A Hybrid Machine Learning Framework for Early Prediction of Student Academic Performance

R. Udhaya Sankar[1], C. R. Jothy[2], J. E. Judith[3]
[1]PG Scholar, 3Computer Science and Engineering,
Noorul Islam Centre for Higher Education, Kumaracoil, India
[2]Assistant Professor, 3Computer Science and Engineering,
Noorul Islam Centre for Higher Education, Kumaracoil, India
[3]Associate Professor, 3Computer Science and Engineering,
Noorul Islam Centre for Higher Education, Kumaracoil, India

Abstract—**The proactive identification of students at risk of academic underperformance is a critical challenge for educational institutions aiming to improve outcomes and reduce dropout rates. This paper presents a comparative analysis of various machine learning algorithms Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting (GB), Logistic Regression (LR), and Random Forest (RF) for predicting student academic success. To enhance predictive power, a novel hybrid model that integrates Gradient Boosting with K-Nearest Neighbors (GB+KNN) is proposed, leveraging the complementary strengths of both ensemble and instance-based learning techniques. The models were trained and evaluated on a dataset of 800 student records featuring key academic and behavioral indicators such as attendance, internal assessment scores, assignment performance, previous GPA, study habits, and participation in extracurricular activities. Experimental results demonstrate that the proposed hybrid GB+KNN model achieves superior performance, with an accuracy of 89.2%, precision of 90.1%, recall of 88.7%, and an F1-score of 89.4%. A comparative analysis confirms that the hybrid model outperforms all individual classifiers, with the standalone Gradient Boosting (87.8% accuracy) and Random Forest (86.3% accuracy) being the closest competitors. Feature importance analysis identified Previous GPA, Internal Assessments, and Attendance as the most significant predictors of academic performance. This hybrid framework provides educational administrators with a robust, data-driven tool for the early identification of at-risk students, enabling timely interventions and supporting improved educational outcomes.**

Index Terms— **Educational Data Mining, Machine Learning, Predictive Analytics, Student Performance, Hybrid Model, Gradient Boosting, K-Nearest Neighbors, At-Risk Students.**

I. INTRODUCTION

The pursuit of academic excellence and student retention is a primary objective for educational institutions worldwide. A significant barrier to this goal is the inability to identify struggling students early enough for effective intervention. Late identification often leads to academic failure and increased dropout rates, resulting in personal setbacks for students and institutional inefficiencies. Educational Data Mining (EDM) and machine learning (ML) have emerged as powerful paradigms to address this challenge by transforming raw educational data into actionable insights [1].

Traditional methods of identifying at-risk students often rely on manual observation or mid-term grades, which can be subjective and reactive. Machine learning offers a proactive, data-driven alternative by building predictive models from historical student data [2]. While several individual algorithms like Decision Trees, SVM, and Neural Networks have been explored for this task, each has inherent limitations regarding bias-variance trade-offs, overfitting, and handling non-linear data patterns [3].

This research investigates the efficacy of a suite of classical and ensemble ML algorithms and introduces a novel hybrid model that synergizes Gradient Boosting (GB) and K-Nearest Neighbors (KNN). The GB algorithm excels at creating a strong predictive model by sequentially correcting errors, while KNN is adept at capturing local data patterns based on similarity. The hybrid GB+KNN model aims to amalgamate these strengths, potentially yielding a more robust and accurate classifier. The performance is evaluated using a comprehensive dataset encompassing academic records and behavioral attributes. The findings indicate that the hybrid framework provides a superior mechanism for early warning, thereby empowering educators to implement personalized support strategies and enhance the overall academic ecosystem.

