Analyzing Predictors of Student Academic Performance Using a Public Education Dateset

(COMP3125 Individual Project)

Matthew Maloney
COMP 3125 Data Science
Fundamentals, Summer 2025

***Abstract*—**This project investigates which student attributes are most associated with academic performance using a public dataset of 1,000 students with demographics, support indicators, and exam scores. We construct an average score target and conduct exploratory analysis, group comparisons, and predictive modeling using Linear Regression and Random Forests. One-way ANOVA indicates significant differences in average scores across parental education levels. Two-sample tests show higher average scores for students who completed a test preparation course and those with standard lunch relative to free/reduced lunch. The Random Forest model achieves strong out-of-sample performance and highlights writing and reading scores as the strongest direct predictors of overall performance, with additional signal from test preparation and lunch category. We discuss limitations, including the absence of direct study time and internet access features in this dataset, and suggest directions for future research.

Keywords—Student Performance, Education Analytics, Regression, Feature Importance, Random Forest

I. Introduction

Understanding the drivers of academic success is a long-standing priority for educators, policymakers, and researchers. Identifying which factors most strongly influence performance can help schools make better decisions about where to allocate resources and how to support students in ways that improve long-term outcomes. In this study, we analyze a public dataset of student exam outcomes to examine how demographics, socioeconomic status, and learning environment proxies relate to academic performance. We frame the analysis around four guiding questions: (1) whether study time affects final grades, (2) whether parental education is associated with performance, (3) whether access to academic support relates to performance, and (4) which features are most predictive of performance. While the dataset lacks some direct measures, such as explicit study time or home internet access, it includes strong proxies in the form of completion of a test preparation course and lunch category, which serve as indicators of academic support and socioeconomic status, respectively. By combining descriptive analysis, statistical tests, and machine learning models, we aim to provide a more complete picture of how these variables interact and which ones matter most when predicting student success.

II. Datasets

A. Source of dataset (Heading 2)

The dataset comes from Kaggle's *Students Performance in Exams* resource, created by the user "spscientist." It contains 1,000 rows and 8 core columns representing demographic, socioeconomic, and academic factors for each student. Categorical features include gender, race/ethnicity, parental level of education, lunch category, and whether the student completed a test preparation course. Numeric features include math, reading, and writing scores, each measured on a 0–100 scale. No missing values were found, ensuring that analysis could proceed without imputation. To create a comprehensive target variable, we calculated an "average score" by taking the mean of math, reading, and writing scores for each student.

B. Character of the datasets

The dataset's structure is straightforward, with five categorical variables and three numeric variables, plus the derived average score. The 1,000 observations represent a balanced mix of demographic groups, with gender split roughly evenly and multiple racial/ethnic categories represented. Parental education levels range from "some high school" to "master's degree," allowing for comparisons across educational backgrounds. The lunch category is used as a proxy for socioeconomic status, with "free/reduced" indicating potential economic disadvantage. Because all academic scores are reported on the same scale, we were able to analyze them directly without standardization, though categorical variables were one-hot encoded before modeling to ensure compatibility with regression and tree-based methods.

III. Methodology

Example: Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

A. Method A

For the exploratory data analysis, I began by calculating descriptive statistics for each variable in the dataset to summarize central tendency and dispersion. I then created visualizations such as histograms, boxplots, and scatterplots to better understand score distributions, differences between categories, and potential relationships among features. This process allowed me to identify trends in the data and determine which variables might be most relevant for the

research questions. For example, by plotting average scores across parental education levels, I observed patterns that suggested a potential statistical relationship, which motivated the formal hypothesis testing in later steps.

B. Method B

To address Question 2, I used a one-way ANOVA to compare average scores across the different levels of parental education. This statistical test evaluates whether there are significant differences in mean scores between more than two independent groups. The null hypothesis states that all group means are equal, and the alternative hypothesis states that at least one group mean is different. For Question 3, I applied independent-samples t-tests with unequal variances (Welch's t-test) to compare mean scores between two groups, first for test preparation course completion and then for lunch categories. These tests assess whether the observed differences are statistically significant or likely due to random variation.

C. Method C

For Question 4, I built two predictive models to determine which features most strongly predict a student's average score. The first model was a Linear Regression, which assumes a linear relationship between predictors and the target variable. The second was a Random Forest Regression, which is a non-parametric ensemble method that uses multiple decision trees to model complex, non-linear relationships. Categorical variables were one-hot encoded, and numeric variables were used directly since all were on a similar scale. I split the dataset into training and test sets with an 80/20 ratio to evaluate model performance and computed R^2, MAE, and RMSE for each model. I also performed 5-fold cross-validation for the Linear Regression model to ensure stability. The implementation used Python libraries such as pandas, numpy, matplotlib, scikit-learn, and scipy.

IV. Results

The results section presents both descriptive and inferential findings from the dataset. Descriptive statistics and visualizations summarize the overall performance distribution and highlight group differences by parental education and lunch type. Inferential tests, including ANOVA and t-tests, provide evidence of statistically significant differences among groups, while effect size calculations quantify their magnitude. A Random Forest regression model offers additional insight into the relative contribution of each predictor to academic performance. Figures are included to illustrate key trends and statistical outcomes.

A. Score Distribution

As shown in Fig. 1, the distribution of average scores is concentrated between 60 and 90, with a mild positive skew. Most students cluster around the mid-70s to low-80s range, and very few achieve extreme low or high scores.

