Aadi Sudan and Sai Musku
Professor Ghosh
CS 183
March 11, 2025

Analyzing Student Performance to Predict Final Grades

Background:

Academic performance is a critical factor of future success, serving as a prediction of. Researchers have long sought to identify the key factors influencing student performance, with early theories focusing on cognitive ability. Spearman's theory of general intelligence (g-factor) proposed that cognitive ability underlies performance in both academic and non-academic tasks. However, as research expanded, it became evident that students with similar IQ levels often performed differently, suggesting that additional factors were at play.

Bloom's Taxonomy of Educational Objectives broadened this understanding by introducing the idea that external factors, such as instructional quality and motivation, also influence learning. This led to a more comprehensive view of academic performance, incorporating both cognitive (e.g., memory, reasoning) and non-cognitive (e.g., motivation, emotional regulation) factors.

Heckman and Kautz (2012) demonstrated that non-cognitive traits, including self-control and motivation, are as important as cognitive ability in predicting academic and career success. Similarly, Duckworth and Seligman (2005) found that self-discipline was a stronger predictor of academic performance than IQ. Parental involvement and socioeconomic status (SES) have also been shown to play crucial roles in shaping student success. Jeynes (2007) conducted a meta-analysis of 77 studies and found a strong positive correlation between parental involvement and academic achievement. Motivation, teacher quality, and peer influence are also key determinants of success. Intrinsic motivation—driven by personal interest and a sense of mastery—produces better learning outcomes than extrinsic motivation (e.g., grades).

This study utilizes a publicly available dataset found on Kaggle, a popular and trustworthy website for data analysts and engineers to get their hands on datasets they can manipulate and draw conclusions from. The dataset has 6607 entries and 20 features. The features are Hours_Studied, Attendance, Parental_Involvement, Access_to_Resources, Extracurricular_Activities, Sleep_Hours, Previous_Scores, Motivation_Level, Internet_Access, Tutoring_Sessions, Family_Income, Teacher_Quality, School_Type, Peer_Influence, Physical_Activity, Learning_Disabilities, Parental_Education_Level, Distance_From_Home, Gender, and our target variable of Exam_Score.

Design:

The study aims to explore the relationship between student performance and various cognitive and non-cognitive factors using machine learning techniques. By applying

Chi-Squared tests, visualization techniques, and linear regression, we seek to uncover correlations and predictive relationships between these variables and student outcomes, and potentially create a predictive model that can levy those relationships to forecast a given student's grades.

We made many different visualizations, ranging from heatmaps to boxplots to scatterplots, to understand which features were independent and which had some level of correlation. Chi-Squared tests were used to supplement the various graphs for categorical data, whereas the basic Pearson Correlation Coefficient test was used for numerical data.

We chose to do two different linear regression tests - one using only numerical data, which we found to encompass mostly cognitive traits and aspects that would impact them, and another using numerical and categorical data, which included non-cognitive traits. By doing this, we would be able to analyze the impact non-cognitive traits have on academic performance when combined with cognitive traits, which are more typically the subject of focus.

Implementation:

The dataset, a CSV file, was loaded into a Jupyter notebook using the pandas data analysis library and displayed in order to ensure successful loading and gauge dimensionality. The next order of business was to handle any missing values or outliers. Fortunately, the dataset does not have many missing values - only three categorical features have roughly 60-90 N/A entries - so we chose to impute the mean values of those features. In addition, the numeric columns had a few outliers that heavily skewed the data and could contribute to error later on, so we chose to remove any values with a z-score of 3 or greater. By doing this, we aimed to ensure that values could only be up to a max of 3 standard deviations away from the mean, which ensures that our data is more tightly packed together and allows us to disregard any extreme values that could skew the data. This method removed 101 entries, which in the long run is a very minor percentage of our dataset, leaving us with 6506 entries that we could work with.

The next thing we did was classify features as either numerical or categorical. This would help with future analysis, such as calculating the Pearson Correlation Coefficients for numerical features or using the Chi-Square test on categorical features, visualization, and creating our models, as we wanted to test a model trained on exclusively numerical data before comparing to one trained on everything. In the end, our numerical data ended up consisting of what our research papers would define as cognitive traits - the previous score feature is the closest we could get to a performance/intelligence check - or factors that directly affected cognitive factors, such as hours slept, amount of time studied, and level of tutoring and class attendance. Physical activity was the only feature that wasn't directly related to cognitive performance in this subset. Conversely, our categorical columns ended up consisting of almost exclusively non-cognitive factors, such as extracurricular activities, motivation level, or teacher quality.

```
categorical_cols = ["Parental_Involvement", "Access_to_Resources", "Extracurricular_Activities",
                    "Internet_Access", "School_Type", "Motivation_Level", "Teacher_Quality",
                    "Family_Income", "Peer_Influence", "Learning_Disabilities", "Parental_Education_Level",
                    "Distance_from_Home", "Gender"]

numerical_cols = ["Hours_Studied", "Attendance", "Previous_Scores", "Sleep_Hours", "Tutoring_Sessions",
                  "Physical_Activity", 'Exam_Score']
```