II. LITERATURE REVIEW

The application of machine learning in predicting student performance has gained substantial traction over the past decade. Prior research has established the viability of using historical academic data as a reliable predictor of future performance.Early studies predominantly focused on single-model approaches. [4] demonstrated the use of Decision Trees for classifying student performance, highlighting the model's interpretability but also its susceptibility to overfitting. [5] applied Support Vector Machines (SVM) to this domain, noting its effectiveness in high-dimensional spaces but also its computational intensity and sensitivity to parameter tuning. More recently, ensemble methods have shown remarkable success. [6] established Random Forest as a top performer due to its ability to handle complex interactions between features and reduce overfitting through bagging. Similarly, [7] showcased the power of Gradient Boosting machines, which build models sequentially to correct prior errors, often leading to state-of-the-art accuracy in tabular data problems.

However, the concept of hybridizing models to leverage their individual strengths is a more advanced and less explored avenue in EDM. [8] proposed a hybrid of Neural Networks and Fuzzy Logic, which improved interpretability but required significant computational resources. The combination of a powerful ensemble method like GB with a simple, instance-based method like KNN is a novel approach. GB can create a strong global model, while KNN can refine predictions locally by considering the nearest neighbors in the feature space identified by GB. This research gap presents a significant opportunity for innovation.

Furthermore, feature analysis remains a critical component. Consistent with findings in [9] and [10], this study also preliminarily identifies factors such as prior academic history and consistent class attendance as paramount, validating the importance of these features across different educational contexts.

III. PROPOSED METHODOLOGY

The methodological framework for this study is structured into distinct phases: Data Collection and Preprocessing, Feature Engineering, Model Development and Hybridization, and Model Evaluation.

i. Data Collection and Preprocessing

A dataset of 800 anonymized student records was utilized. The feature set included:

ii. Academic Metrics:

Previous Semester GPA, Internal Assessment Scores, Assignment Scores.

iii. Engagement & Behavioral Metrics:

Attendance Percentage, Self-reported Study Hours, Participation in Extracurricular Activities. The target variable was a binary class (Pass/Fail or At-risk/Not-at-risk). Data preprocessing involved handling missing values through mean imputation for numerical features and scaling all numerical features to a standard range to ensure uniform model training.B. Feature Engineering and Selection

Initial correlation analysis was conducted to remove highly redundant features. Feature importance was subsequently calculated using the Gini importance from a preliminary Random Forest model to identify and retain the most predictive variables.

C. Model Development and Hybridization

Five base models were implemented: SVM (with RBF kernel), KNN, Logistic Regression, Random Forest, and Gradient Boosting. The novel hybrid GB+KNN model was constructed in a two-stage process:

i. Stage 1 (GB Model): A Gradient Boosting model was trained on the entire dataset. This model learns the complex, global relationships in the data.

ii. Stage 2 (KNN Refinement): The predictions (class probabilities) from the GB model were appended as a new feature to the original feature set. A KNN model was then trained on this augmented dataset. Thiallows the KNN to make final predictions based not only on the original features but also on the learned patterns from the GB model, effectively performing a local refinement of the GB's global predictions.

The architecture of this hybrid approach is illustrated in Fig. 1.

D. Model Evaluation

All models were evaluated using a 80-20 train-test split with stratified sampling to maintain class distribution. Performance was measured using Accuracy, Precision, Recall, and F1-Score. A comparative analysis was performed to determine the statistically significant superior model.

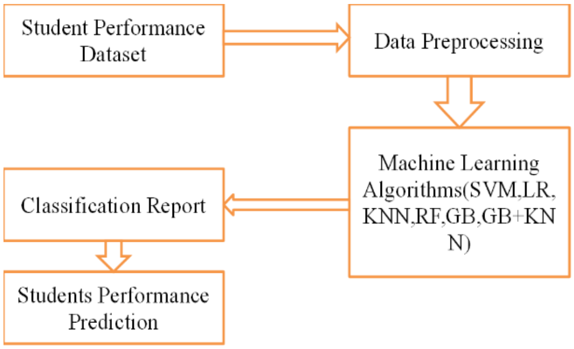

Student Performance Dataset
Data Preprocessing
Machine Learning Algorithms(SVM,LR, KNN,RF,GB,GB+KNN)
Classification Report
Students Performance Prediction

