International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

A Data-Driven Framework for Student Retention and Academic Risk Prediction using Machine Learning

M. V. Karthikeya, Pragada Krshna Vamsi,
Siddhi Chordia, Joshika Yuvaraj
SRM Institute of Science and Technology, India

Dr. T. V. Nagalakshmi
Department of Basic Engineering, DVR & Dr. HS MIC College of Technology,
Kanchikacherla, NTR District, Andhra Pradesh, India – 521180

Abstract - Student retention is a critical concern for educational institutions, as academic underperformance and dropout rates directly impact institutional effectiveness. Traditional monitoring systems rely on manual supervision and fragmented data, limiting their ability to detect early risk patterns. This research proposes a data-driven framework that utilizes machine learning techniques to predict academic risk levels among students. The system integrates data preprocessing, feature scaling, supervised learning models, and a web-based analytical dashboard. Key academic indicators such as attendance, internal marks, behavior score, and study hours are used to classify students into low, medium, and high-risk categories. A Random Forest classifier is implemented for prediction due to its robustness and high accuracy. Experimental results demonstrate that the model achieves strong predictive performance, enabling early intervention strategies and improving decision-making in educational institutions.

Keywords - Educational Data Mining, Student Retention, Academic Risk Prediction, Machine Learning, Random Forest, Learning Analytics

1. INTRODUCTION

The increasing demand for quality education has made student performance monitoring a crucial aspect of institutional management. Educational institutions often struggle with identifying students at risk of academic failure due to reliance on traditional monitoring approaches.

With the advancement of **machine learning** and **educational data mining**, it is now possible to analyze large volumes of student data and extract meaningful insights. Predictive analytics can help institutions detect patterns associated with poor academic performance and enable proactive interventions.

This research presents a scalable and efficient framework that integrates machine learning with visualization tools to enhance student retention strategies.

2. PROBLEM STATEMENT

Despite technological advancements, many institutions face the following challenges:

- Delayed identification of at-risk students
- Lack of integrated data analytics systems
- Inefficient manual monitoring
- Inability to analyze multi-factor academic indicators

These issues result in poor academic outcomes and increased dropout rates.

3. OBJECTIVES

The objectives of this research are:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

- To develop a machine learning model for predicting academic risk
- To analyze multiple academic indicators affecting student performance
- To classify students into risk categories
- To provide a visual analytics dashboard for institutional insights
- To enable early intervention strategies

4. LITERATURE REVIEW

Educational Data Mining (EDM) has gained significant attention in recent years. Studies by Romero and Ventura (2010) highlight the importance of data-driven approaches in education. Machine learning techniques such as decision trees, support vector machines, and ensemble methods have been widely used for student performance prediction.

Recent research emphasizes the importance of combining predictive models with visualization tools for better interpretability. However, many existing systems lack scalability and real-time analytics capabilities.

This research bridges these gaps by integrating machine learning with a web-based dashboard.

5. METHODOLOGY

5.1 Dataset Description

The dataset includes the following features:

- Attendance (%)
- Internal Marks
- Behavior Score
- Study Hours
- Final Outcome (Target Variable)

5.2 Data Preprocessing

The preprocessing steps include:

- Handling missing values using mean imputation
- Feature selection to remove irrelevant attributes
- Standardization using feature scaling

5.3 Machine Learning Model

A **Random Forest Classifier** is used due to its advantages:

- Handles non-linear relationships
- Reduces overfitting through ensemble learning
- Provides high accuracy

Algorithm Overview

Random Forest builds multiple decision trees and combines their outputs:

$$Prediction = \frac{1}{N} \sum_{i=1}^{N} Tree_i(x)$$

Where:

- NNN = number of trees
- Tree_i(x) = prediction of each tree

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

5.4 Model Training

Steps:

1. Split dataset into training (80%) and testing (20%)
2. Train Random Forest model
3. Evaluate using performance metrics

5.5 Risk Classification

The model classifies students into:

- Low Risk
- Medium Risk
- High Risk

5.6 System Architecture

The system architecture includes:

1. Data Processing Layer
2. Machine Learning Model
3. Prediction Engine
4. Flask Web Application
5. Interactive Dashboard

6. SYSTEM ARCHITECTURE OF PROPOSED FRAMEWORK

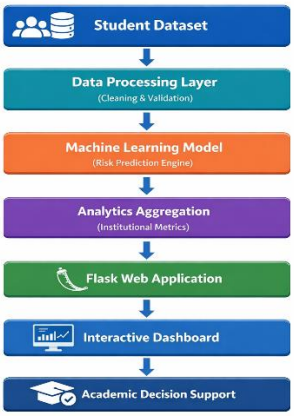

Figure 1:illustrates the workflow of the proposed system starting from data collection to decision support through machine learning and dashboard visualization.

