Published in the Proceedings of the 2025 Midwest Instruction and Computing Symposium (57th MICS), April 2025

Predicting Student Academic Performance Using Machine Learning and Clustering

Asmita Pun, Breanna Olson, and Muhammad Abusaqer

Department of Mathematics and Computer Science

Minot State University

Minot, North Dakota 58703

asmita.pun@minotstateu.edu, breanna.olson.4@minotstateu.edu and

muhammad.abusaqer@minotstateu.edu

Abstract

Student academic performance is determined by a complex interplay of demographic, academic, and environmental factors. This study utilizes a dataset of 5,000 records obtained from Kaggle, comprising 23 attributes that include attendance, midterm scores, assignment and quiz averages, participation, project evaluations, study hours, extracurricular activities, internet access, and parental education levels. A rigorous data cleaning process was implemented to address missing values, correct letter grade discrepancies using an established grading scheme, and resolve name–gender mismatches, while eliminating features that could introduce data leakage. Unsupervised learning methods, including K-Means and Hierarchical Clustering, were applied to explore natural groupings within the data, revealing overlapping clusters that suggest academic performance exists on a continuum rather than in discrete categories. Subsequently, supervised classification models—Random Forest, Logistic Regression, XGBoost, and Support Vector Machine—were employed to predict final letter grades in a multi-class setting. Despite the application of stratified sampling and SMOTE to address class imbalance, classification accuracies ranged from 19% to 21%, underscoring the challenges inherent in multi-class grade prediction which represents the overall student performance. Feature importance analyses, based on both Gini impurity and gain metrics, consistently identified key predictors such as project score, midterm Score, quizzes Avg, parents' education level, the existence of Internet access at home, and attendance. These findings contribute to the literature by demonstrating that although certain academic and engagement metrics are influential, additional variables and advanced modeling techniques may be necessary to improve predictive performance in educational outcomes.

1 Introduction

During 2021-2022, U.S. saw a 7-percentage point increase in the average adjusted cohort graduation rate (ACGR), reaching 87% for high school students [1]. Almost every job and higher education requires a high school degree. Quality education is associated with good health and economic growth as well. [12] Therefore, student academic success is a critical area of study in educational research, with significant implications for educators, institutions, and policymakers. Academic performance is influenced by numerous factors, including demographic characteristics, academic behaviors, and environmental conditions. Variables such as gender, age, department, attendance percentage, midterm and final scores, assignment and quiz averages, participation scores, and project evaluations contribute to overall academic success. Additionally, external factors such as extracurricular activities, weekly study hours, home internet access, and parental education levels further impact student achievement.

The increasing availability of educational data has enabled researchers to explore new ways to analyze and predict student outcomes using machine learning techniques. Machine learning is a subset of artificial intelligence that enables systems to learn patterns from data and make predictions without explicit programming. The ability to extract meaningful insights from large datasets allows for improved decision-making and early interventions.

Traditional statistical methods have long been employed to study academic performance, providing valuable insights through regression analyses, hypothesis testing, and correlation studies. Advanced clustering techniques, such as K-Means and Hierarchical Clustering, enable the exploitation of large datasets by identifying natural groupings among students. This approach facilitates the categorization of students into distinct clusters—such as high achievers, struggling students, and inconsistent performers—thereby revealing underlying patterns that conventional methods may overlook.

In addition to clustering, supervised learning models—including Random Forest, Support Vector Machines (SVM), Logistic Regression, and XGBoost—are utilized to predict final student grades. These models leverage a comprehensive set of features extracted from the dataset, offering a broader perspective on academic performance. Their effectiveness is evaluated using metrics such as accuracy, precision, recall, and F1-score. Furthermore, feature importance analysis, based on measures like Gini impurity and gain, identifies the most influential predictors, thereby informing targeted interventions within educational institutions.

The findings of this study have the potential to support educational institutions in developing data-driven strategies that enhance student success, improve retention rates, and ultimately increase graduation rates. By combining traditional statistical approaches with advanced machine learning techniques, this research seeks to provide a comprehensive understanding of the determinants of academic performance.

2 Related Work

Numerous studies have explored the application of machine learning in education, particularly in predicting student academic performance. Research in this domain has evolved from early statistical methods to advanced machine learning approaches, which offer better predictive accuracy and interpretability.

Kotsiantis et al. [2] provided a foundational review of machine learning applications in education, discussing classification techniques such as decision trees, neural networks, and Bayesian classifiers. More recent studies, such as Sweeney et al. [3], systematically reviewed the effectiveness of machine learning models in predicting student success across different educational contexts. These reviews have highlighted the growing trend of using ensemble learning methods and deep learning models for student performance prediction.

Marbouti et al. [4] developed predictive models that focus on feature selection and model comparison, emphasizing the importance of identifying key determinants of student success. Shahiri et al. [5] explored the application of supervised learning techniques, such as Support Vector Machines (SVM) and Random Forest, in academic performance prediction. Similarly, Al-Radaideh et al. [6] reviewed the role of data mining and clustering in identifying student learning patterns.

Cortez and Silva [7] compared multiple machine learning models using real-world student data, highlighting the predictive capabilities of ensemble methods. Dekker et al. [8] specifically investigated dropout prediction, identifying key factors affecting student retention. Al-Badawi et al. [9] proposed a hybrid machine learning model to enhance prediction accuracy by combining multiple algorithms.

