© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1714882

Student Retention Risk Prediction Using Machine Learning And Educational Analytics

BHAVESH PRAKASH DALVI[1], MONISH GULATI[2], GAYATRI BAKHTIANI[3]

[1,2]*Department of Data Science Thakur Shyamnarayan Degree College*
[3]*Head of Department, Data Science, Thakur Shyamnarayan Degree College*

Abstract- ***Student retention is a major challenge faced by educational institutions worldwide. Early identification of students who are at risk of academic failure or dropout enables institutions to provide timely interventions and improve student success rates. This research proposes a machine learning-based system for predicting student academic risk using historical academic data. The system analyzes key academic attributes including attendance percentage, semester grade point averages, backlog count, and student participation in events. A Random Forest Regression model is used to estimate a risk probability score for each student. The predicted score is then used to classify students into low-risk, medium-risk, and high-risk categories. The system also includes an interactive dashboard that visualizes student risk patterns and highlights at-risk students for early intervention. Experimental results demonstrate that the proposed model achieves high predictive accuracy with strong generalization capability. The developed system provides a practical decision-support tool for educational institutions to monitor student performance and improve retention outcomes through data-driven strategies.***

Index Terms- ***Student Retention, Machine Learning, Random Forest, Educational Data Mining, Academic Analytics***

I. INTRODUCTION

Student dropout and academic underperformance are significant challenges faced by educational institutions. Identifying students who may face academic difficulties at an early stage allows institutions to provide academic support and improve retention rates.

Traditional methods of identifying at-risk students rely on manual academic monitoring and periodic performance evaluation. However, these methods often fail to detect early warning signs. With the increasing availability of educational data, machine learning techniques provide an effective approach for predicting student outcomes.

Machine learning models can analyze multiple academic attributes such as attendance, GPA trends, backlog history, and engagement indicators. By identifying patterns in historical student data, predictive models can estimate the likelihood of academic risk.

This research presents a Student Retention Risk Scoring System that uses machine learning algorithms to analyze academic performance data and generate risk predictions. The system also integrates an interactive dashboard to visualize student performance patterns and support decision-making for educators.

II. IDENTIFY, RESEARCH AND COLLECT IDEA

The concept of predicting student academic risk originates from the field of Educational Data Mining and Learning Analytics. Many researchers have explored the use of machine learning algorithms to analyze student performance and predict dropout probabilities.

Previous research has shown that academic indicators such as attendance percentage, grade point average, and backlog history are strong predictors of student success. Machine learning models such as Logistic Regression, Decision Trees, Support Vector Machines, and Random Forest algorithms have been used to analyze these factors.

The idea behind this research is to develop a predictive system that combines machine learning techniques with academic data visualization. By integrating predictive analytics with an interactive

dashboard, institutions can monitor student performance more effectively and identify students who require academic intervention.

III. WRITE DOWN YOUR STUDIES AND FINDINGS

The research methodology consists of several stages including data preprocessing, feature engineering, machine learning model training, and evaluation.

The dataset used in this study contains academic records of students including attributes such as attendance percentage, semester-wise GPA, backlog count, event participation, course, gender, year, and age.

Data preprocessing techniques are applied to clean and normalize the dataset. Missing values are handled using median imputation and categorical attributes are encoded using label encoding. Feature engineering is performed to compute additional attributes such as average GPA and event participation score.

A risk score formula is designed based on the influence of academic factors such as attendance, GPA, backlog count, and participation levels. This score is normalized to produce a probability value between 0 and 1.

The dataset is then divided into training and testing sets using an 80:20 split ratio.

IV. EXPERIMENTAL RESULTS

The predictive model used in this research is the Random Forest Regression algorithm. Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.

The model was trained using several academic attributes including attendance, average GPA, backlog count, event participation score, course, gender, academic year, and age.

Model performance was evaluated using statistical metrics including R^2 Score, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and cross-validation.

The experimental results are summarized below.

Metric	Value
Training R^2 Score	0.9965
Testing R^2 Score	0.9850
MAE	0.0100
RMSE	0.0159
Cross Validation Mean R^2	0.9817

The results indicate that the model achieves high prediction accuracy and strong generalization performance.

An interactive dashboard was also developed to visualize student risk distribution, attendance correlation, backlog impact, and feature importance.

V. FUTURE WORK

The proposed system can be further improved by integrating real-time academic data from institutional databases or ERP systems. Incorporating additional student attributes such as financial background, psychological factors, and engagement metrics may further enhance prediction accuracy.

Future research may also explore advanced machine learning models such as Gradient Boosting, XGBoost, or Deep Learning architectures for improved predictive performance.

Additionally, the system could include an automated recommendation module that suggests personalized academic interventions for students identified as high risk.

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1714882

VI. CONCLUSION

This research presented a machine learning-based system for predicting student academic risk using historical academic performance data. The proposed system uses a Random Forest Regression model to estimate a risk probability score for each student.

The integration of predictive modeling with an interactive dashboard enables educators to monitor student performance and identify at-risk students more effectively.

The experimental results demonstrate that the proposed model achieves high prediction accuracy and can serve as a valuable decision-support tool for improving student retention.

The developed system highlights the potential of machine learning and educational analytics in supporting data-driven decision making within educational institutions.

REFERENCES

[1] Breiman, L. Random Forests. Machine Learning Journal, 2001.

[2] Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research

[3] Han, J., Kamber, M., Pei, J. Data Mining: Concepts and Techniques.

[4] Kaggle Dataset: Student Performance Dataset.

[5] React Documentation – https://react.dev

[6] Flask Documentation – https://flask.palletsprojects.com