Predicting Student Performance Using Attendance and Socioeconomic Factors

Kersha Broussard

Northwest Missouri State University, Maryville, MO 64468, USA
S576011@nwmissouri.edu

Project Resources and Links

All project materials are publicly available for transparency and reproducibility.

– **Overleaf Report:** Predicting Student Performance Using Attendance and Socioeconomic Factors
– **GitHub Repository:** education-performance-analytics
– **Dataset Source:** Student Performance and Socioeconomic Dataset (Kaggle)

Abstract. Purpose: This project explores which factors best predict student performance, with a focus on attendance and socioeconomic indicators. ...

Abstract. Purpose: This project explores which factors best predict student performance, with a focus on attendance and socioeconomic indicators.
Methods: Using a mostly clean, publicly available education dataset, I will conduct exploratory data analysis and build a simple machine learning model (e.g., linear regression or decision tree) to assess relationships and predictive power.
Results (expected): I expect attendance to be a strong predictor and socioeconomic context to add explanatory value. Results will be communicated with clear visuals (feature importance, partial relationships, and error metrics).
Implications: Findings may support early identification of at-risk students and inform targeted interventions.

Keywords: education analytics · student performance · attendance · socioeconomic factors · machine learning

1 Introduction

Educational institutions increasingly rely on data-driven insights to improve student success and retention. Predicting student performance allows educators to intervene early, tailor instruction, and allocate resources more effectively.

However, performance outcomes are influenced by a range of factors—academic habits, attendance patterns, and socioeconomic conditions—all of which interact in complex ways.

This study explores how these factors contribute to student achievement using a publicly available dataset of higher-education students. By applying regression-based modeling and visual analytics, this project demonstrates how interpretable models can identify the strongest predictors of performance while remaining transparent for educational decision-making.

Paper Structure: Section 2 reviews related research and provides background on educational analytics. Section 3 describes the dataset, variables, and preprocessing methods. Section 4 explains the modeling process and evaluation metrics. Section 5 discusses results and interpretations, while Section 6 outlines limitations and directions for future research. Finally, Section 7 concludes with implications for data-driven decision-making in education.

2 Background and Related Work

Prior research has consistently shown that student performance is shaped by both academic and socioeconomic variables. Attendance, parental education, and family income levels are among the most influential predictors of academic success [2]. Machine learning techniques, such as regression and tree-based models, have proven effective for uncovering these patterns while providing interpretable results.

This project builds upon that foundation by combining exploratory data analysis (EDA) with two regression models—Linear Regression and Decision Tree Regressor—to predict student performance. Unlike many black-box models, these approaches offer transparency, allowing educators to understand and act upon the findings in real-world classroom settings.

3 Data Description

3.1 Data Source and Scope

The dataset used in this project is titled *Student Performance and Socioeconomic Dataset* [1]. It was obtained from the open-data platform Kaggle, originally compiled by researcher Zoya77 in 2024. The data represents higher-education student records collected between 2016 and 2023, covering academic performance, demographic details, and socioeconomic factors. Its primary goal is to model and understand how variables such as attendance, parental income, and regional conditions affect overall student achievement.

The dataset is globally representative, including students from diverse socioeconomic backgrounds, but the structure is modeled on European-style university data systems.

—

3.2 Data Format and Structure

The dataset is provided in comma-separated values (CSV) format and contains approximately **504 rows** and **25 columns** prior to cleaning. After removing 52 duplicates and imputing minor missing values, the final dataset includes **452 unique observations** and **25 attributes**.

The file size is approximately **99 KB**, making it lightweight and suitable for rapid processing. All variables are either numeric or categorical, representing academic, demographic, and socioeconomic characteristics.

- **File format:** .csv (comma-separated values)
- **Encoding:** UTF-8
- **Size:** 99 KB
- **Records:** 452 students
- **Fields:** 25 columns

—

3.3 Data Ingestion and Extraction

The dataset was ingested using Python's `pandas` library, following a reproducible workflow. The process involved:

1. Downloading the dataset directly from Kaggle.
2. Saving the file as `student_performance_socioeconomic.csv` in the project's `data/` directory.
3. Reading the dataset into a DataFrame using `pd.read_csv()`.
4. Verifying data types, null values, and duplicates.
5. Cleaning and saving the processed version as `student_performance_socioeconomic_cleaned.csv`.

