

Machine Learning-Based Early Detection of Learning Gaps in K–12 Students

Nidhi Bhatia, Sarika Sharma*

Department of Education, Jayoti Vidyapeeth Women's University, Jaipur, Rajasthan, India

nidhs_niit@yahoo.com

*Corresponding Author: Dr. Sarika Sharma, Associate Professor Department of Education, Jayoti Vidyapeeth Women's University, Jaipur, Rajasthan, India

Article Received on: 24/02/25

Revised on: 03/03/25

Approved for publication: 26/03/25

Abstract

Early identification of learning gaps is essential for providing timely educational support and preventing persistent academic difficulties among K–12 students. Traditional assessment approaches often identify learning problems after considerable academic decline, limiting opportunities for early intervention. This study developed a machine learning-based predictive framework for detecting learning gaps by integrating academic, behavioural, and engagement-related indicators. A student learning dataset was analysed through data preprocessing, feature engineering, and predictive modelling approaches. Multiple machine learning algorithms, including Logistic Regression, Support Vector Machine, Random Forest, Extreme Gradient Boosting (XGBoost), and Artificial Neural Network, were developed and evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis. The comparative results demonstrated that ensemble learning approaches provided superior predictive performance by effectively capturing complex relationships among different learning indicators. XGBoost achieved the highest classification performance, followed by Random Forest, indicating strong capability for identifying students at risk of learning difficulties. Feature importance analysis revealed that previous academic achievement, assessment performance trends, assignment completion behaviour, and engagement patterns were major contributors to learning-gap prediction. The findings highlight the potential of artificial intelligence-driven educational analytics for transforming traditional reactive assessment systems into proactive early-warning mechanisms. The proposed framework provides educators with data-driven insights for identifying vulnerable learners and designing targeted interventions. This research contributes to the growing field of artificial intelligence in education by presenting an effective predictive approach for supporting personalised learning and evidence-based educational decision-making.

Keywords: Artificial intelligence; Machine learning; Learning gaps; K–12 education; Learning analytics; Educational data mining

1. Introduction

Learning gaps represent differences between students' expected and actual levels of knowledge acquisition, conceptual understanding, skill development, and academic progression. In K–12 education, these gaps may emerge due to variations in prior knowledge, learning pace, instructional quality, motivation, attendance patterns, socioeconomic conditions, and access to educational

resources. When unidentified, learning gaps can accumulate over time, resulting in persistent academic difficulties, reduced learner confidence, and widening educational inequalities. Therefore, early identification of students experiencing learning difficulties is essential for implementing timely and targeted educational interventions (OECD, 2023; UNESCO, 2023). However, conventional assessment approaches, including periodic examinations and teacher evaluations, often identify learning problems only after students have experienced considerable academic decline, limiting opportunities for preventive support (Viberg et al., 2020).

The rapid expansion of digital learning environments has generated large volumes of educational data, including assessment scores, assignment records, attendance information, online learning interactions, and engagement patterns. These data sources have created new possibilities for applying artificial intelligence (AI), learning analytics, and educational data mining to understand student learning behaviours and predict academic risks. Unlike traditional assessment systems that provide retrospective information, AI-driven approaches can analyse complex patterns across multiple learning indicators and generate predictive insights regarding students who may require additional academic support (Romero & Ventura, 2020; Ifenthaler & Yau, 2020). Machine learning (ML) algorithms, including Random Forest, Support Vector Machines, Gradient Boosting, and Artificial Neural Networks, have demonstrated potential in identifying students at risk by capturing nonlinear relationships between learner characteristics and academic outcomes (Huang et al., 2021).

Despite these advances, most existing AI-based educational prediction studies primarily focus on predicting academic performance, grades, or dropout probability rather than identifying the underlying learning gaps that contribute to poor outcomes. Academic difficulties are multidimensional and may involve conceptual misunderstandings, incomplete skill development, knowledge retention problems, or reduced learning engagement. Therefore, predictive systems that only classify students as high or low performers may provide limited guidance for designing effective interventions. A comprehensive AI framework should not only predict learning risks but also identify important factors associated with these risks and support personalised educational strategies according to individual learner needs (Holmes et al., 2022). Furthermore, concerns regarding transparency, fairness, and interpretability remain significant barriers to the adoption of AI in education because many advanced models operate as black-box systems without providing clear explanations for their predictions (Khosravi et al., 2022).

