A Comprehensive Machine Learning Framework for Predicting Student Academic Performance Using Feature Selection and Multi-Model Evaluation

Joud Alsuhaibani,Reema Alelaiyt,Maysam Alhuthail,Sana marran
Department of Computer Science
Princesss Nourah bint Abdularhaman University

***Abstract*—Predicting student academic performance is a central problem in educational data mining, with applications in early intervention systems, personalized learning, and academic planning. This study presents a robust machine learning framework using a dataset of 6,605 student records, encompassing academic, behavioral, environmental, socioeconomic, and psychological factors. The proposed workflow includes preprocessing, outlier detection, categorical encoding, grade categorization, model training, hyperparameter optimization, and interpretability analysis using feature importance. Five classifiers—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—were trained on full features. Feature importance was computed via Random Forest and permutation importance, and the top 10 features were used to build reduced models. Results show SVM achieved the highest performance on full features (Accuracy = 96.24%, F1 = 0.9584). After feature reduction, SVM and Logistic Regression maintained strong performance around 86%. Key predictors include Attendance, Hours Studied, Previous Scores, Tutoring Sessions, Sleep Hours, Parental Involvement, Access to Resources, Motivation Level, Family Income, and Peer Influence. These findings demonstrate the critical role of academic habits and environmental factors, and the importance of interpretable machine learning in educational analytics.**

I. Introduction

Educational data mining (EDM) has become a transformative discipline that enables institutions to leverage rich datasets to improve student outcomes. The rise of digital learning platforms produces vast data capturing academic performance, behavioral trends, and environmental conditions. Machine learning (ML) provides the tools to analyze these multidimensional datasets and predict student performance effectively.

Accurate prediction of academic outcomes supports early intervention and personalized learning, but challenges include missing data, heterogeneous features, and nonlinear interactions. This work addresses these challenges through a multi-stage ML pipeline, incorporating feature selection, model comparison, and interpretability analysis.

This study seeks to answer the following questions:

- Which ML models most accurately predict student grade categories?
- Which features are the strongest predictors of exam performance?
- How does model performance change after reducing to the top predictors?

The remainder of this paper is organized as follows: Section II reviews related work. Section III details the dataset and preprocessing. Section IV presents the methodology. Section V discusses feature importance. Section VI shows experimental results. Section VII provides a discussion, and Section VIII concludes.

II. Related Work

Prior research consistently demonstrates the superiority of ML models over traditional statistical techniques in predicting academic outcomes. Behavioral metrics such as study hours, attendance, and previous performance are strong predictors, along with environmental variables like parental involvement and access to resources [1], [2]. Feature selection methods, particularly ensemble-based techniques like Random Forest importance, improve interpretability and reduce overfitting, enabling more actionable insights for educators [6], [7].

III. Dataset and Preprocessing

A. Dataset Overview

The dataset contains 6,605 records with variables from academic, behavioral, environmental, socioeconomic, and psychological domains. The target variable is the student exam score, later mapped to categorical grades.

B. Handling Missing Values

Three columns had missing values: *Teacher Quality*, *Parental Education Level*, and *Distance from Home*. Mode imputation filled these missing entries. Remaining missing rows were dropped.

C. Outlier Detection

Outliers in numerical features (*Sleep Hours, Attendance, Previous Scores*) were identified via Z-score filtering ($|Z| > 3$) and removed to reduce bias.

D. Text Cleaning

Categorical text data were stripped of whitespace and converted to lowercase to maintain consistency.

E. Grade Categorization

Exam scores were binned into nine grade categories: F, D, D+, C, C+, B, B+, A, A+, using increments of 5–10 points.

F. Encoding and Scaling

Categorical features were one-hot encoded. Numerical features were standardized:

$$x' = \frac{x - \mu}{\sigma}$$

G. Train-Test Split

A stratified 75/25 split preserved the distribution of grade categories.

IV. Methodology

A. Machine Learning Models

Five classifiers were used:

- **Logistic Regression**: Optimizes log-loss.
- **Support Vector Machine**: Maximizes the class separation margin.
- **Decision Tree**: Splits features to minimize Gini impurity.
- **Random Forest**: Ensemble of decorrelated decision trees.
- **K-Nearest Neighbors**: Predicts based on proximity in feature space.

B. Hyperparameter Tuning

GridSearchCV with 5-fold StratifiedKFold cross-validation tuned hyperparameters, selecting optimal values for each model.

C. Evaluation Metrics

Models were evaluated using Accuracy, Precision, Recall, and F1-score.

V. Feature Importance Analysis

Feature importance was computed using:

- **Random Forest Importance**: Mean decrease in Gini impurity.
- **Permutation Importance**: Change in weighted F1-score when shuffling feature values.