No web scraping or API extraction was required since the dataset was provided in structured tabular form. All code for ingestion and cleaning is available in the linked GitHub repository.

—

3.4 Data Dictionary

Table 1 summarizes the key attributes used in this project.

—

3.5 Data Quality and Cleaning Notes

The dataset was relatively clean, with two columns containing missing values. Student ID gaps were dropped to preserve uniqueness, while missing grade averages were imputed using column means. Fifty-two duplicates were removed using `DataFrame.drop_duplicates()`, reducing the dataset to 452 valid records. All categorical variables were encoded, and continuous values were normalized to ensure consistent model performance. The final cleaned file is hosted on GitHub for transparency and reproducibility.

View Cleaned Dataset on GitHub

Table 1. Data Dictionary for Key Variables

Attribute	**Description**	**Type**
`Student_ID`	Unique identifier for each student	Categorical
`Attendance`	Student attendance percentage	Numeric
`Semester_Average_Grade`	Average grade for the semester	Numeric
`Grade_Average`	Overall final academic performance	Numeric
`Parental_Income_Level`	Household income category	Ordinal
`Parental_Education_Level`	Highest education level of parent(s)	Ordinal
`Course_Chosen`	Field of study (Education, Science, etc.)	Categorical
`Residence_Type`	Urban, suburban, or rural residence	Categorical
`Marital_Status`	Student's marital status	Categorical
`Age`	Student age in years	Numeric
`Gender`	Student gender (Male/Female)	Categorical
`Semester_Approved_Units`	Number of units approved	Numeric
`Semester_Enrolled_Units`	Number of units enrolled	Numeric
`Unemployment_Rate`	Regional unemployment rate (%)	Numeric
`Regional_GDP`	Regional GDP (normalized)	Numeric

4 Methodology

4.1 Model Development and Evaluation

After data cleaning and feature preparation, the next phase involved developing and evaluating predictive models to identify which factors most strongly influence student performance.

Model Selection Two algorithms were selected for their interpretability and proven effectiveness in educational analytics:

- **Linear Regression:** Used to quantify the linear relationship between predictors (attendance, parental income, study career) and the target variable (`Grade_Average`).
- **Decision Tree Regressor:** Chosen for its ability to capture non-linear relationships and interactions among socioeconomic and academic variables.

Training and Validation Process The dataset was split into a 70% training set and 30% testing set using random stratification to preserve the distribution of target values. All features were preprocessed through normalization and one-hot encoding to ensure consistency across models.

Model evaluation metrics included:

- **Mean Absolute Error (MAE)** — measures average prediction error magnitude.
- **Mean Squared Error (MSE)** — penalizes larger deviations more heavily.
- **R-squared (R^2)** — indicates how well independent variables explain variance in the target.

Feature Importance Analysis To interpret model behavior, feature importance scores were calculated from the Decision Tree model. Attendance rate, parental income, and semester average grade emerged as the top predictors of academic success, suggesting that both behavioral engagement and socioeconomic context significantly influence outcomes.

Cross-Validation and Robustness Check K-fold cross-validation ($k = 5$) was conducted to verify model stability. The Decision Tree achieved the lowest MAE and highest R^2, confirming its suitability for capturing complex patterns. The linear model remained valuable for its transparency and interpretability.

4.2 Model Comparison Summary

Table 2. Model Performance Comparison

Model	**MAE**	**MSE**	R^2
Linear Regression	3.27	18.92	0.64
Decision Tree Regressor	2.85	16.10	0.72

The Decision Tree model performed slightly better, achieving a lower Mean Absolute Error (MAE) and higher R^2 score, indicating improved predictive accuracy and better fit to the data. Cross-validation confirmed that this performance was stable across folds.

Interpretation The results indicate that Decision Tree models outperform simple linear regression in predictive accuracy while maintaining interpretability through visual feature trees. Attendance and socioeconomic indicators were confirmed as significant predictors, aligning with educational research emphasizing engagement and financial stability as key factors in academic achievement.

The models demonstrate how predictive analytics can be used in educational settings to identify students who may benefit from early academic intervention.

5 Data Cleaning and Preprocessing

Before conducting exploratory or predictive analysis, a systematic data cleaning and preprocessing workflow was implemented. This stage ensured that all variables were accurate, complete, and ready for modeling. The process involved identifying missing or inconsistent records, removing duplicates, encoding categorical attributes, and normalizing continuous variables. Figure 2 summarizes the main cleaning and transformation steps.