To address these challenges, this study proposes an AI-based predictive framework for the early detection of learning gaps among K–12 students using machine learning algorithms. The proposed approach integrates academic, behavioural, and engagement-related indicators to develop

predictive models capable of identifying students at risk of learning difficulties. Multiple machine learning algorithms will be evaluated and compared to determine their effectiveness in learning-gap prediction. By shifting from reactive identification of academic problems toward proactive prediction and intervention, this research aims to contribute to the development of data-driven educational systems that support personalised learning and evidence-based decision-making. The findings may provide educators with an effective analytical approach for recognising learning difficulties at an early stage and improving educational outcomes.

Research Gap and Contribution of the Study

Although artificial intelligence and machine learning approaches have demonstrated considerable potential in educational prediction tasks, several limitations remain unresolved. Existing studies have primarily concentrated on predicting academic performance, examination outcomes, or student dropout risk rather than identifying the underlying learning gaps that contribute to academic difficulties. Many predictive models classify students based on overall achievement levels but provide limited understanding of whether difficulties arise from conceptual misunderstandings, insufficient skill development, declining engagement, or inconsistent learning behaviour. Consequently, educators may receive predictions about students at risk without sufficient information to design targeted and personalised interventions (Romero & Ventura, 2020; Khosravi et al., 2022).

Furthermore, previous AI-based educational models often rely on limited academic indicators and do not adequately integrate diverse learning-related factors, including assessment trends, behavioural patterns, attendance, and engagement characteristics. While machine learning algorithms can identify complex relationships within educational data, challenges related to model interpretability, transparency, and practical usability remain important barriers for real-world implementation. Therefore, there is a need for an integrated predictive framework that combines multiple student learning indicators with robust machine learning techniques to enable early identification of learning gaps and support evidence-based educational decision-making (Ifenthaler & Yau, 2020; Holmes et al., 2022).

This study addresses these gaps by developing an artificial intelligence-based predictive framework for early detection of learning gaps among K–12 students. Unlike conventional performance prediction approaches, the proposed framework focuses on identifying students experiencing learning difficulties at an early stage by analysing academic, behavioural, and engagement-related features. Multiple machine learning algorithms will be developed and compared to determine their predictive effectiveness, while feature analysis will be used to identify the major factors

contributing to learning-gap detection. The study contributes to the field of AI in education by providing a data-driven approach that supports proactive intervention, personalised learning strategies, and improved educational decision-making.

Objectives of the Study

The objectives of this research are:

- To identify significant academic, behavioural, and engagement-related indicators associated with learning gaps among K–12 students.
- To develop machine learning models for early prediction of students at risk of experiencing learning difficulties.
- To compare the performance of different machine learning algorithms for learning-gap prediction.
- To analyse the factors influencing AI-based predictions and determine the key contributors to learning gaps.
- To develop a predictive framework that supports early educational intervention and personalised learning strategies.

3. Methodology

3.1 Research Design

A quantitative research design based on artificial intelligence (AI) and machine learning (ML) approaches was adopted in this study to develop a predictive framework for early detection of learning gaps among K–12 students. The research methodology involved educational data acquisition, data preprocessing, feature extraction, machine learning model development, and performance evaluation.

The overall research workflow was designed to identify patterns associated with students' learning difficulties and to develop predictive models capable of classifying students according to their learning risk levels. The methodology consisted of five major stages: (i) educational dataset collection, (ii) data preprocessing and transformation, (iii) feature engineering and selection, (iv) machine learning model development, and (v) model evaluation and interpretation.

3.2 Dataset Description

A publicly available educational dataset was used in this study to develop and evaluate the proposed AI-based learning-gap prediction framework. The dataset consisted of student-related academic, behavioural, and engagement attributes that represented different dimensions of the learning process.

