International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 3, May-June 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Early Identification of At-Risk Students Using Machine Learning

T.Thirunavukarasu*, P.Logaiyan**
*(Student, Department of Master Computer Application, Sri Manakula Vinayagar Engineering College, Pondicherry, India
Email:thirunavakarasut10@gmail.com)
** (Professor, Department of Master Computer Application, Sri Manakula Vinayagar Engineering College, Pondicherry, India
Email: logaiyanmca@smvec.ac.in)

----------**********----------

Abstract:

Early identification of at-risk students is a critical challenge in academic institutions seeking to improve student retention and outcomes. Traditional approaches to monitoring student performance rely on manual assessment, which is often subjective and fails to capture the complex interplay of factors affecting academic risk.

This paper presents the development of an early warning system for at-risk students using performance analytics and machine learning models. The proposed system employs classification models including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost to analyze student academic data. The system evaluates student risk based on key indicators such as assignment completion rates, attendance, assessment scores, and engagement levels, and classifies each student into High-Risk, Medium-Risk, and Low-Risk categories.

The proposed early identification system enables timely academic interventions, reduces manual monitoring effort, and improves the reliability of the at-risk identification process.

Keywords: *At-risk students, machine learning, early warning system, classification, Random Forest, XGBoost, educational data mining.*

----------**********----------

I. INTRODUCTION

Student retention and academic success are among the most important indicators of the effectiveness of any educational institution. The ability to identify students who are at risk of failing or dropping out early in the academic cycle is crucial for enabling timely interventions. However, in most institutions, the monitoring of student performance is carried out using conventional, often subjective methods. These approaches rely heavily on the judgment of academic advisors and are frequently constrained by the volume and variety of available student data.

Traditional forecasting methods such as Moving

Recent advances in data analytics and machine learning offer new opportunities to address these limitations. Machine Learning (ML) models can systematically analyze large volumes of student data—encompassing academic performance, attendance, and engagement—and detect patterns that are difficult for human evaluators to identify manually.

This paper presents the development of an early student risk identification system leveraging

performance analytics and machine learning. The system considers multiple student attributes, including assignment completion, attendance, assessment scores, and learning engagement, and classifies students into risk tiers: High-Risk, Medium-Risk, and Low-Risk.

The system applies and compares several supervised classification algorithms: Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and XGBoost. Each model is trained and evaluated for accuracy and generalizability before the best-performing model is selected for deployment.

II. LITERATURE SURVEY

The early identification of at-risk students has attracted increasing research interest in recent years. Traditional methods based on end-of-term grade reviews and advisor intuition have proven inadequate for timely intervention. Researchers have turned to data-driven and machine learning approaches to improve early warning systems.

A study [1] explored the application of machine learning techniques such as Decision Trees and Logistic Regression for predicting student academic failure. The results demonstrated improvements over manual review approaches, though limitations in handling non-linear relationships among academic features were noted.

Further research [2] compared ensemble and non-ensemble classification algorithms—including Random Forest, Support Vector Machines, and K-Nearest Neighbors—for student performance prediction. Ensemble methods, especially Random Forest, showed superior accuracy; however, challenges related to computational overhead were identified.

Work on feature engineering [3] highlighted the importance of selecting meaningful academic indicators such as attendance, assignment submission rates, and mid-term scores. Careful feature selection was found to substantially improve prediction performance, though many early warning systems still lacked real-time capabilities.

More recent research [4] applied gradient boosting methods, including XGBoost, to student risk prediction, achieving state-of-the-art results. Challenges related to model interpretability and scalability across diverse institutional contexts remained open research questions.

Based on the literature, it is evident that while individual ML approaches offer promise, systems combining multiple models with systematic evaluation are more likely to generalize effectively across diverse student populations. This paper addresses that gap.

III. THEORETICAL FRAMEWORK

The proposed system is grounded in data-driven decision-making principles, applying supervised machine learning to identify at-risk students based on historical academic data. Rather than relying on human intuition, the framework extracts patterns from student performance indicators and uses these to produce reliable risk predictions.

Classification algorithms form the core of this framework, mapping student feature vectors to risk categories: High-Risk, Medium-Risk, and Low-Risk. Six classification approaches are considered: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost. Each algorithm processes the data differently, allowing for a comparative analysis of their suitability for the task.

The framework also incorporates data preprocessing techniques to handle missing values, normalize features, and encode categorical variables. Key predictive features—including assignment completion rate, attendance percentage, assessment scores, and student engagement metrics—are prioritized based on their predictive relevance. This theoretical foundation supports the development of a robust and interpretable early warning system.

IV. METHODOLOGY

The proposed system follows a systematic methodology for early at-risk student identification using machine learning. Initially, student academic data—including assignment completion rates, attendance records, assessment scores, and engagement metrics—is collected from institutional sources. Data preprocessing is applied to handle missing values, correct inconsistencies, and normalize features for model compatibility.