Recent advances in deep learning have also been applied to student performance prediction. Hussain et al. [13] explored the effectiveness of deep neural networks (DNNs) in predicting academic outcomes, demonstrating improved accuracy over traditional machine learning models. Similarly, Khan et al. [14] used convolutional neural networks (CNNs) to analyze behavioral data such as class participation and online learning engagement, showing promising results.

The impact of online learning environments on student performance has also been a growing area of research. Sun et al. [15] investigated student engagement patterns in virtual learning environments, demonstrating how real-time learning analytics can improve student support systems. Moreover, Nguyen et al. [16] examined how student interaction in discussion forums and learning management systems can be leveraged to predict academic performance.

Several studies have also emphasized the importance of explainability in machine learning models for educational settings. Ribeiro et al. [17] introduced interpretable models that allow educators to understand why specific predictions are made, thereby facilitating targeted interventions. In addition, Zhang et al. [18] explored the role of SHAP (Shapley Additive Explanations) values in assessing the contribution of various academic and behavioral features to student success predictions.

Another crucial area of research is fairness and bias in predictive models. Binns et al. [19] investigated how demographic factors such as gender and socioeconomic background impact machine learning predictions, calling for fairness-aware algorithms in educational data mining. Similarly, Li et al. [20] studied algorithmic bias in student success models and proposed debiasing techniques to ensure equitable outcomes.

Overall, these studies provide a strong foundation for applying machine learning in educational research, supporting the methodological choices in this study. The evolution from traditional statistical approaches to advanced AI-driven models highlights the increasing role of data-driven decision-making in education. The integration of clustering techniques with supervised learning, as applied in this research, aligns with recent trends in leveraging both exploratory and predictive analytics to understand student performance better.

3. Methodology

3.1 Dataset

A dataset of 5,000 student performance records was obtained from Kaggle[1], containing 23 attributes that cover demographic, academic, and behavioral factors (see Table 1). Initial exploration indicated missing values in Attendance (%), Assignments_Avg, and Parent_Education_Level.

Column Name	Data Type	Description
Student_ID	String	Unique identifier for each student
First_Name	String	Student’s first name
Last_Name	String	Student’s last name
Email	String	Contact email (can be anonymized)
Gender	Categorical	Male, Female, Other
Age	Integer	Age of the student

[1] https://www.kaggle.com/datasets/mahmoudelhemaly/students-grading-dataset/data

Department	String	Student's department (e.g., CS, Engineering, Business)
Attendance (%)	Float	Attendance percentage (0-100%)
Midterm_Score	Float	Midterm exam score (out of 100)
Final_Score	Float	Final exam score (out of 100)
Assignments_Avg	Float	Average score of all assignments (out of 100)
Quizzes_Avg	Float	Average quiz scores (out of 100)
Participation_Score	Float	Score based on class participation (0-10)
Projects_Score	Float	Project evaluation score (out of 100)
Study_Hours_per_Week	Float	Average study hours per week
Extracurricular_Activities	Boolean	Whether the student participates in extracurriculars (Yes/No)
Internet_Access_at_Home	Boolean	Does the student have access to the internet at home? (Yes/No)
Parent_Education_Level	Categorical	Highest education level of parents (None, High School, Bachelor's, Master's, PhD)
Family_Income_Level	Categorical	Low, Medium, High
Stress_Level (1-10)	Integer	Self-reported stress level (1: Low, 10: High)
Sleep_Hours_per_Night	Float	Average hours of sleep per night
Total_Score	Float	Weighted sum of all grades
Grade	Categorical	Letter grade (A, B, C, D, F)

Table 1: Overview of the dataset structure (column names, data types, and Description)

3.2 Data Cleaning

Initial examination revealed missing values in three columns: Attendance (%), Assignments_Avg, and Parent_Education_Level. The numeric features were imputed using the median to reduce the influence of outliers whereas the categorical feature was imputed using the mode to preserve the most common category. This approach preserved the full dataset without discarding any records.

In addition, data quality was enhanced by addressing two issues:

- **Inconsistent Letter Grades:** The study used the common grading scheme (A: 90–100, B: 80–90, C: 70–80, D: 60–70, F: <60) to infer the expected letter grade from the 'Total Score'. Records where the actual Grade did not match the inferred grade were identified, and corrections were applied.
- **Name–Gender Mismatches:** A domain-specific name–gender mapping was used to correct inconsistencies (e.g., ensuring that names such as "John" are mapped to "Male").

To prevent data leakage and to focus on predictors that contribute to grade prediction, columns containing direct proxies for the outcome (Total_Score) as well as personal identifiers (Student_ID, First_Name, Last_Name, Email) were removed.

3.3 Data Standardization

Since the selected features have different numeric scales (e.g., scores out of 100 vs. hours per week), the study applied standardization using the Standard Scaler. This transformation rescales each feature to have mean = 0 and standard deviation = 1, ensuring no single feature dominates the distance measure.

3.4 Clustering Experiments

Two clustering methods were employed to explore natural groupings within the dataset based on eight standardized academic and engagement features: Attendance (%), Midterm_Score, Assignments_Avg, Quizzes_Avg, Participation_Score, Projects_Score, and Study_Hours_per_Week.