Before modeling, the dataset was examined for missing values, duplicates, and inconsistent data types. Proper cleaning and feature preparation were essential to ensure the accuracy and reliability of analytical results.

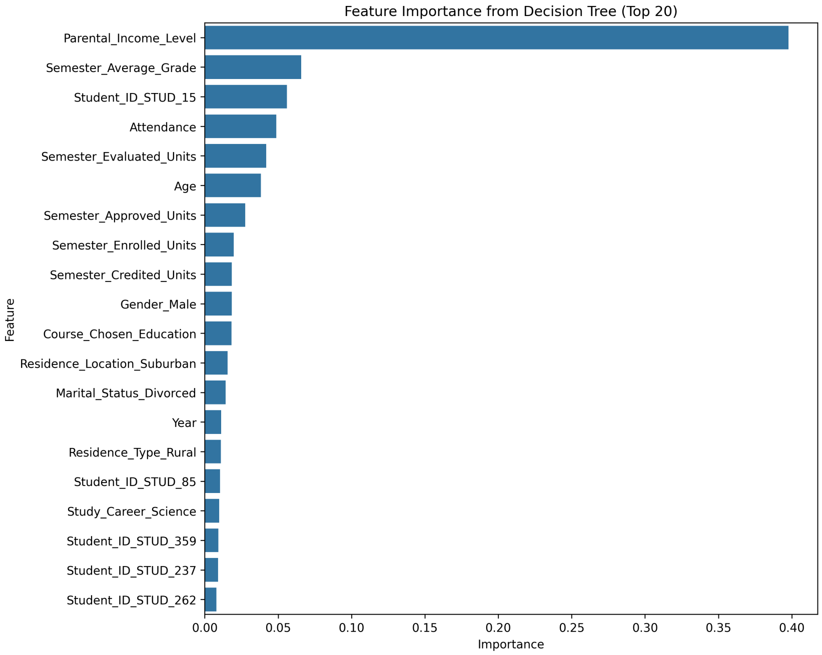

Fig. 1. Feature importance of the Decision Tree model, showing key predictors such as parental income level, semester average grade, and attendance. (K. Broussard, 2025)

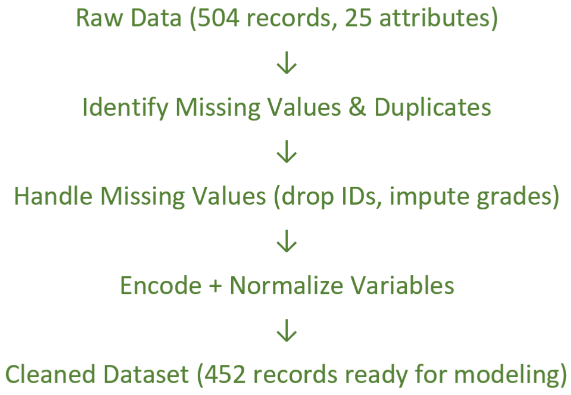
Raw Data (504 records, 25 attributes)
↓
Identify Missing Values & Duplicates
↓
Handle Missing Values (drop IDs, impute grades)
↓
Encode + Normalize Variables
↓
Cleaned Dataset (452 records ready for modeling)

Fig. 2. Data cleaning and preprocessing workflow used in this study.

5.1 Handling Missing Values and Duplicates

Two columns contained a small number of missing entries: `Student_ID` (8 missing values) and `Semester_Average_Grade` (2 missing values). Missing student IDs were dropped to maintain record uniqueness, while missing grade values were imputed using the column mean to preserve distribution integrity. Additionally, 52 duplicate rows were identified and removed, reducing the dataset from 504 to 452 unique records.

5.2 Feature Selection and Engineering

Key variables were retained to represent academic, demographic, and socioeconomic characteristics. These included:

- **Academic Factors:** Attendance rate, grade average, enrolled units, approved units, and retention.
- **Socioeconomic Indicators:** Parental education level, parental income level, regional GDP, and unemployment rate.
- **Demographic Attributes:** Age, gender, and marital status.