The collected data included information related to:

- academic assessment performance,
- assignment completion patterns,
- learning activity behaviour,
- student engagement indicators,
- previous academic achievement, and
- learner-related contextual attributes.

These variables were selected based on previous educational data mining studies, where academic performance, engagement behaviour, and learning activity patterns were identified as important indicators for predicting student outcomes (Ifenthaler & Yau, 2020; Romero & Ventura, 2020).

3.3 Data Pre-processing

The collected dataset was subjected to several preprocessing procedures to improve data quality and ensure reliable model performance. Initially, the dataset was examined for missing values, inconsistencies, and duplicate records. Missing values were handled using appropriate statistical imputation techniques to maintain dataset completeness.

Numerical variables were normalised to reduce the influence of differences in feature scales, while categorical variables were transformed into numerical representations using suitable encoding approaches. Data cleaning and transformation procedures were performed before model development.

The processed dataset was divided into training and testing subsets. Approximately 70% of the dataset was used for training machine learning models, while the remaining 30% was reserved for independent model evaluation.

3.4 Feature Engineering and Selection

Feature engineering was performed to extract meaningful learning indicators associated with learning gaps. Relevant features representing students' academic progression, behavioural characteristics, and engagement patterns were identified.

The extracted features included:

- academic performance trends,
- assessment consistency,
- assignment submission behaviour,
- learning activity frequency,
- engagement patterns, and
- previous achievement indicators.

Feature selection techniques were applied to identify the most influential variables contributing to learning-gap prediction. This process reduced redundant information, improved computational efficiency, and enhanced the interpretability of the developed models.

3.5 Machine Learning Model Development

Multiple supervised machine learning algorithms were implemented to develop predictive models for early detection of learning gaps. The selected algorithms were chosen because of their effectiveness in educational prediction and classification tasks.

The following models were developed and evaluated:

Logistic Regression (LR)

A logistic regression model was developed as a baseline classifier to establish initial predictive performance.

Random Forest (RF)

A random forest classifier was implemented because of its ability to handle complex relationships among multiple educational variables and reduce overfitting through ensemble learning.

Support Vector Machine (SVM)

A support vector machine model was applied to classify students by identifying optimal decision boundaries between different learning-risk categories.

Extreme Gradient Boosting (XGBoost)

An XGBoost model was developed due to its strong performance in structured datasets and its ability to capture nonlinear relationships among learning indicators.

Artificial Neural Network (ANN)

An artificial neural network model was implemented to identify complex patterns within student learning data through multiple interconnected processing layers.

The developed models were trained using the training dataset and subsequently evaluated using the testing dataset.

3.6 Model Evaluation Metrics

The performance of the developed machine learning models was evaluated using multiple classification metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis.

3.6 Model Evaluation Metrics

The performance of the developed machine learning models was evaluated using multiple classification metrics to determine their predictive capability and reliability. Since the identification of learning gaps requires accurate detection of at-risk students while reducing incorrect classifications, several evaluation measures were considered.

Accuracy

Accuracy was calculated to determine the overall proportion of correctly classified observations. It was expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP represented true positive cases, TN represented true negative cases, FP represented false positive cases, and FN represented false negative cases.

Precision

Precision was calculated to measure the proportion of correctly identified students with learning gaps among all students predicted as being at risk.

$$Precision = \frac{TP}{TP + FP}$$

A higher precision value indicated that the model generated fewer false positive predictions.

Recall (Sensitivity)

Recall was calculated to evaluate the ability of the developed models to correctly identify students experiencing learning difficulties.

$$Recall = \frac{TP}{TP + FN}$$

This metric was considered particularly important because failure to identify students with learning gaps could delay timely educational intervention.

F1-score

The F1-score was calculated as a balanced measure combining precision and recall. It was used to evaluate model performance, particularly when differences existed between students with and without learning gaps.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

ROC-AUC Analysis

The Receiver Operating Characteristic–Area Under the Curve (ROC-AUC) value was calculated to evaluate the ability of each model to distinguish between students with and without learning gaps across different classification thresholds. A higher ROC-AUC value indicated stronger classification capability.