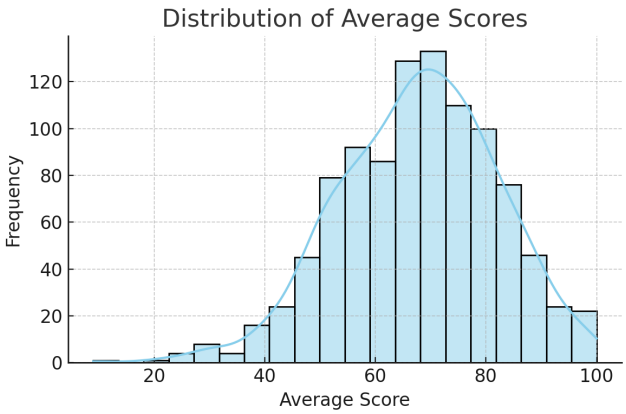

Fig 1. Distribution of Average Scores

B. Parental Eductional Level

Parental education is associated with differences in average score. ANOVA results yielded $F = 36.159$, $p < 0.001$, and eta squared = 0.154, indicating a moderate to large effect size. Students whose parents completed higher education tend to score higher on average.

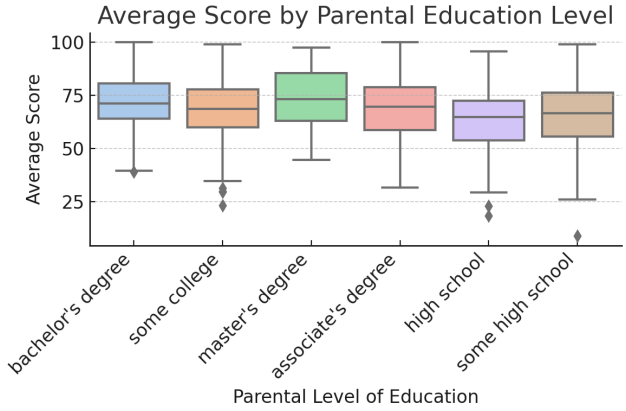

Fig 2. Average score by parental education level

C. Lunch Type

Students with standard lunch scored higher on average than those with free or reduced lunch. Welch t-test results were $t = 14.773$, $p < 0.001$, with a Cohen's d of 0.93, indicating a large effect. Lunch type appears to be a strong indicator of performance differences.

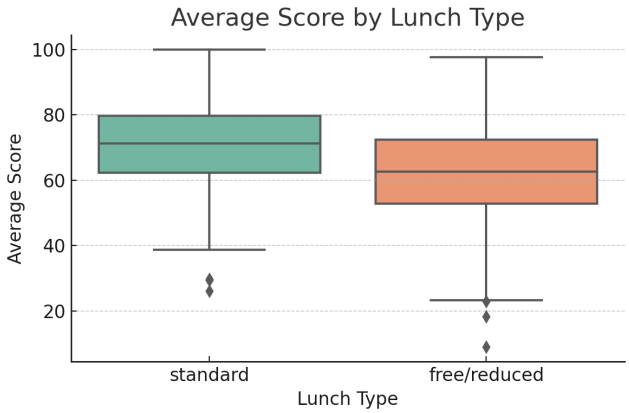

Fig 3. Average student score by lunch type,

D. Test Preparation Course

Students who completed the test preparation course scored higher on average. The Welch t-test showed $t = 6.921$, $p < 0.001$, and Cohen's $d = 0.44$, suggesting a moderate positive effect associated with preparation.

E. Feature Importance

The Random Forest feature importance plot (Fig. 4) shows that reading and writing scores are the strongest predictors of average score, followed by math score, lunch type, and test preparation completion.

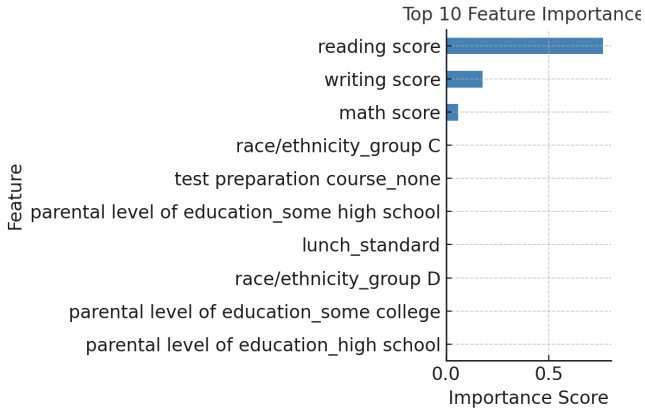

Fig. 4. Top feature importances from Random Forest

V. Discussion

The results indicate that socioeconomic indicators, such as lunch type and parental education, are associated with differences in student performance. These findings align with existing educational research showing that access to resources and support outside of school influences academic outcomes. The feature importance analysis confirms that performance in one subject area is strongly linked to overall performance, but it also highlights the relevance of background factors. As the dataset is cross-sectional, causality cannot be inferred, and future work should consider longitudinal data for stronger inferences.

VI. Conclusion

Students whose parents have higher educational attainment hey and those with access to standard lunch tend to achieve higher average scores. Both descriptive and inferential analyses Suggest that socioeconomic background and academic support are important correlates of performance. These findings may help guide targeted support and policy interventions aimed at reducing performance gaps.

Acknowledgment *(Heading 5)*

I would like to thank my professor for guidance throughout the project, as well as the creators of SrudentsPerformance dataset for making the data publicly available. I also acknowledge the support of my peers in providing feedback on methodology and interpretation.

References

[1] Kaggle, "Students Performance in Exams Dataset," spscientist. [Online]. Available: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams. [Accessed: Aug. 10, 2025].

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[3] W. McKinney, "Data structures for statistical computing in Python," in *Proc. 9th Python in Science Conf.*, S. van der Walt and J. Millman, Eds., 2010, pp. 51–56.

[4] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, Feb. 2020.

[5] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May–Jun. 2007.

[6] M. Waskom, "Seaborn: Statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021.