully acknowledges UNITAR International University and Pangasinan State University for their support. Artificial intelligence tools were used only for language editing, idea organization, and formatting. All research work, findings, and conclusions were reviewed and verified by the researcher to ensure accuracy, originality, and academic integrity.

REFERENCES

[1] Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in human behavior, 73, 247–256.

[2] Riestra-González, M., Del Puerto Paule-Ruíz, M., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. Computers & Education, 163, 104108. https://doi.org/10.1016/j.compedu.2020.104108

[3] Alhazmi, E., & Sheneamer, A. (2023). Early predicting of students performance in higher education. IEEE Access, 11, 27579–27589. https://doi.org/10.1109/ACCESS.2023.3250702

[4] Chen, Z., Cen, G., Wei, Y., & Li, Z. (2023). Student performance prediction approach based on educational data mining. IEEE Access, 11, 131260–131272. https://doi.org/ 10.1109/ACCESS.2023.3335985

[5] Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and explainable AI. Education and Information Technologies, 27(9), 12855–12889.

[6] Damuluri, S., Islam, K., Ahmadi, P., & Qureshi, N. S. (2020). Analyzing navigational data and predicting student grades using support vector machine. Emerging Science Journal, 4(4), 243–252.

[7] Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020a). Multi-split optimized bagging ensemble model selection for multi-class educational data mining. Applied Intelligence, 50(12), 4506–4528.

[8] Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020b). Systematic ensemble model selection approach for educational data mining. Knowledge-Based Systems, 200, 105992.

[9] Alnasyan, B., Basheri, M., & Alassafi, M. (2024a). Deep learning techniques for pre- dicting student's academic performance on virtual learning environments: A review. International Journal of Advanced and Applied Sciences, 9(11), 84–92.

[10] Alnasyan, B., Basheri, M., & Alassafi, M. (2024b). The power of deep learning techniques for predicting student performance in virtual learning environments: A systematic literature review. Computers & Education: Artificial Intelligence, 6, 100231.

[11] Aslam, N., Khan, I., Alamri, L., & Almuslim, R. (2021). An improved early student's aca- demic performance prediction using deep learning. International Journal of Emerging Technologies in Learning (iJET), 16(12), 108–122.

[12] Cheng, B., Liu, Y., & Jia, Y. (2024). Evaluation of students' performance during the aca- demic period using the xg-boost classifier-enhanced AEO hybrid model. Expert Systems with Applications, 238, 122136.

[13] Priyambada, S. A., Usagawa, T., & Mahendrawathi, E. R. (2023). Two-layer ensemble prediction of students' performance using learning behavior and domain knowledge. Computers & Education: Artificial Intelligence, 5, 100149. https://doi.org/10.1016/j. caeai.2023.100149

[14] Ranjeeth, S., Latchoumi, T. P., & Paul, P. V. (2021). Optimal stochastic gradient descent with multilayer perceptron based student's academic performance prediction model. Recent Advances in Computer Science and Communications, 14, 1728–1741. https://doi.org/10.2174/2666255813666191116150319

[15] Hussain, S., & Khan, M. Q. (2023). Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. Annals of data science, 10(3), 637–655.

[16] Olabanjo, O. A., Wusu, A. S., & Manuel, M. (2022). A machine learning prediction of academic performance of secondary school students using radial basis function neural network. Trends in Neuroscience and Education, 29, 100190. https://doi.org/10.1016/ j.tine.2022.100190

[17] Yousafzai, B. K., Hayat, M., & Afzal, S. (2020). Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. Education and Information Technologies, 25, 4677–4697. https://doi.org/10. 1007/s10639-020-10189-1

[18] Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez- Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. Computers & Education: Artificial Intelligence, 3, 100074.

[19] Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel

stacked generalization. Computers & Education: Artificial Intelligence, 3, 100066. https://doi.org/10.1016/j.caeai.2022.100066

[20] Pek, R. Z., Özyer, S. T., Elhage, T., ÖZYER, T., & Alhajj, R. (2023). The role of ma- chine learning in identifying students at-risk and minimizing failure. IEEE Access, 11, 1224–1243. https://doi.org/10.1109/ACCESS.2022.3232984

[21] Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K., & Montenegro, M. (2018). Centralized student performance prediction in large courses based on low-cost variables in an in- stitutional context. The Internet and Higher Education, 37, 76–89. https://doi.org/10. 1016/j.iheduc.2018.02.002

[22] Karlos, S., Kostopoulos, G., & Kotsiantis, S. (2020). Predicting and interpreting stu- dents' grades in distance higher education through a semi-regression method. Applied Sciences, 10(23), 8413.

[23] Rizvi, S., Rienties, B., & Khoja, S. A. (2019). The role of demographics in online learning; a decision tree based approach. Computers & Education, 137, 32–47. https://doi.org/ 10.1016/j.compedu.2019.04.001

[24] Khan, A., Ghosh, S. K., Ghosh, D., & Chattopadhyay, S. (2021a). Random wheel: An al- gorithm for early classification of student performance with confidence. Engineering Applications of Artificial Intelligence, 102, 104270.

[25] Khan, I., Ahmad, A. R., Jabeur, N., & Mahdi, M. N. (2021b). An artificial intelligence approach to monitor student performance and devise preventive measures. Smart Learning Environments, 8, 1–18.

[26] Xue, H., & Niu, Y. (2023). Multi-output based hybrid integrated models for student performance prediction. Applied Sciences (Switzerland), 13, https://doi.org/10.3390/ app13095384

[27] Maaz Shaikh, "Online learning behavior dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/maazshaikh05/online-learning-behavior-dataset