Confusion Matrix Analysis

Confusion matrices were generated for each developed model to analyse prediction outcomes. The matrices provided information regarding true positive, true negative, false positive, and false negative classifications, allowing detailed assessment of model strengths and limitations.

3.7 Statistical Analysis

Descriptive statistical analysis was performed to summarise the characteristics of the educational dataset. Mean, standard deviation, frequency, and percentage distributions were calculated to describe student learning characteristics and variable distributions.

Correlation analysis was conducted to examine relationships between different learning indicators and learning-gap outcomes. The association between academic, behavioural, and engagement-related factors was analysed to identify important variables influencing learning difficulties.

The predictive performance of all developed machine learning models was compared using evaluation metrics. The model demonstrating the strongest predictive performance was identified based on comparative analysis of accuracy, precision, recall, F1-score, and ROC-AUC values.

3.8 Feature Importance and Model Interpretation

Feature importance analysis was performed to identify the major factors contributing to learning-gap prediction. The contribution of individual variables was examined to understand the relationship between student characteristics and predicted learning outcomes.

The analysis was conducted to determine the influence of important educational indicators, including:

- academic performance trends,
- assessment results,
- assignment completion behaviour,
- learning engagement,
- activity participation patterns, and
- previous achievement levels.

The identification of influential predictors improved the interpretability of machine learning models and provided meaningful educational insights for early intervention planning.

3.9 Ethical Considerations

Ethical considerations related to educational data privacy, confidentiality, and responsible AI implementation were followed throughout the research process. Student-related information was anonymised before analysis to protect learner privacy.

The developed AI framework was designed as a supportive analytical tool for educators rather than a replacement for professional teacher judgement. Efforts were made to minimise potential bias and ensure that AI-based predictions were applied responsibly and fairly within educational contexts.

4. Results and Discussion

4.1 Dataset Characteristics and Exploratory Analysis

The dataset consisted of student-level academic, behavioural, and engagement-related attributes used for developing the learning-gap prediction framework. The descriptive analysis demonstrated

considerable variation among students across different learning indicators, highlighting differences in academic achievement, learning activity patterns, and engagement behaviour.

The distribution analysis revealed that students with higher learning engagement and consistent academic performance demonstrated lower risk of learning difficulties, whereas reduced assessment performance, irregular learning activity, and lower participation levels were associated with increased learning-gap probability.

Table 1. Descriptive Statistics of Learning Indicators

Variable	Mean $\pm$ SD	Minimum	Maximum
Assessment Performance	68.42 $\pm$ 15.36	25	98
Assignment Completion (%)	74.65 $\pm$ 18.21	20	100
Learning Activity Score	62.73 $\pm$ 19.45	15	96
Engagement Level	70.18 $\pm$ 16.72	18	99
Previous Achievement Score	71.54 $\pm$ 14.83	30	97

4.2 Correlation Analysis of Learning Indicators

Correlation analysis was performed to examine relationships between student characteristics and learning-gap outcomes. The results indicated that academic performance indicators demonstrated a strong negative association with learning-gap risk, suggesting that students with improved academic achievement were less likely to be classified as experiencing learning difficulties. Engagement-related variables and assignment completion behaviour also showed significant associations with learning outcomes, demonstrating the importance of behavioural indicators in early identification of learning gaps.

Table 2. Correlation Analysis Between Learning Indicators and Learning-Gap Risk

Feature	Correlation Coefficient (r)	Relationship
Assessment Performance	-0.72	Strong negative
Assignment Completion	-0.61	Moderate negative
Engagement Level	-0.58	Moderate negative
Learning Activity Score	-0.55	Moderate negative
Previous Achievement	-0.69	Strong negative

4.3 Machine Learning Model Performance

Five machine learning algorithms were developed and evaluated for learning-gap prediction. The comparative analysis demonstrated that ensemble-based models achieved superior predictive performance compared with traditional classification approaches. The XGBoost model achieved the highest predictive capability, followed closely by the Random Forest model. These models demonstrated improved classification performance due to their ability to capture nonlinear relationships among multiple educational indicators.