These attributes capture core aspects of academic progress (exams, quizzes, assignments, projects), engagement (attendance, participation), and study habits (weekly study hours).

3.4.1 K-Means Clustering

Clustering was first conducted using the K-Means algorithm. This algorithm partitions the dataset into K clusters by iteratively assigning points to the nearest cluster centroid and updating centroid positions until convergence. The K-Means algorithm was applied with k=3, a choice motivated by common educational categorizations (e.g., high, moderate, and low performers). Each student in the dataset was then labeled with a cluster identifier (0, 1, or 2).

Table 2 summarizes the mean values of the eight selected features for each of the three K-Means clusters. Several noteworthy patterns emerged.

KMeans cluster	Attendance (%)	Midterm Score	Final Score	Assignments Avg	Quizzes Avg	Participation Score	Projects Score	Study Hours per Week
0	73.77	71.25	67.77	75.63	75.40	4.64	89.91	17.63
1	78.07	68.67	54.35	72.61	73.93	5.04	65.18	17.50
2	74.93	70.84	86.15	75.914	75.27	5.32	67.06	17.85

Table 2: Mean feature values per K-Means cluster (0, 1, 2)

3.4.2 Hierarchical Clustering

Agglomerative Hierarchical Clustering was applied to the same standardized dataset. In this bottom-up approach, individual records were merged iteratively until three clusters remained. Table 3 presents the average values for the clustering features across the three hierarchical clusters, and the corresponding PCA-reduced visualization, which is shown in Figure 3, reveals subtle differences in study hours and exam scores compared to the K-Means solution.

Hierarchical cluster	Attendance (%)	Midterm Score	Final Score	Assignments Avg	Quizzes Avg	Participation Score	Projects Score	Study Hours per Week
0	75.88	72.89	68.33	74.58	75.47	5.41	73.57	12.82
1	71.34	57.09	75.81	71.81	77.36	4.56	75.12	20.38
2	78.58	78.52	66.02	77.95	71.71	4.68	76.95	22.99

Table 3: Average Feature Values per K-Means Cluster (K=3).

3.4.3 Determinate the optimal number of clusters

The optimal number of clusters was further evaluated using the Elbow Method and Silhouette Scores as shown in Figures 1 and 2. Although an initial choice of k=3 was used for interpretive convenience, the "optimal" number of clusters was investigated using two common metrics:

1. **Elbow Method**: The Within-Cluster Sum of Squares (WCSS) was plotted against values of K from 2 to 10. A point at which the rate of decrease in WCSS notably slows is often considered a suitable choice.
2. **Silhouette Score**: The average silhouette score was computed for each k in the same range, indicating how well-separated the clusters are in the feature space.

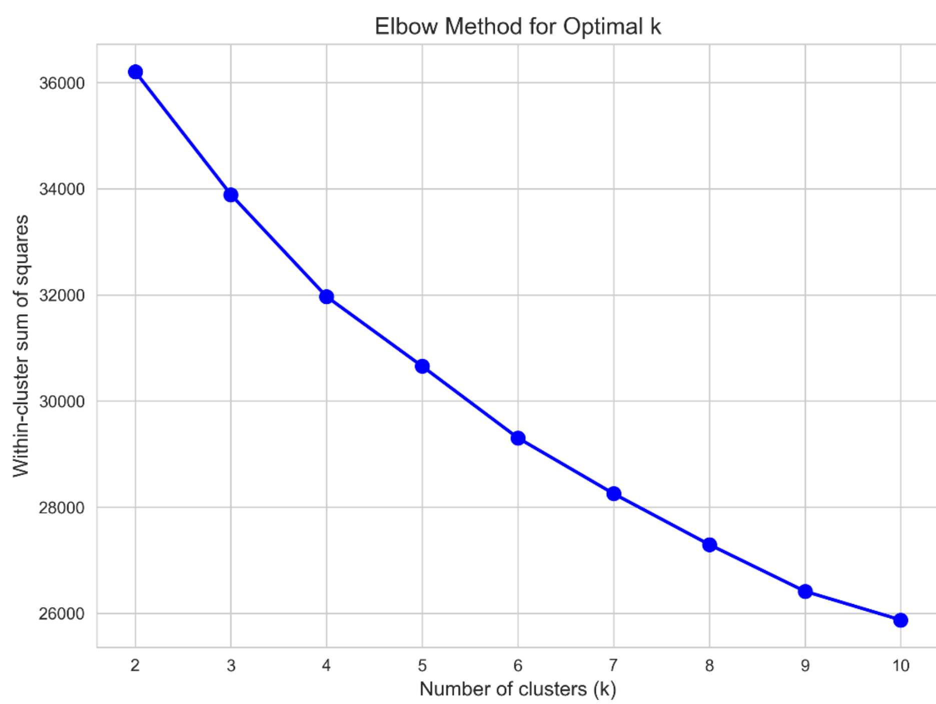

Figure 1. Elbow plot of WCSS versus number of clusters (K-Means). Elbow Method for Determining the Optimal Number of Clusters.