Once preprocessed, the dataset is split into training and testing subsets. Multiple classification models—Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and XGBoost—are trained on the training data and evaluated on the testing data. Model performance is compared using standard evaluation metrics including accuracy, precision, recall, and F1-score.

The best-performing model is selected and used to classify students into risk categories: High-Risk, Medium-Risk, and Low-Risk. These classifications are made available to academic advisors and institutional decision-makers to inform timely interventions.

V. EXISTING SYSTEM

The prevailing approach to identifying at-risk students in most academic institutions relies on conventional, end-of-semester review processes. Academic advisors and faculty evaluate student performance based on grade reports and subjective observation. This process is inherently retrospective—risk is typically identified only after poor performance has already occurred, leaving limited time for effective intervention.

A significant limitation of existing systems is evaluation bias. Inconsistent criteria across instructors and departments result in unequal treatment of students who may exhibit similar risk profiles. Furthermore, performance reviews are conducted at fixed intervals, preventing real-time identification of emerging academic difficulties.

Another critical shortcoming is the fragmentation of student data. Attendance records, assessment scores, assignment submission data, and engagement metrics are often maintained in separate systems, making it difficult to obtain a holistic view of student academic health. These limitations collectively underscore the need for a more integrated, data-driven approach to early risk identification.

VI. PROPOSED SYSTEM

The proposed system offers a machine learning-based framework for the early identification of at-risk students. Unlike conventional evaluation methods that depend on periodic, subjective reviews, the proposed data-driven system enables objective, continuous assessment of student academic health.

Student risk is evaluated based on a comprehensive set of academic indicators: assignment completion rates, attendance frequency, assessment performance, and engagement consistency. Data from these sources is integrated, preprocessed, and fed into a suite of supervised classification models.

Six classification algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, and XGBoost—are trained and evaluated for their ability to accurately identify at-risk students. The model demonstrating the best performance on metrics such as accuracy, Finally, the selected model classifies students into risk tiers, and results are presented to academic advisors in the form of visualized reports, supporting timely and targeted intervention.

VII. SYSTEM ARCHITECTURE

The system architecture describes the end-to-end flow of the proposed early warning system, from raw data collection to actionable risk classification outputs.

The first stage is data collection, where academic data is gathered from institutional sources such as

learning management systems, attendance registers, and assessment records. These data sources provide both quantitative and qualitative indicators of student engagement and performance.

This is followed by data preprocessing, in which missing values are handled, inconsistencies corrected, and features normalized and encoded to ensure compatibility with machine learning algorithms.

The processed dataset is then partitioned into training and testing sets. Multiple classification models—Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and XGBoost—are trained on the training partition to learn predictive patterns from historical student data.

Each trained model is evaluated using performance metrics: accuracy, precision, recall, and F1-score. Based on this evaluation, the best-performing model is selected for student risk prediction.

The results of this study validate the effectiveness of machine learning approaches for the early identification of at-risk students. By incorporating multiple academic performance indicators, the proposed system delivers a structured and objective basis for risk assessment that surpasses conventional manual review.

Ensemble methods such as Random Forest and XGBoost consistently outperformed other algorithms, attributed to their ability to aggregate multiple decision trees and model complex feature interactions while mitigating overfitting. In contrast, Logistic Regression and SVM showed comparatively lower accuracy, reflecting the non-linear nature of academic risk patterns.

The choice of predictive features proved critical to system performance. Assignment completion rates, attendance, and assessment scores were found to be highly discriminative indicators of academic risk. Rigorous data preprocessing further enhanced model reliability.

The continuous assessment capability of the ML-based approach represents a major practical advantage over periodic manual reviews. However, limitations remain: prediction quality is contingent on data completeness and accuracy, and institutional privacy considerations must be carefully managed when handling student data.

[Fig. 1: System Architecture Diagram]