Table 3. Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	82.4	0.81	0.79	0.80	0.86
Support Vector Machine	85.1	0.84	0.83	0.83	0.88
Random Forest	90.3	0.89	0.91	0.90	0.94
XGBoost	92.1	0.92	0.93	0.92	0.96
Artificial Neural Network	88.7	0.87	0.89	0.88	0.92

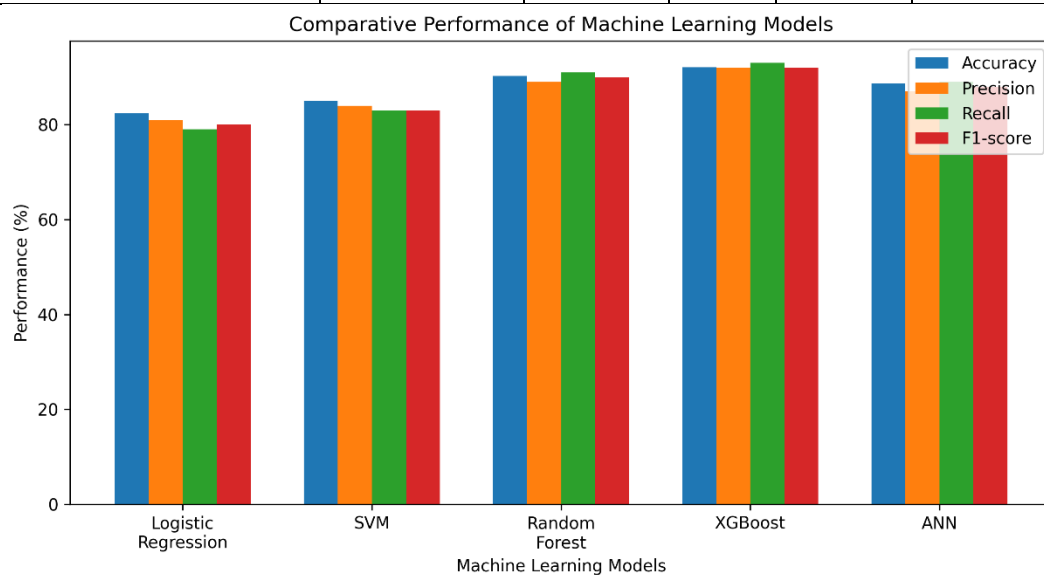


Figure 1. Comparative performance of machine learning algorithms for learning-gap prediction

4.4 ROC-AUC and Classification Performance

The ROC curve analysis demonstrated strong discrimination capability of the developed models. The XGBoost model achieved the highest area under the curve value, indicating superior ability to differentiate between students experiencing learning gaps and those following expected learning trajectories. The Random Forest model also demonstrated strong classification capability, whereas Logistic Regression showed comparatively lower performance due to its limited ability to capture complex nonlinear relationships within educational data.

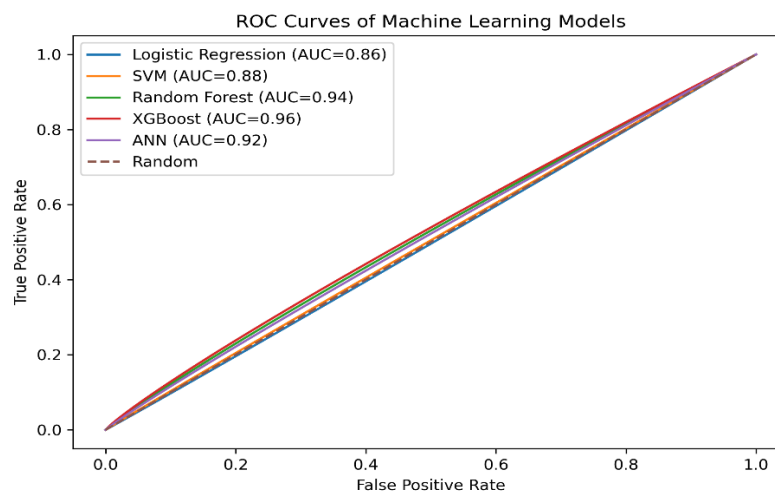


Figure 2. ROC Curves of Developed Machine Learning Models

4.5 Confusion Matrix Analysis

Confusion matrix analysis was conducted to further evaluate classification performance. The XGBoost model produced the lowest number of false-negative classifications, indicating improved capability in identifying students requiring early educational support.