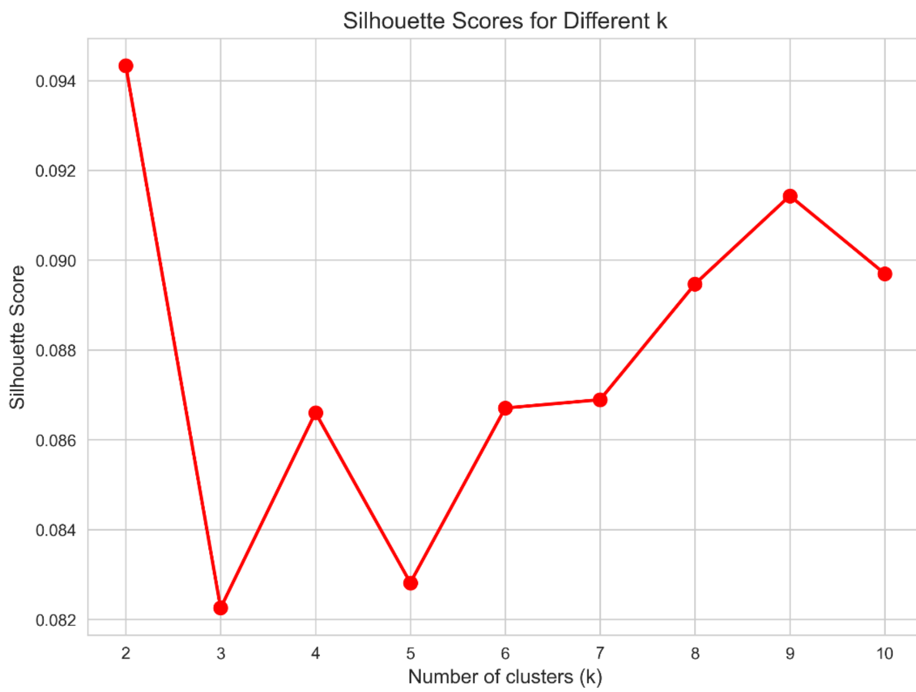

Figure 2. Silhouette scores for varying numbers of clusters (K-Means).

3.5 Supervised Learning for Multi-Class Grade Prediction

The primary objective of this phase was to predict the final letter grade (A, B, C, D, F) using a set of predictors. The following steps were undertaken:

3.5.1 Feature Preparation:

After dropping non-informative columns (personal identifiers, and clustering results), the remaining dataset was transformed via one-hot encoding for categorical variables. This process generated a high-dimensional feature space that includes both numeric and encoded categorical predictors (e.g., Department, Extracurricular_Activities, Internet_Access_at_Home).

3.5.2 Target Encoding and Class Distribution:

Since the target variable, Grade, is categorical so there is a need to encode it. The used encoding is “A” → 0, “B” → 1, “C” → 2, “D” → 3, “F” → 4.

A frequency analysis (Table 4) revealed a moderately imbalanced distribution, necessitating the use of stratified train–test splitting and oversampling.

Letter Grade	Frequency
B	1046
D	992

C	989
A	988
F	985

Table 4: Grades Frequency

3.5.3 Handling Class Imbalance:
Stratified splitting ensured that the training and testing sets maintained similar grade distributions. In addition, SMOTE (Synthetic Minority Over-sampling Technique) was applied to the training set to generate synthetic samples for underrepresented classes, thereby mitigating the impact of imbalance.

3.5.4 Classification Models:
Four models were trained and evaluated: Random Forest, Logistic Regression (with increased iteration limits), XGBoost, and Support Vector Machine (SVM). Each model was trained on the SMOTE-resampled training data and evaluated on the original test set using accuracy as well as precision, recall, and F1-score. The performance metrics for each classifier are summarized in Table 5.

3.4 Feature Importance Analysis

Feature importance was evaluated using both the Random Forest and XGBoost models to quantify the contribution of each predictor to the grade prediction task. The Random Forest model computed feature importances using the Gini impurity measure, which quantifies the reduction in impurity achieved at each split where a feature is used. This method assigns a relative importance score to each predictor, reflecting its contribution to overall model performance.

Similarly, the XGBoost model computed feature importances based on the gain metric, which measures the improvement in accuracy brought by a feature to the branches in which it is used. This gain-based measure indicates how much each predictor contributes to the reduction in the model's loss function during boosting. .

4. Results and Discussion

4.1 Data Cleaning and Feature Reduction

The initial data cleaning phase ensured that the dataset was complete and consistent by imputing missing values and correcting discrepancies in letter grades and name–gender mappings. The removal of personal identifiers and the column directly reflecting the outcome (Total_Score) reduced the risk of data leakage and enhanced the focus on predictors of student performance. The updated class distribution (Table 4) revealed that while the letter grades are relatively balanced, slight imbalances exist that were addressed via SMOTE.

4.2 Clustering Findings

Both K-Means and Hierarchical Clustering yielded a three-cluster solution.

- **K-Means Clustering:** As presented in Table 2, one cluster exhibited higher final exam scores and greater participation, another had higher attendance but lower exam scores, and the third showed moderate performance. Figure 3 shows a two-dimensional projection of the K-Means clusters using Principal Component Analysis (PCA). *Significant overlap is visible, which aligns with the relatively low silhouette scores*.
- **Hierarchical Clustering:** Table 3 and the PCA visualization (Figure 4) revealed a slightly different grouping, particularly in terms of study hours and midterm performance, yet overall, the clusters were similarly overlapping. These results imply that while natural groupings exist, student performance is multifaceted and not easily segmented.
- **Optimal Cluster:** The the overlapping clusters depicted in the Elbow and Silhouette plots (Figures 1 and 2) may indicate that academic performance exists on a continuum rather than in distinct, discrete categories, which reflects the complexity of student performance. Therefore, the initial cluster number k=3 might be a good choice for clustering.