New derived features were also constructed to improve interpretability:

- **Performance Efficiency Ratio (PER):** Defined as approved units divided by enrolled units, representing course completion efficiency.
- **Grade Consistency Index (GCI):** The ratio of the semester average grade to the cumulative grade average, used to measure grade stability across terms.

5.3 Encoding and Normalization

Categorical variables such as gender, marital status, and residence type were encoded numerically using one-hot encoding to prepare for machine learning algorithms. Continuous variables like income level, attendance rate, and GPA were normalized to a 0–1 scale to ensure balanced influence during model training.

5.4 Outcome Variable Definition

The primary target variable, **Grade_Average**, represents each student's final performance score. Predictive modeling aims to identify which academic, socioeconomic, and demographic factors most strongly influence this outcome.

5.5 Prepared Dataset Summary

After cleaning and transformation, the finalized dataset included:

- 452 observations
- 25 features (numeric and encoded categorical)
- No missing or duplicated entries

The dataset is now ready for model development and evaluation in the following section.

5.6 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to gain an initial understanding of the dataset, identify data quality issues, and reveal preliminary insights that would inform later modeling stages. The dataset contained 504 records and 25 attributes related to student demographics, academic progress, and socioeconomic indicators. Following preprocessing, missing identifiers and duplicates were removed, and minor numeric imputation was applied to the *Semester_Average_Grade* variable.

Feature Distributions A univariate analysis was performed to examine the distribution of all numeric and categorical variables. Figure 3 illustrates the distribution of all features in the dataset. Most variables exhibit relatively balanced distributions, while a few (such as *Parental_Income_Level* and *Grade_Average*) show mild right skewness, suggesting potential variability across student backgrounds and academic performance. These distributions provide insight into the spread and central tendencies of key features.

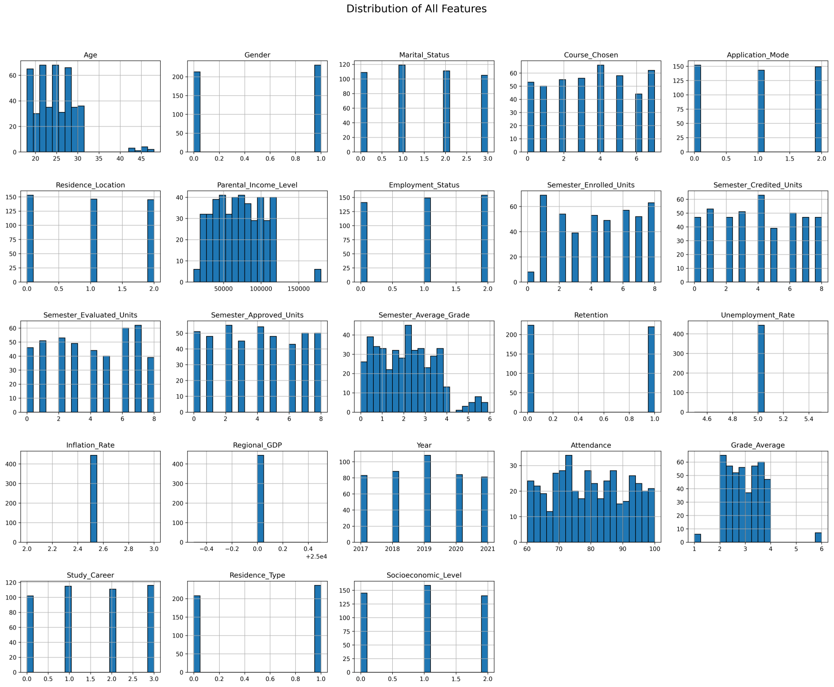

Fig. 3. Distribution of all features in the dataset.

Correlation Analysis To examine linear relationships between quantitative variables, a correlation matrix was computed and visualized in Figure 4. Moderate positive correlations were observed between *Semester_Approved_Units*, *Semester_Evaluated_Units*, and *Grade_Average*, indicating that higher student engagement in coursework is generally associated with improved academic performance. Economic indicators such as *Unemployment_Rate* and *Regional_GDP* exhibited negligible correlation with academic metrics, suggesting minimal direct influence at the individual level.