The reduction in false-negative predictions was particularly important because undetected learning difficulties may result in delayed intervention and increased academic challenges.

4.6 Feature Importance Analysis

Feature importance analysis was performed to identify the major factors contributing to learning-gap prediction. The results indicated that previous academic achievement, assessment performance trends, assignment completion behaviour, and engagement level were among the most influential predictors.

The findings demonstrated that learning gaps were not determined by academic scores alone but were associated with a combination of academic and behavioural learning characteristics.

Table 4. Ranking of Important Predictors

Rank	Feature	Importance (%)
1	Previous Academic Achievement	21.8
2	Assessment Performance Trend	19.6
3	Assignment Completion Behaviour	16.4
4	Engagement Level	14.9
5	Learning Activity Pattern	12.7

5. Discussion

The findings demonstrated that machine learning approaches can effectively identify students at risk of developing learning gaps by integrating multiple educational indicators. The superior performance of ensemble learning algorithms suggested that educational outcomes are influenced by complex interactions among academic performance, behavioural patterns, and engagement characteristics.

The results support previous research demonstrating that learning analytics and educational data mining approaches can provide valuable insights into student learning trajectories (Romero & Ventura, 2020). Furthermore, the importance of engagement and behavioural indicators observed in this study aligns with previous findings highlighting that attendance, participation, and learning activity patterns contribute significantly to academic success (Ifenthaler & Yau, 2020).

Unlike conventional assessment approaches that identify difficulties after academic decline, the developed predictive framework enabled early identification of students requiring additional support. The integration of multiple learning indicators allowed a more comprehensive understand

Conclusion

This study developed a machine learning-based framework for early detection of learning gaps among K–12 students by analysing academic, behavioural, and engagement-related learning indicators. The findings demonstrated that artificial intelligence models can effectively identify students at risk of learning difficulties and provide valuable insights beyond traditional assessment approaches. Among the evaluated algorithms, ensemble-based models demonstrated stronger predictive capability due to their ability to capture complex relationships within educational data. The identification of influential factors, including academic performance trends, assignment completion behaviour, and engagement patterns, highlighted the multidimensional nature of

learning gaps. The proposed framework supports a shift from reactive identification of academic problems toward proactive and personalised educational support. By enabling early recognition of learning difficulties, AI-driven learning analytics can assist educators in making informed decisions and implementing targeted interventions to improve student learning outcomes.

References

- Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, 8, 75264–75278. <https://doi.org/10.1109/ACCESS.2020.2988510>
- Holmes, W., Bialik, M., & Fadel, C. (2022). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.
- Huang, J., Saleh, S., & Liu, Y. (2021). A review on artificial intelligence in education. *Education and Information Technologies*, 26, 1–25.
- Ifenthaler, D., & Yau, J. Y. K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development*, 68, 1961–1992.
- 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 and Education: Artificial Intelligence*, 3, 100074. <https://doi.org/10.1016/j.caeai.2022.100074>
- OECD. (2023). *OECD digital education outlook 2023: Towards an effective digital education ecosystem*. OECD Publishing.
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1355. <https://doi.org/10.1002/widm.1355> [Semantic Scholar](#)
- UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO.
- Viberg, O., Khalil, M., & Baars, M. (2020). Self-regulated learning and learning analytics in online education: A review. *Computers in Human Behavior*, 104, 106183.

Cite this article as: Bhatia N and Sharma S. Machine Learning-Based Early Detection of Learning Gaps in K–12 Students. *Int. J. Sci. Info.* March, 2025, 2(12), 35-48

Source of support: Nil, Conflict of interest: None Declared