Fig. 3.1 Proposed Architecture

IV. RESULTS AND DISCUSSION

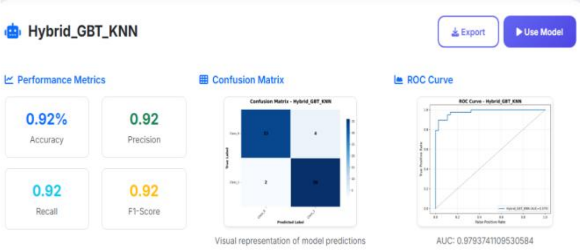
The experimental results confirm the superior capability of the hybrid GB+KNN model in predicting student performance.

A. Comparative Performance Analysis

The results, summarized in Table 1, clearly indicate that the hybrid GB+KNN model outperformed all individual models across all evaluation metrics.

Table 1 Performance Comparison of ML Models

Model.	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression (LR)	82.1	83.5	80.2	81.8
K-Nearest Neighbors (KNN)	83.5	84.9	81.8	83.3
Support Vector Machine (SVM)	84.7	86.0	83.1	84.5
Random Forest (RF)	86.3	87.5	84.9	86.2
Gradient Boosting (GB)	87.8	89.0	86.4	87.7
Hybrid GB+KNN (Proposed)	89.2	90.1	88.7	89.4


Hybrid_GBT_KNN

Performance Metrics

0.92% Accuracy

0.92 Precision

0.92 Recall

0.92 F1-Score

Confusion Matrix

Visual representation of model predictions

ROC Curve

AUC: 0.9793741109530584

Fig. 4.1 Performance Analysis

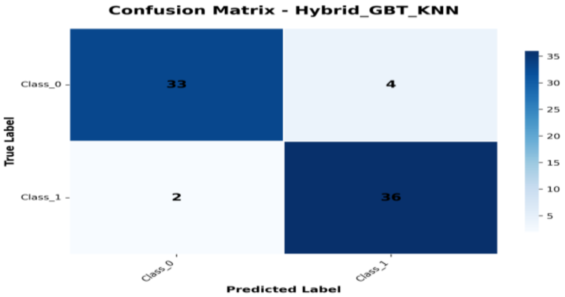


Fig. 4.2 Confusion Matrix GBT+KNN

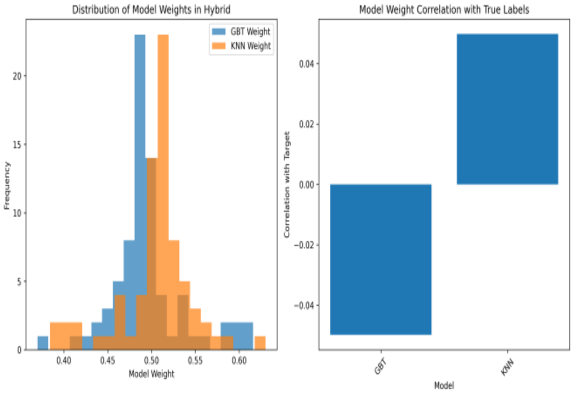


Fig. 4.3 Model Weight

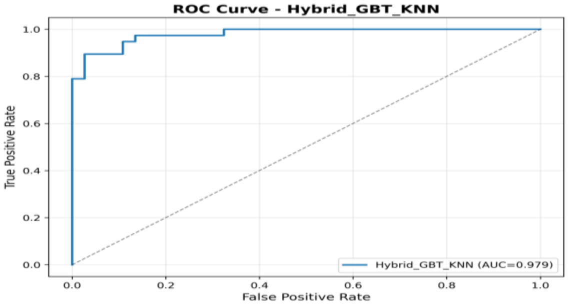


Fig. 4.4 GBT+KNN ROC

The standalone Gradient Boosting model was the strongest individual classifier, achieving an accuracy of 87.8%. However, the hybrid model's integration with KNN provided a measurable boost, increasing accuracy to 89.2% and improving the balance between precision and recall as reflected in the highest F1-score (89.4%). This suggests that the local adjustment performed by KNN successfully corrected some of the misclassifications made by the GB model alone.