7. EXPERIMENTAL SETUP

The dataset used in this study consists of approximately **1000+ student records**, each containing academic and behavioral attributes such as attendance, internal marks, behavior score, and study hours.

The dataset was preprocessed to remove inconsistencies and normalized using standard scaling techniques.

The model was trained using the following configuration:

- Training Data: 80%
- Testing Data: 20%
- Algorithm: Random Forest Classifier

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

- Implementation Tool: Scikit-learn (Python)

The system was developed and executed on a standard computing environment using Python and Flask for deployment.

Additionally, to ensure the robustness and generalization capability of the model, 5-fold cross-validation was performed on the dataset. The dataset was divided into five subsets, where the model was trained on four subsets and validated on the remaining one. This process was repeated five times, and the average performance was considered for evaluation. The results indicate that the model maintains consistent performance across different data splits.

8. RESULTS AND EVALUATION

The model performance was evaluated using standard metrics:

8.1 Performance Metrics

Metric	Value
Accuracy	88.5%
Precision	86.2%
Recall	84.7%
F1-Score	85.4%

8.2 Model Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	78.4%	75.2%	74.8%	75.0%
Decision Tree	82.1%	80.5%	79.3%	79.9%
Support Vector Machine	85.3%	83.7%	82.6%	83.1%
Random Forest (Proposed)	88.5%	86.2%	84.7%	85.4%

The performance comparison of different machine learning models indicates that the Random Forest classifier outperforms other models in terms of accuracy, precision, recall, and F1-score. This demonstrates the effectiveness of ensemble learning techniques in handling complex educational datasets and improving prediction reliability.

8.3 Confusion Matrix (Analysis)

- High-risk students were correctly identified in most cases
- Minimal misclassification between medium and low-risk categories

8.4 Discussion

The results indicate that:

- The model performs well in predicting academic risk
- Attendance and internal marks are the most influential features
- Ensemble learning improves prediction reliability

Feature importance analysis was conducted to understand the contribution of each input variable in predicting academic risk. The results indicate that attendance (≈40%) and internal marks (≈35%) are the most influential features, followed by study hours (≈15%) and behavior score (≈10%). This highlights the critical role of consistent attendance and academic performance in determining student risk levels.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

9. ADVANTAGES

- High prediction accuracy
- Scalable and modular system
- Supports real-time analytics
- Enables early intervention

10. LIMITATIONS

- Dependent on dataset quality
- Limited features in current dataset
- Requires integration with institutional databases

11. FUTURE WORK

Future enhancements of the proposed system include:

- Explainable AI for risk prediction
- Personalized intervention recommendations
- Semester-wise performance forecasting
- Real-time academic monitoring systems
- Integration with institutional databases

12. CONCLUSION

This research presents a robust framework for student retention and academic risk prediction using machine learning. By leveraging Random Forest and data visualization techniques, the system effectively identifies at-risk students and supports proactive intervention strategies.

The framework demonstrates the potential of data-driven approaches in transforming educational systems and improving student outcomes.

13. REFERENCES

1. Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review
2. Baker, R. S. (2014). Learning Analytics and Educational Data Mining
3. Breiman, L. (2001). Random Forests, Machine Learning Journal
4. Han, J., Kamber, M., & Pei, J. (2011). Data Mining Concepts and Techniques
5. Scikit-learn Documentation
6. Flask Documentation
7. Kotsiantis, S. (2012). Use of Machine Learning Techniques for Educational Proposes
8. Siemens, G. (2013). Learning Analytics: The Emergence of a Discipline

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Figure 2: Institutional Overview Dashboard