One-hot encoded features were aggregated to their original variables. The top 10 features were:

1) Attendance
2) Hours Studied
3) Previous Scores
4) Tutoring Sessions
5) Sleep Hours
6) Parental Involvement
7) Access to Resources
8) Motivation Level
9) Family Income
10) Peer Influence

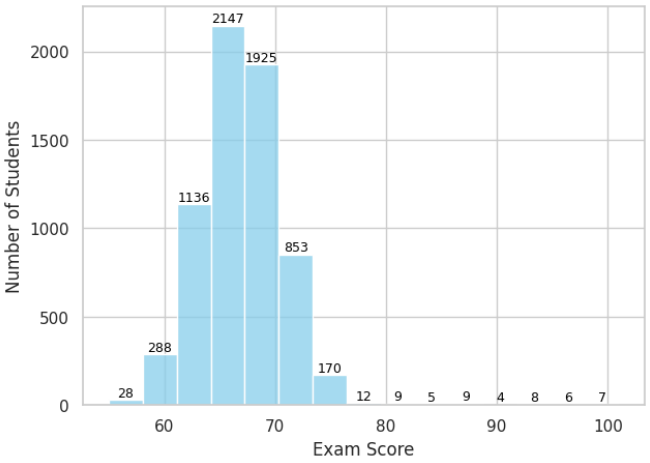

Fig. 1: Histogram showing the distribution of student exam scores. Most students scored between 70 and 90.

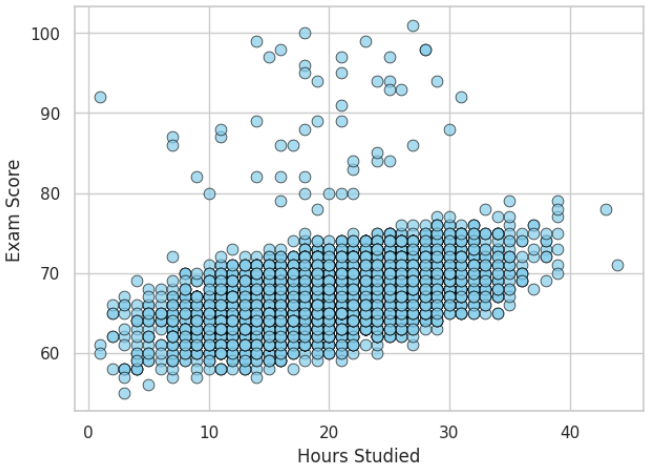

Fig. 2: Scatter plot of Hours Studied vs Exam Score. Positive correlation indicates increased study hours are associated with higher scores.

TABLE I: Performance of tuned models on full dataset

Model	Accuracy	Precision	Recall	F1-score
SVM	**0.962**	0.956	0.962	**0.958**
Logistic Regression	0.915	0.898	0.915	0.906
Random Forest	0.788	0.778	0.788	0.769
Decision Tree	0.623	0.610	0.623	0.615
KNN	0.581	0.570	0.581	0.575

TABLE II: Performance using only top 10 features

Model	Accuracy	F1-score
Logistic Regression	**0.862**	0.858
SVM	0.858	0.855
Random Forest	0.803	0.789

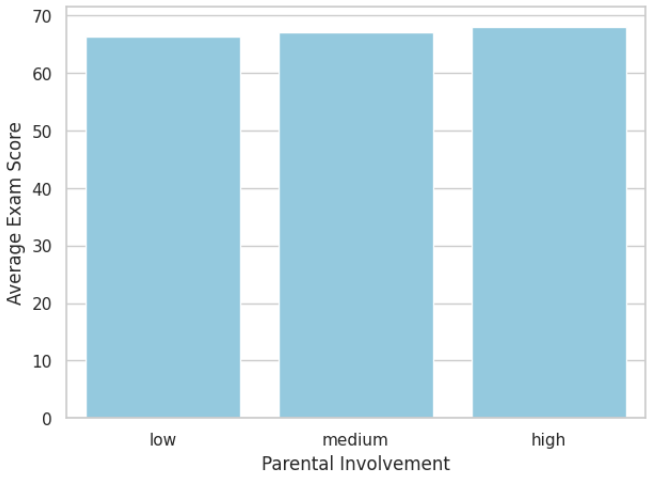

Fig. 3: Average exam score per level of parental involvement. Higher involvement correlates with better performance.

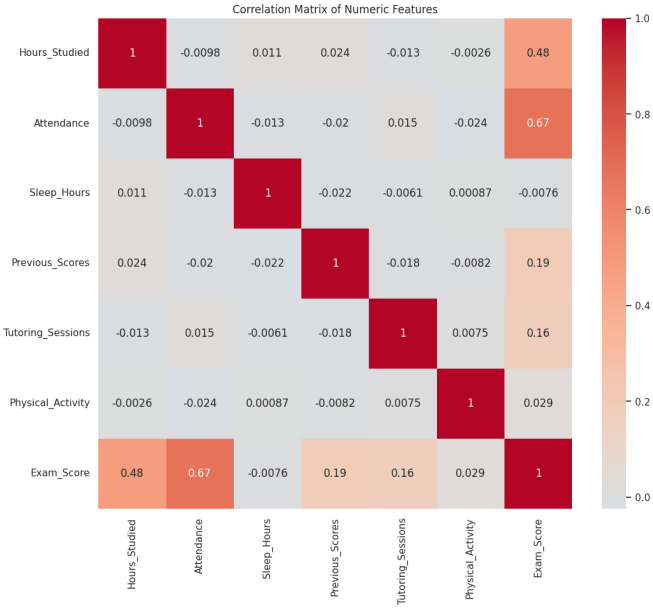

Fig. 4: Correlation heatmap of numerical variables. Strong correlations exist between Attendance, Hours Studied, and Exam Score.