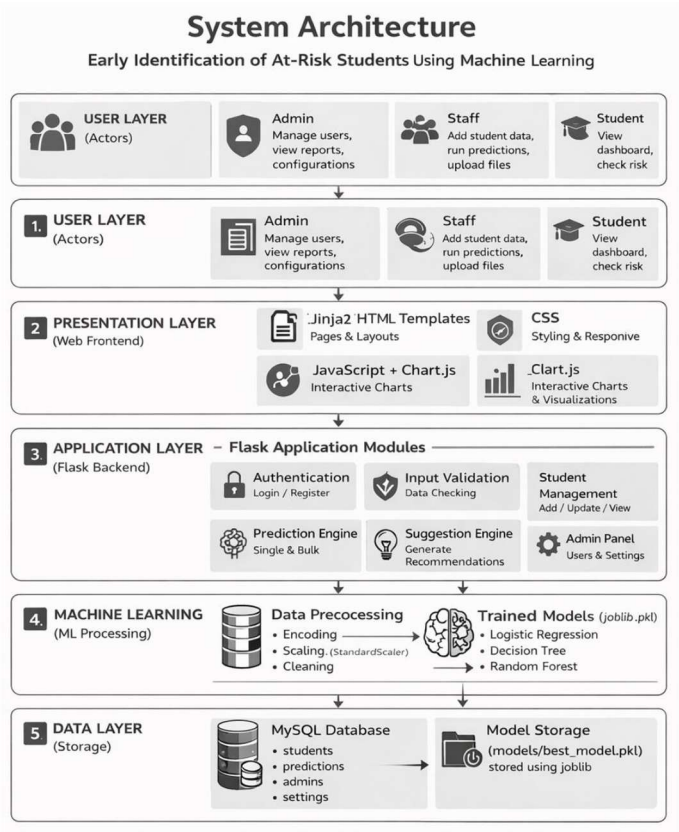

Fig. 1 System Architecture

VIII. RESULT AND DISCUSSION

Several machine learning classification algorithms were applied to evaluate their effectiveness in identifying at-risk students based on academic performance indicators. An initial exploratory data analysis was conducted to identify the most significant features contributing to student academic risk.

Features such as assignment completion rates, attendance consistency, mid-semester assessment scores, and student engagement metrics demonstrated strong associations with academic outcomes and were prioritized for model training.

The dataset was divided into training and testing subsets. Logistic Regression, Decision Tree,

Random Forest, SVM, KNN, and XGBoost models were trained and evaluated. Each model was assessed using accuracy, precision, recall, and F1-score on the held-out testing set.

Among all models evaluated, Random Forest and XGBoost achieved the highest classification accuracy, demonstrating their ability to capture complex, non-linear relationships within the student academic data. These ensemble methods also exhibited stronger resistance to overfitting compared to simpler models.

The analysis confirms that ensemble-based approaches are particularly well-suited for early student risk identification tasks and that feature quality plays a critical role in achieving reliable predictions.

IX. DISCUSSION

The results of this study validate the effectiveness of machine learning approaches for the early identification of at-risk students. By incorporating multiple academic performance indicators, the proposed system delivers a structured and objective basis for risk assessment that surpasses conventional manual review.

Ensemble methods such as Random Forest and XGBoost consistently outperformed other algorithms, attributed to their ability to aggregate multiple decision trees and model complex feature interactions while mitigating overfitting. In contrast, Logistic Regression and SVM showed comparatively lower accuracy, reflecting the non-linear nature of academic risk patterns.

The choice of predictive features proved critical to system performance. Assignment completion rates, attendance, and assessment scores were found to be highly discriminative indicators of academic risk. Rigorous data preprocessing further enhanced model reliability.

The continuous assessment capability of the ML-based approach represents a major practical advantage over periodic manual reviews. However, limitations remain: prediction quality is contingent on data completeness and accuracy, and institutional privacy considerations must be carefully managed when handling student data.

X. CONCLUSION

This paper presented an early identification system for at-risk students leveraging machine learning and performance analytics. The proposed system offers a more objective, consistent, and data-driven alternative to conventional academic monitoring methods, which are often retrospective and subjective.

The system evaluates students based on key academic indicators—assignment completion, attendance, assessment performance, and engagement—and classifies them into High-Risk, Medium-Risk, and Low-Risk categories using multiple supervised classification algorithms.

Experimental results demonstrated that ensemble methods, particularly Random Forest and XGBoost, achieved superior performance compared to other algorithms, highlighting the value of ensemble approaches for capturing complex academic risk patterns.

The proposed system equips academic institutions with a practical tool for proactive intervention, enabling counselors and faculty to direct support resources to students most in need. Continuous monitoring capability further enhances its utility for long-term academic planning.

Future work will explore the integration of real-time data, deep learning techniques, and a broader set of student attributes to further improve the accuracy and practical applicability of the early warning system.

XI. REFERENCES

[1] S. Gupta and A. Sharma, "Student Performance Prediction using Machine Learning Techniques," International Journal of Computer Applications, vol. 178, no. 32, pp. 20–25, 2021.

[2] R. Kumar and P. Singh, "Data Analytics in Educational Performance Management," IEEE International Conference on Data Science and Advanced Analytics, pp. 150–156, 2020.

[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, 2017.
[4] J. Smith and L. Brown, "Predicting At-Risk Students using Machine Learning Models," International Journal of Advanced Research in Computer Science, vol. 10, no. 3, pp. 45–50, 2019.
[5] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[6] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[7] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[8] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the ACM SIGKDD Conference, pp. 785–794, 2016.
[9] V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.
[10] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2019.
[11] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media, 2019.
[12] . K. P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012.