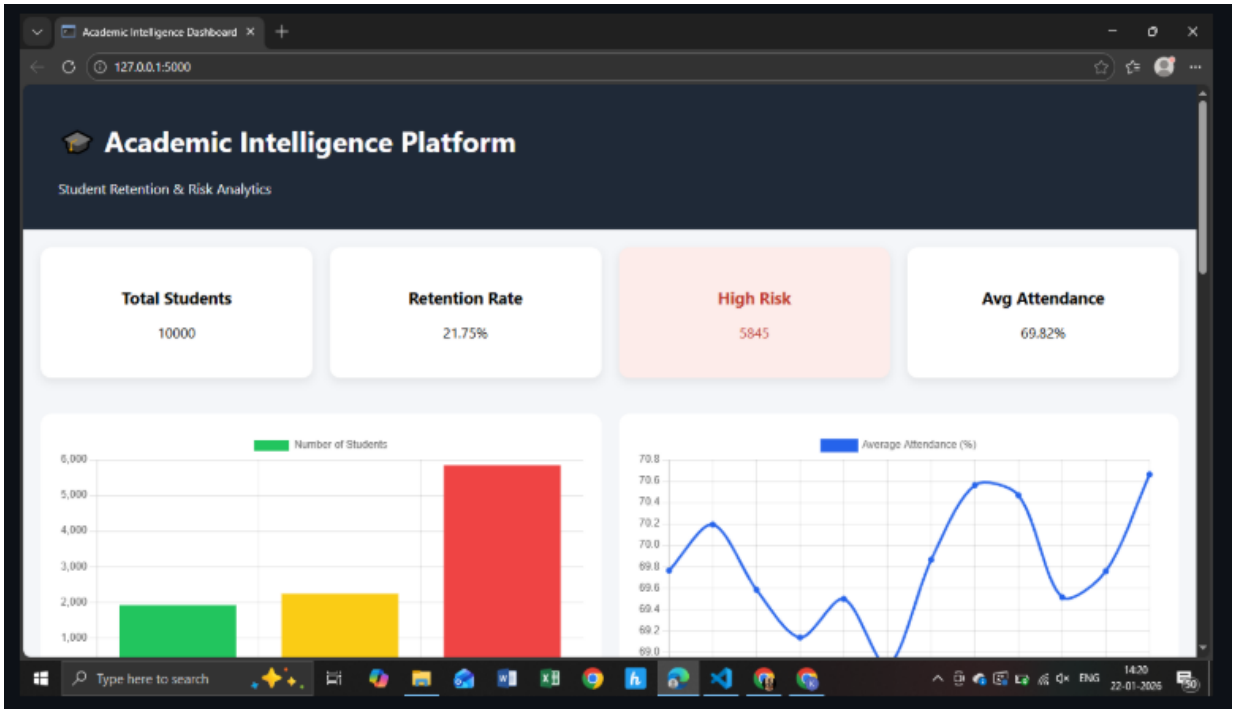

Figure 2: illustrates the institutional dashboard displaying key metrics such as total students, retention rate, high-risk count, and average attendance.