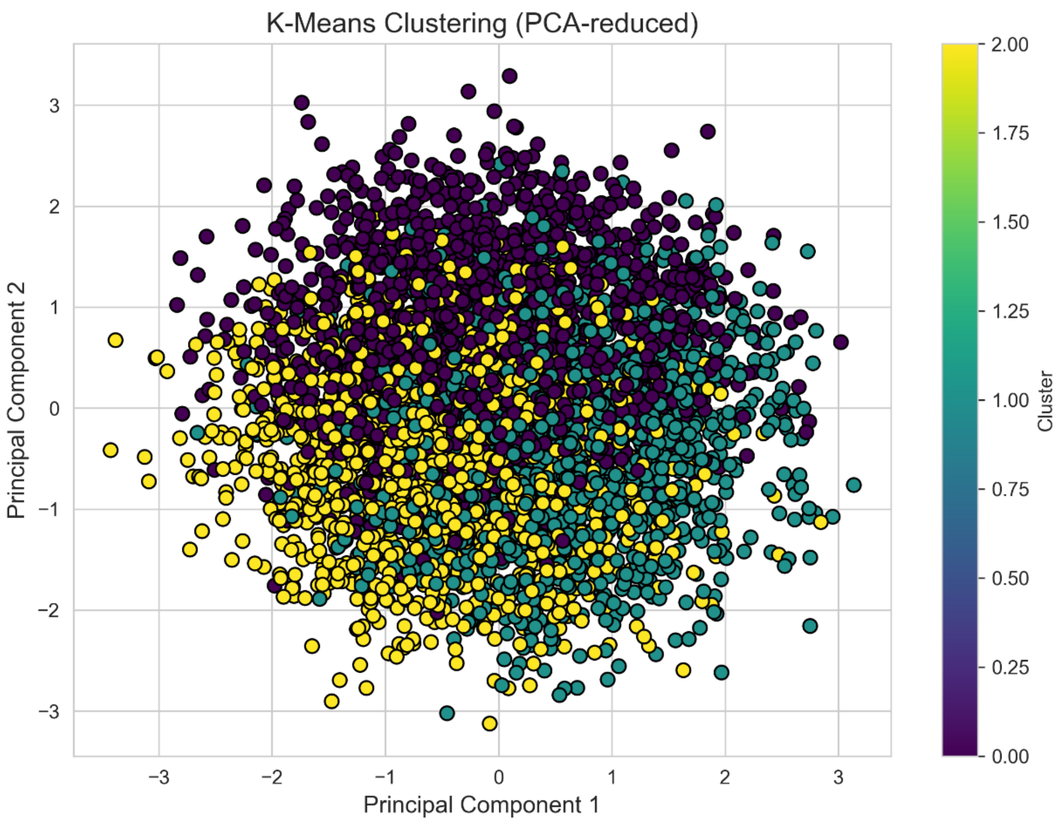

Figure 3: PCA-reduced visualization of K-Means clusters

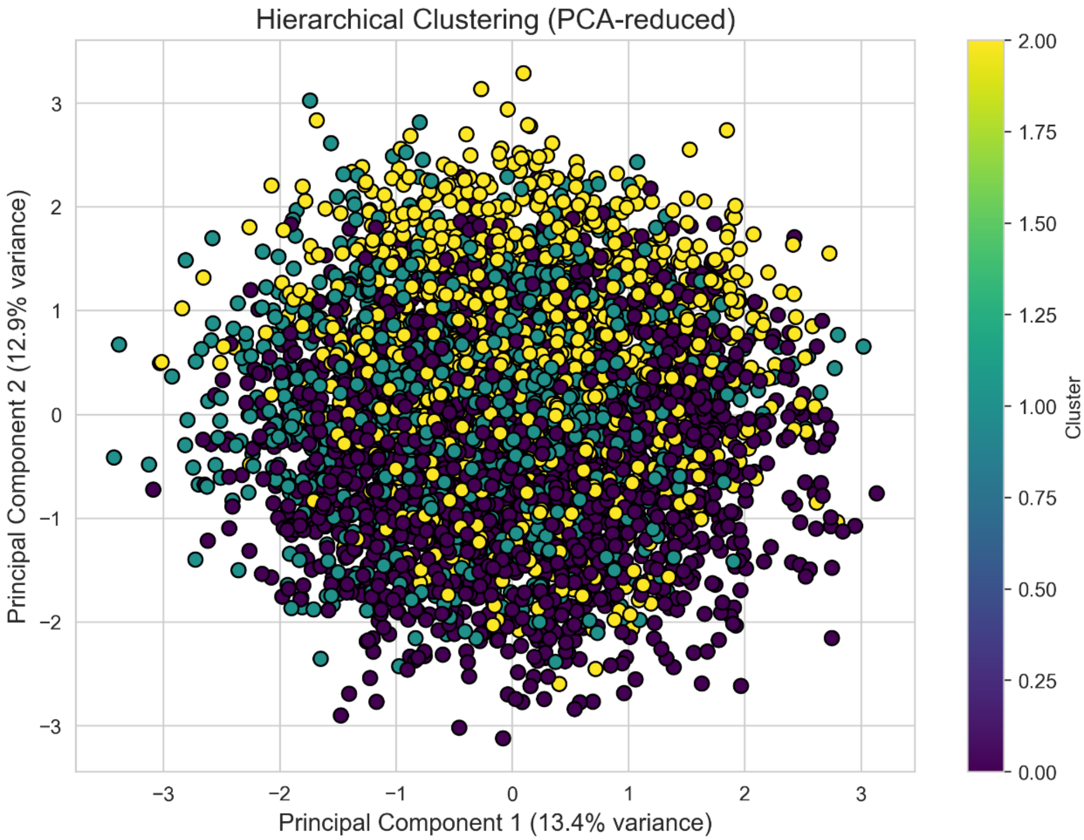

Figure 4: Average Feature Values per Hierarchical Cluster (K=3).

4.3 Supervised Classification Performance

The supervised learning experiments addressed a multi-class prediction problem with five letter grades. Despite the improved data quality and dimensionality reduction, the performance of all models was modest. Table 5 summarizes the evaluation metrics for the four classifiers:

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.20	0.20	0.20	0.20
Logistic Regression	0.20	0.20	0.21	0.20
XGBoost	0.19	0.19	0.19	0.19
SVM	0.21	0.22	0.21	0.19

Table 5: The metrics evaluation for the four classifiers.

The overall accuracies ranged between 19% and 21%, indicating significant challenges in distinguishing among the letter grades. Low F1-scores underscore difficulties in predicting minority classes. These findings suggest that the available predictors, while relevant, capture only part of the complex determinants of academic outcomes. Although the models used SMOTE and class weighting, the inherent overlap in the data makes it difficult for the models to achieve high accuracy which reflects the limitation of machine learning models to capture the nuance of student performance. These results suggest the need to explore other models such as deep learning models and other features such as motivation.

4.4 Feature Importance and Contributions