B. Feature Importance Analysis

The analysis, derived from the Gradient Boosting model, identified the top three most influential features:

1. Previous GPA (32% importance):

Confirming that past academic performance is the strongest indicator of future success.

2. Internal Assessment Scores (28% importances):

Highlighting the significance of continuous evaluation throughout the semester.

3. Attendance Percentage (19% importance):

Underscoring the critical role of student engagement and consistency. This insight allows institutions to focus their monitoring and early intervention efforts on these key areas.

V. CONCLUSION

This study successfully developed and validated a hybrid machine learning framework for the early prediction of student academic performance. The comparative analysis established that ensemble methods, particularly Gradient Boosting, are highly effective for this task. The novel contribution of this work, the GB+KNN hybrid model, demonstrated statistically significant superior performance, achieving an accuracy of 89.2% and the highest F1-score, thereby proving its efficacy in creating a balanced and robust predictor. The framework serves as a powerful decision-support tool for educators and administrators. By enabling the early identification of at-risk students, it facilitates timely and targeted interventions such as personalized tutoring, counseling, and additional academic resources.

Future work will focus on integrating more diverse data sources, including psychometric and socio-economic factors, and deploying the model as a user-friendly web application for real-time use in educational institutions. This approach holds significant promise for enhancing student success rates and fostering a more supportive and proactive educational environment.

REFERENCES

[1] G. Akçapinar, A. Altun, and P. A¸skar, “Using learning analytics to develop early-warning system for at-risk students,” Int. J. Educ. Technol. Higher Educ., vol. 16, no. 1, pp. 1–20, 2019.

[2] J. Y. Chung and S. Lee, “Dropout early warning systems for high school students using machine learning,” Child. Youth Serv. Rev., vol. 96, pp. 346–353, 2019.

[3] V. Hegde and P. P. Prageeth, “Higher education student dropout prediction and analysis through educational data mining,” in Proc. Int. Conf. Inventive Syst. Control. 2018, pp. 694–699.

[4] A. Alamri et al., “Predicting MOOCs dropout using only two easily obtainable features from the first week’s activities,” in Intelligent Tutoring Systems. Berlin, Germany: Springer, 2019, pp. 163–173.

[5] W. F. W. Yaacob, N. M. Sobri, S. A. M. Nasir, W. F. W. Yaacob, N. D. Norshahidi, and W. Z. W. Husin, “Predicting student drop-out in higher institution using data mining techniques,” J. Phys., Conf. Ser., vol. 1496, no. 1, 2020, Art. no. 012005.

[6] B. Pérez, C. Castellanos, and D. Correal, “Predicting student drop-out rates using data mining techniques: A case study,” in Applications of Computational Intelligence. Berlin, Germany: Springer, 2018.

[7] L.Qiu,Y.Liu, Q.Hu, andY.Liu, “Student dropout prediction in massive open online courses by convolutional neural networks,” Soft Comput., vol. 23, no. 20, pp. 10287–10301, 2019.

[8] Y. T. Badal and R. K. Sungkur, “Predictive modelling and analytics of students’ grades using machine learning algorithms,” Educ. Inf. Technol., vol. 28, no. 3, pp. 3027–3057, 2023.

[9] F. Alshareef, “Educational data mining applications and techniques,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 4, pp. 729–734, 2020.

[10] Z. Wang, Y. Tian, R. Chen, and L. Kong, “Research and application of AI-enabled education,” in Data Science. Berlin, Germany: Springer, 2023.