VI. Experimental Results

A. Exploratory Figures

B. Model Performance (Full Features)

C. Model Performance (Top 10 Features Only)

VII. Discussion

SVM achieved the highest accuracy on the full feature set. Feature reduction to the top 10 predictors caused only a slight drop in performance. Attendance, Hours Studied, and Previous Scores were the most influential features. Environmental and socioeconomic factors (Parental Involvement, Access to Resources, Family Income) also contributed significantly. These

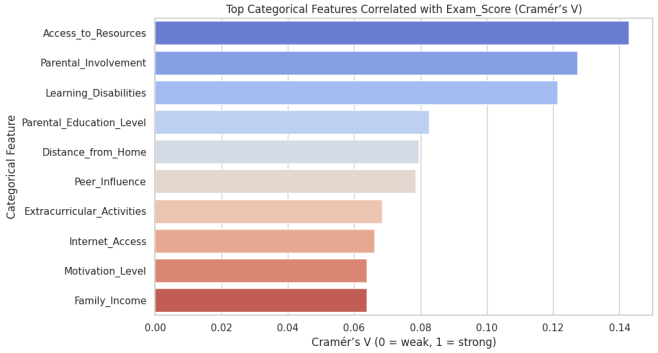

Fig. 5: Categorical feature correlation ,Access to resources appeared to be the highest among all .

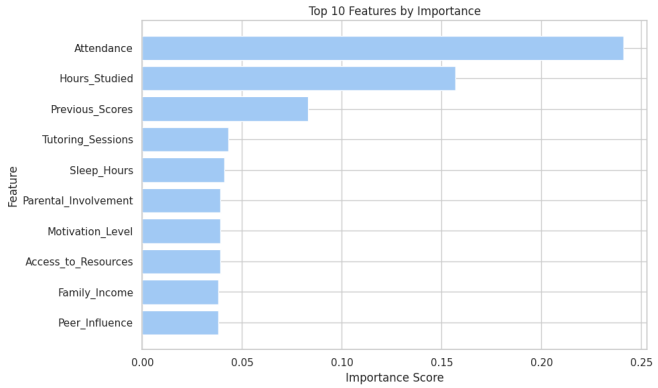

Fig. 6: Horizontal bar plot of top 10 features by aggregated importance. Attendance, Hours Studied, and Previous Scores dominate the predictive signal.

insights suggest that interventions targeting study habits and supportive environments may improve student outcomes.

VIII. Conclusion

This research presents a complete ML pipeline for predicting student academic performance. SVM outperformed other classifiers on full features, and reduced models maintained high accuracy, demonstrating the effectiveness of feature selection. Combining Random Forest and permutation importance ensures interpretability, guiding educators in identifying the most impactful factors. Future work may include deep learning approaches, longitudinal modeling, and real-time prediction.

IX. References

References

[1] A. Kotsiantis, et al., "Predicting Students' Grades Using Machine Learning Techniques," *Educational Technology & Society*, 2007.

[2] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE Transactions on Systems, Man, and Cybernetics (SMC)*, 2013.

[3] I. Goodfellow, et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[4] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2014.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.
[6] M. Yağcı, "Educational data mining: Prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, p. 11, 2022.
[7] D. Khairy, N. Alharbi, and M. A. Amasha, "Prediction of student exam performance using data mining classification algorithms," *Education and Information Technologies*, 2024.
[8] Y. Jedidi, A. Ibriz, and M. Benslimane, "Predicting students' academic performance and modeling using data mining techniques," *Mathematical Modeling and Computing*, vol. 11, no. 3, pp. 814–825, 2024.
[9] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and predicting students' academic performance using data mining techniques," *International Journal of Modern Education and Computer Science (IJMECS)*, vol. 8, no. 11, pp. 36–42, 2016.
[10] Y. Zhang, et al., "Educational data mining techniques for student performance prediction: Method review and comparison analysis," *Frontiers in Psychology*, 2022.
[11] N. Alshanqiti, "Predicting student performance with data mining and learning analytics techniques: A systematic literature review," *The American Journal of Applied Sciences*, vol. 5, no. 6, pp. 5–8, 2023.
[12] K. S. Al Shibli and A. S. S. Al Abri, "Model for prediction of student grades using data mining algorithms," *European Journal of Information Technologies and Computer Science*, 2022.
[13] A. Alhassan, "Data mining approach to predict success of secondary school students: A Saudi Arabian case study," *Education Sciences*, vol. 13, no. 3, p. 293, 2020.
[14] M. N. Injadat, et al., "Systematic ensemble model selection approach for educational data mining," *arXiv preprint arXiv:2005.06647*, 2020.