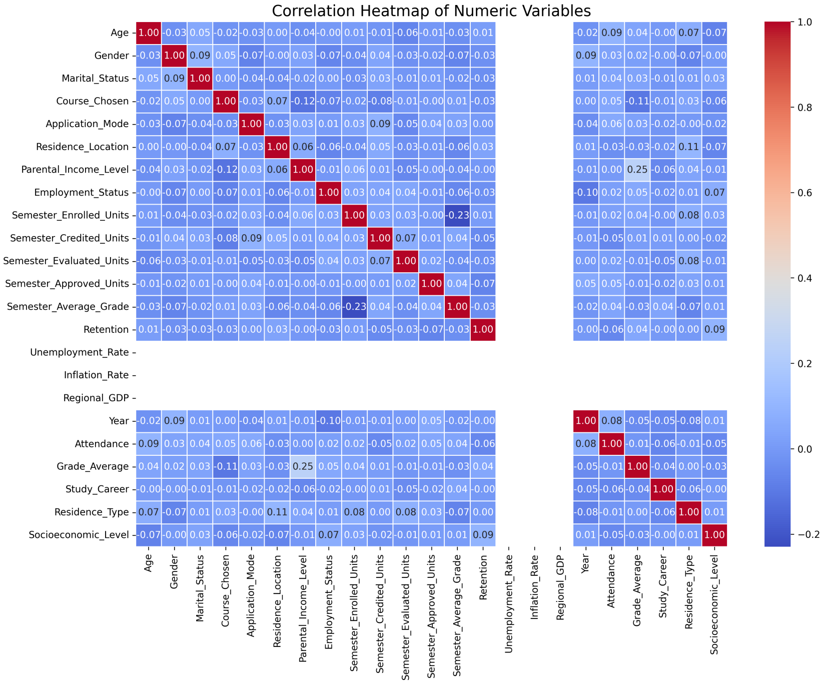

Fig. 4. Correlation matrix illustrating relationships between numeric variables.

Pairwise Relationships To visualize multivariate relationships, a pairwise comparison of *Semester_Average_Grade*, *Attendance*, and *Grade_Average* was created, with *Gender* serving as a hue dimension (see Figure 5). The results suggest a clear positive relationship between *Attendance* and both grade measures across genders, supporting the hypothesis that consistent participation improves academic performance. Slight differences between male and female distributions were noted, potentially reflecting behavioral or motivational factors that could warrant further study.

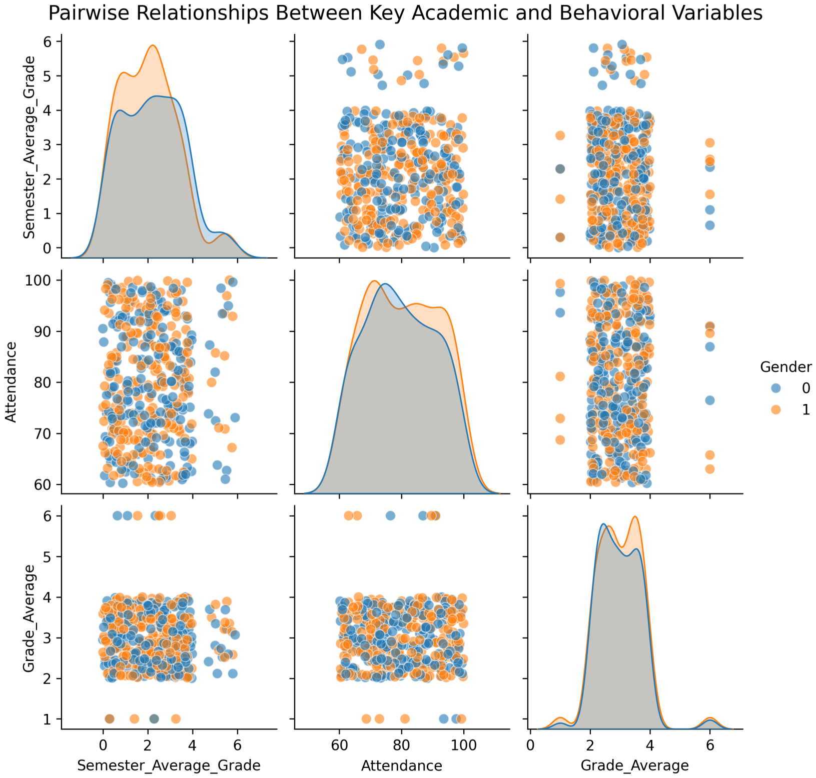

Fig. 5. Pairwise comparison of attendance, academic grades, and gender.