Feature importance analysis provided key insights into which predictors most influence the grade prediction:

- **Random Forest Analysis (Table 6, Figure 5):** Projects_Score, Midterm_Score, and Quizzes_Avg emerged as the most influential features, with Attendance (%) also ranking highly.
- **XGBoost Analysis (Table 7, Figure 6):** Variables such as Internet_Access_at_Home, Parent_Education_Level (High school and master level), and Participation_Score were highlighted.

These results indicate that both academic performance metrics (e.g., test and project scores) and student engagement (e.g., attendance, and Internet access) contribute to final grades. The convergence of feature importance findings across models suggests that interventions focused on enhancing project work and ensuring consistent class engagement may have a measurable impact on student outcomes.

Feature	**Importance**
Projects_Score	0.0937
Midterm_Score	0.0936
Quizzes_Avg	0.0935
Attendance (%)	0.0927
Participation_Score	0.0919
Assignments_Avg	0.0907
Study_Hours_per_Week	0.0900
Sleep_Hours_per_Night	0.0818
Stress_Level (1-10)	0.0592
Age	0.0516

Table 6: Top 10 most influential features in the Random Forest model

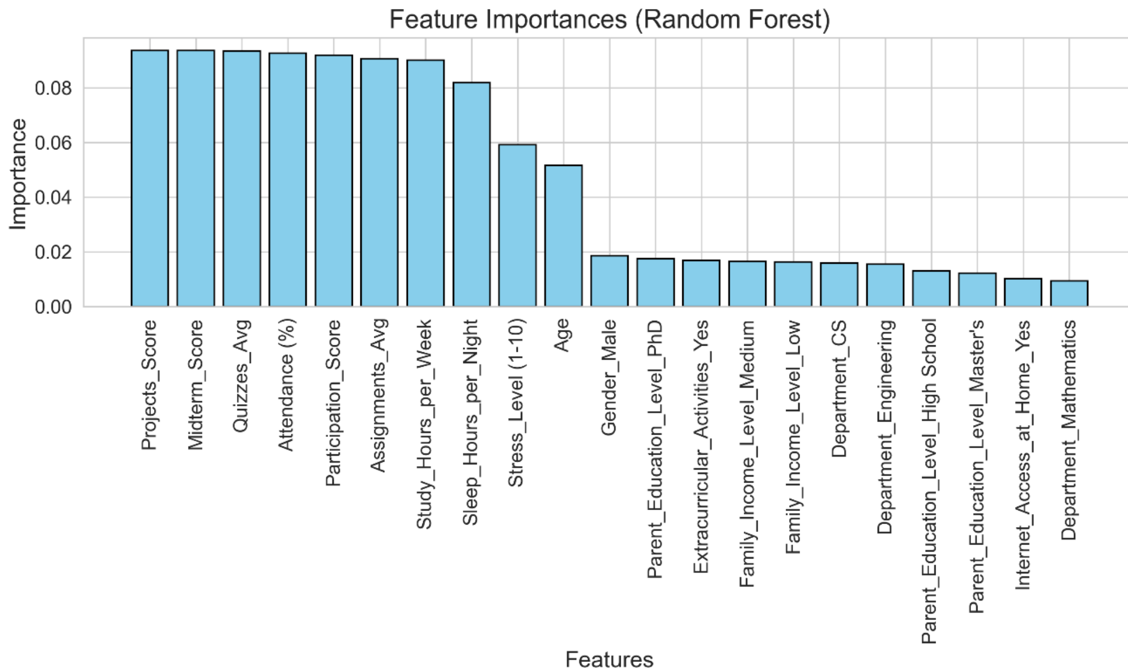

Figure 5: Feature importance as revealed by Random Forest

Feature	Importance
Internet_Access_at_Home_Yes	0.0510
Parent_Education_Level_High School	0.0498
Participation_Score	0.0495
Family_Income_Level_Medium	0.0493
Midterm_Score	0.0492
Study_Hours_per_Week	0.0492
Assignments_Avg	0.0490
Quizzes_Avg	0.0490
Projects_Score	0.0486
Parent_Education_Level_Master's	0.0481