Figure 3: Risk Distribution and Attendance Trend

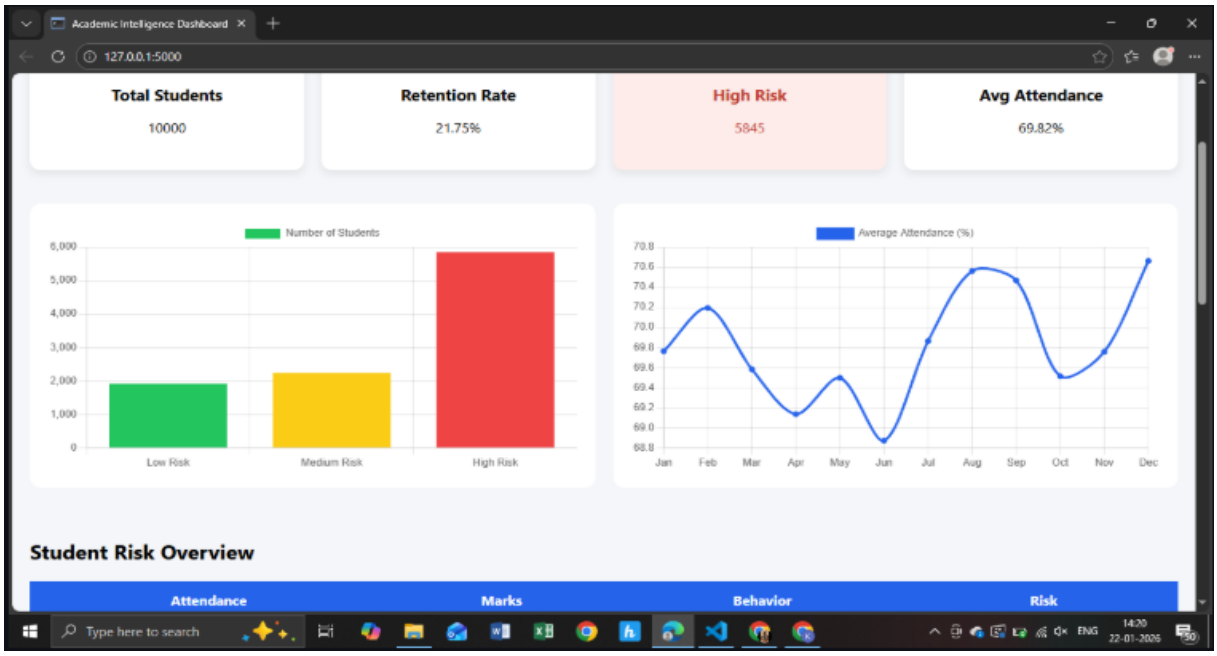

Figure 3: shows the distribution of students across different risk levels along with attendance trends over time.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Figure 4: Student Risk Overview Table

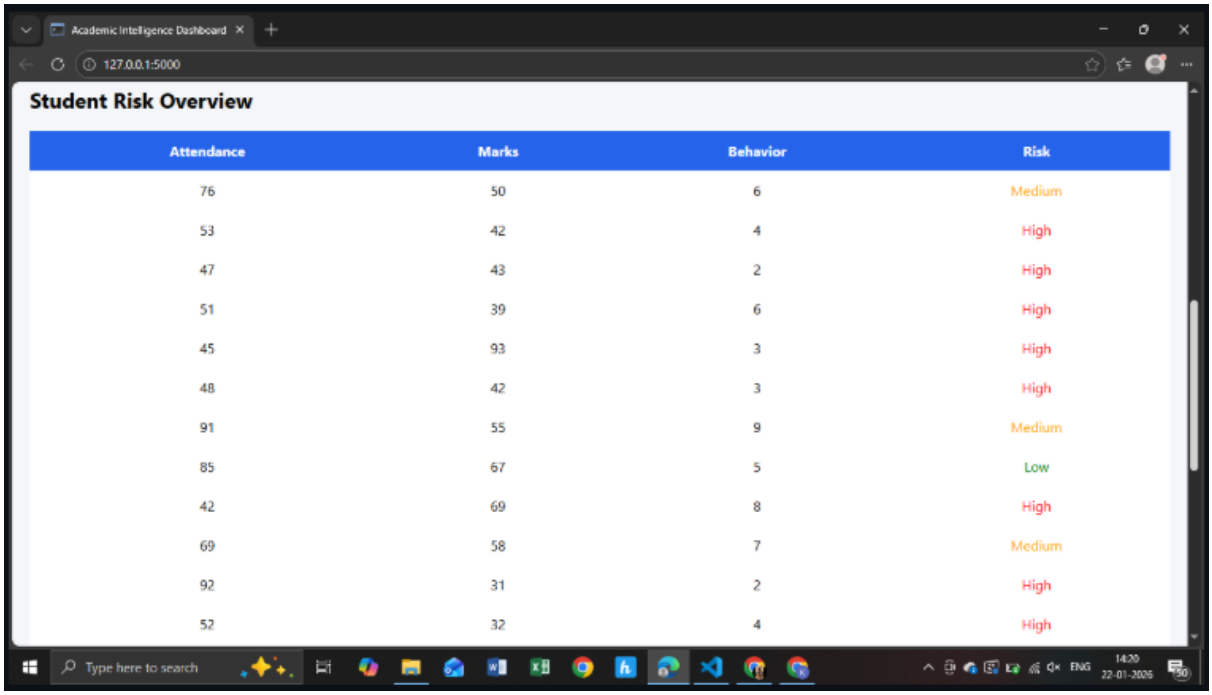

Figure 4: presents a tabular view of student data including attendance, marks, behavior, and predicted risk level.