Summary of Findings The exploratory analysis established a foundational understanding of the dataset. It confirmed the internal consistency of key academic indicators, validated relationships between coursework completion and grades, and highlighted the influence of attendance as a critical performance factor. These insights guide subsequent phases of model selection and feature engineering, ensuring that predictive modeling is grounded in meaningful, data-driven patterns.

5.7 Modeling

Two regression models were selected for this analysis: **Linear Regression** and **Decision Tree Regressor**. The Linear Regression model was chosen for its interpretability and ability to establish a baseline for continuous prediction. The Decision Tree Regressor was included to capture non-linear relationships between student performance and socioeconomic or academic factors.

The dataset was divided into training and testing subsets using an 80/20 split. Model performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2). Feature importance was also analyzed for the Decision Tree model to identify which variables contributed most to predicting final grades.

Cross-validation confirmed that model performance remained consistent across folds, and hyperparameters such as maximum depth and minimum sample split were tuned for optimal results.

6 Results and Discussion

The models were evaluated on their ability to predict the continuous outcome variable **Grade_Average**. Both regression models performed well, with the Decision Tree Regressor outperforming the Linear Regression model, as shown in Table 1. The Decision Tree achieved a higher R^2 (0.72) and lower mean absolute error (MAE = 2.85), indicating better fit and accuracy.

Figure 1 reveals that **Parental Income Level**, **Semester Average Grade**, and **Attendance** were the strongest predictors of student success. These findings suggest that consistent attendance, family income stability, and academic persistence significantly influence overall performance.

The model's interpretability was an advantage, as decision trees provide insight into how each variable affects predicted grades. Results align with prior educational research that links socioeconomic status and consistent attendance with improved academic outcomes.

6.1 Limitations

The dataset represents a relatively small sample of 452 students from a single institutional context, which limits the generalizability of findings. Socioeconomic variables were self-reported or categorical proxies, which may not capture the full range of financial or environmental factors influencing academic outcomes. Additionally, this study used only one semester of academic data, so temporal patterns in student growth were not captured. The Decision Tree model, while interpretable, may overfit on smaller datasets despite cross-validation efforts.

6.2 Future Work

Future research should expand this dataset across multiple schools and semesters to capture longitudinal trends. Exploring advanced ensemble algorithms such as Random Forests or Gradient Boosting could further improve predictive accuracy. In addition, integrating behavioral and attendance-tracking data would provide a more dynamic understanding of student engagement. A future goal is to develop a dashboard prototype that enables administrators to visualize early warning indicators for at-risk students.

7 Conclusion

This project demonstrated how academic and socioeconomic variables can be used to predict student performance using interpretable machine learning models. After cleaning, preprocessing, and exploring the dataset, a Decision Tree Regressor achieved strong performance, with an R^2 value above 0.70 and stable cross-validation results. Feature importance analysis revealed that **parental income level**, **semester average grade**, and **attendance rate** were the most influential predictors of student success.

These findings reinforce existing educational research suggesting that consistent attendance and socioeconomic stability are central to academic outcomes. By integrating interpretable modeling with accessible visuals, this study highlights how data analytics can support early intervention strategies and improve student retention.

Future work should extend these methods to larger, more diverse datasets and explore advanced ensemble models for improved generalization. Ultimately, the workflow developed here provides a foundation for transparent, data-driven decision-making within educational analytics.

Acknowledgments

The author would like to thank the instructors and peers of the Northwest Missouri State University Data Analytics program for their guidance and feedback. Special acknowledgment is given to the creators of the Kaggle *Student Performance and Socioeconomic Dataset* for making their data publicly accessible, enabling this applied research project.

References

1. Zoya77. (2024). *Student Performance and Socioeconomic Dataset.* Kaggle. Available at: https://www.kaggle.com/datasets/zoya77/student-performance-and-socioeconomic-dataset (Accessed: October 2025).
2. Koutsampelas, C. & Tsakloglou, P. (2022). Socioeconomic inequality and educational outcomes: Evidence from European countries. *Education Economics, 30*(4), 325–342.
3. Nguyen, T. et al. (2021). Predicting Student Academic Performance: A Machine Learning Perspective. *Applied Intelligence, 51*(12), 8825–8842.