Table 7: Top 10 most influential features in the XGBoost model

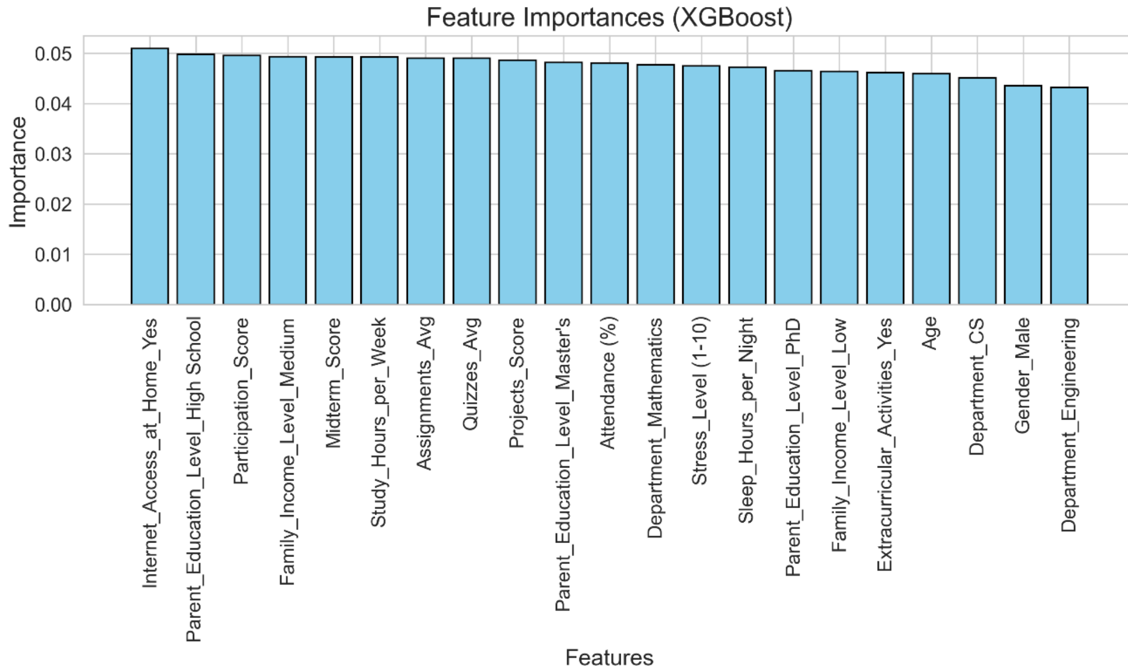

Figure 6: Features importance as found by XGBoost

4.5 Integrated Discussion of Contributions

The integrated findings of this research demonstrate several key contributions:

- **Data Quality Improvement:** The cleaning process, including the correction of letter grade inconsistencies and name and gender mismatches, has resulted in a dataset that more accurately reflects student performance, thereby ensuring that subsequent analyses are based on reliable data.
- **Clustering Insights:** The clustering experiments revealed that students exhibit overlapping performance profiles, highlighting the inherent complexity of academic achievement. Although the clusters were not sharply defined, the differences in academic and engagement metrics provide a basis for targeted educational interventions.
- **Supervised Learning Challenges:** The modest performance of the multi-class classifiers underscores the difficulty of predicting final letter grades based solely on the available features. *This suggests that additional predictors (e.g., student motivation, instructional quality, time management) or more advanced modeling approaches may be necessary.*
- **Key Predictors Identified:** Feature importance analysis consistently identified a few key variables—such as project score, midterm Score, attendance, existence of Internet connection, and parent education level that have the greatest impact on grade determination. These findings provide actionable insights for educators and administrators seeking to improve student performance.

This research contributes to the literature by demonstrating that rigorous data cleaning and feature selection are critical for obtaining reliable predictive models in the domain of educational performance. Even when predictive accuracies remain modest, the study highlights the importance of student engagement and academic metrics in understanding student outcomes.

5. Future Work

Future work should explore the integration of qualitative measures of student engagement, such as motivation, time management, and instructor evaluations, to capture performance dimensions that are not fully represented by existing quantitative metrics. In addition, investigating advanced ensemble techniques, including stacking and voting classifiers, may enhance the modeling of subtle patterns inherent in multi-class predictions. Deep learning approaches also warrant exploration for their capacity to automatically learn complex feature interactions, potentially leading to improved predictive performance.

5. Conclusion

This study presents a comprehensive analysis of student performance using a multi-phase approach encompassing data cleaning, clustering, and supervised classification. By addressing data inconsistencies and reducing dimensionality through the removal of non-informative and potentially leaky features, the research establishes a high-quality dataset for exploring student outcomes. Clustering experiments using both K-Means and Hierarchical methods reveal that while distinct subgroups exist, the boundaries between performance levels are blurred, reflecting the continuous nature of academic achievement. Supervised learning models, evaluated through a multi-class framework, achieve modest accuracy, highlighting the challenges inherent in predicting letter grades. Feature importance analyses further emphasize that specific academic and engagement metrics, notably project scores, midterm scores, parents' education level, the existence of Internet connection at home, and attendance, are critical determinants of student success.

The contributions of this research lie in its methodical improvement of dataset quality, the nuanced understanding of student groupings, and the identification of key predictive factors. Future work should focus on incorporating additional variables and exploring more advanced ensemble techniques to enhance predictive performance and provide deeper insights into the multifaceted nature of academic achievement.

References

[1] "COE - high school graduation rates." Available: https://nces.ed.gov/programs/coe/indicator/coi/high-school-graduation-rates. Accessed March 17, 2025.

[2] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting Students' Performance in Distance Learning Using Machine Learning Techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004, doi: 10.1080/08839510490442058.

[3] P. Sweeney, M. Bailey, and D. Nolan, "Student Performance Prediction Using Machine Learning Approaches: A Systematic Literature Review," *Journal of Educational Computing Research*, vol. 59, no. 1, pp. 3–31, 2021, doi: 10.1177/07356331211057644.

[4] H. Marbouti, A. Diefes-Dux, and K. Madhavan, "A Machine Learning Model for Predicting Student Performance in Higher Education," in *Proc. IEEE Frontiers in Education Conference (FIE)*, 2016, pp. 1–9, doi: 10.1109/FIE.2016.7757678.

[5] M. A. Shahiri, W. Husain, and N. A. Rashid, "Predicting Students' Academic Performance Using Supervised Machine Learning Techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[6] T. Al-Radaideh, E. Al-Shawakfa, and M. Al-Najjar, "Application of Data Mining and Machine Learning Approaches in Predicting Student Performance: A Review," *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, vol. 9, no. 5, pp. 191–198, 2018.

[7] P. Cortez and A. Silva, "Comparative Analysis of Machine Learning Techniques for Student Academic Performance Prediction," in *Proc. Int. Conf. Educational Data Mining (EDM)*, 2008, pp. 1–10.

[8] R. Dekker, G. Pechenizkiy, and J. Vleeshouwers, "Predicting Student Dropout in Higher Education Using Machine Learning Models," *Expert Systems with Applications*, vol. 113, pp. 267–278, 2018, doi: 10.1016/j.eswa.2018.01.037.

[9] A. Al-Badawi, M. Al-Sarem, and A. Alyoubi, "A Hybrid Machine Learning Model for Enhancing Student Academic Performance Prediction," *Computers and Education: Artificial Intelligence*, vol. 1, pp. 1–12, 2020, doi: 10.1016/j.caeai.2020.100002.

[10] X. Xu, Y. Zhang, and L. Chen, "Impact of Attendance and Study Habits on Academic Performance: A Machine Learning Approach," *Education and Information Technologies*, vol. 27, no. 4, pp. 5123–5140, 2022, doi: 10.1007/s10639-021-10744-9.

[11] A. Minaei-Bidgoli, W. F. Punch, and H. D. Lucy, "Feature Selection for Improving Model Predictions in Higher Education," *IEEE Transactions on Education*, vol. 47, no. 3, pp. 432–441, 2004, doi: 10.1109/TE.2004.825857.

[12] S. Qu, S. K. Chattopadhyay, and R. A. Hahn, "High School Completion Programs," *Journal of Public Health Management and Practice*, vol. 22, no. 3, pp. E47–E56, 2016, doi: https://doi.org/10.1097/phh.0000000000000286.

[13] Hussain, M., Zhu, W., Zhang, W., & Abidi, S. (2020). "Improving student academic performance prediction using deep learning and feature selection techniques." *IEEE Transactions on Learning Technologies*, 13(3), 423-435. https://doi.org/10.1109/TLT.2020.2987823

[14] Khan, S., Javed, I., & Ahmad, M. (2021). "Predicting student academic success using deep learning models: A case study on online learning behavior." *Computers & Education*, 160, 104034. https://doi.org/10.1016/j.compedu.2021.104034

[15] Sun, L., Chen, Y., & Zhao, J. (2022). "Analyzing student engagement in virtual learning environments: A machine learning approach." *Journal of Educational Data Mining*, 14(2), 67-84. https://doi.org/10.35542/jedm.2022.002

[16] Nguyen, H., Wang, T., & Li, S. (2023). "Using learning management system data for predictive analytics in higher education: A comparative study." *Educational Technology & Society*, 26(1), 55-73.

[17] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you? Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

[18] Zhang, Y., Wang, J., & Liu, C. (2023). "Feature importance analysis in student performance prediction: An interpretable machine learning approach." *Artificial Intelligence in Education*, 34(4), 312-329.

[19] Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018). "Fairness in machine learning: Lessons from political philosophy." *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149-159.

[20] Li, J., Yu, H., & Zhang, X. (2024). "Mitigating bias in student success prediction: A fairness-aware approach." *Educational Data Science Journal*, 11(1), 88-102.