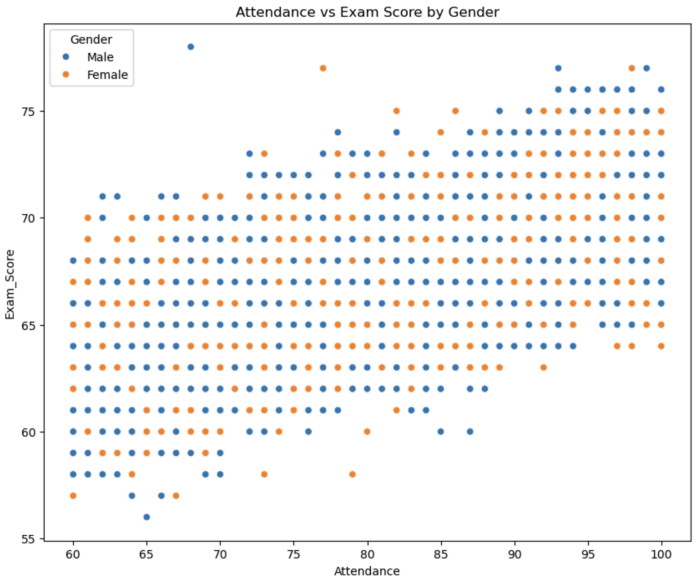
Visualizations played an important role in analyzing and understanding the relationships between different variables. A heatmap with all numeric variables was made to calculate correlation coefficients. Scatterplots for our two most correlated pairs - hours studied vs exam score and attendance vs exam score - were designed to clarify their relationship and linear dependence. A boxplot was made for every single categorical variable to gauge the 5-number summary and distribution of every single feature's different classes. A colored scatterplot was made to chart the correlation between exam score, previous score, and motivation level, which we designed not due to a high correlation coefficient, but because we believed the three to be directly connected.

Contingency tables were made for four pairs of variables - parental involvement vs access to resources, access to resources vs school type, extracurricular activities vs peer influence, and motivation level vs teacher quality - that the Chi-Square test deemed as not completely independent.

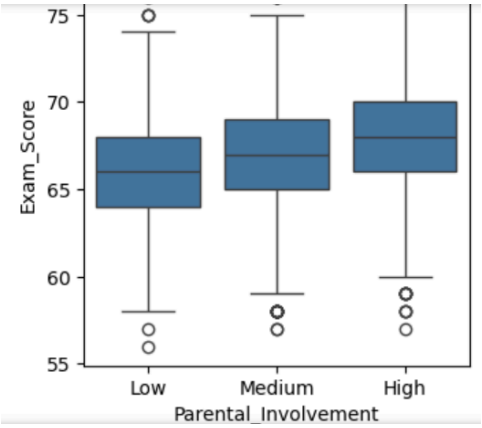
Following the visual analysis, the categorical columns were encoded using one-hot encoding, and two different linear regression models were created - one with only the numerical columns (to analyze how easy it would be to predict final grades given exclusively cognitive traits) and another with the numerical columns and the encoded categorical columns (to analyze how much non-cognitive traits improve or detract from model accuracy).

Results:

When analyzing the numerical variables in our data set by computing pairwise Pearson Correlation coefficients, we only identified two relationships that were moderately correlated. One statistically significant relationship was between Hours_Studied and Exam_Score, which had a Pearson Correlation coefficient of 0.51. The second and most statistically significant relationship we found was between Attendance and Exam_Score, which had a Pearson Correlation coefficient of 0.68. Our two scatterplots further helped visualize and depict the relationships we noted as statistically significant, both taking a roughly positive linear shape. One thing to note, however, was that the gender was more or less evenly distributed, which hinted that gender may not play a large role in impacting our target variable.



To understand the relationship between our categorical variables and our target variable, we employed numerous different visualizations, particularly box plots. From the box plot with Parental_Involvement and Exam_Score, we observed an incremental positive trend in the median exam scores as Parental_Involvement increased. When we looked at the box plot with Access_to_Resources and Exam_Score, we noticed that, similar to Parental_Involvement, the median exam scores were higher if a student had an increased access to resources. Our box plot with Extracurricular_Activities and Exam_Score showed a pretty level median between those who did participate in extracurricular activities and those that didn't. However, the upper quartile for students who participated in extracurricular activities was a little higher than students that didn't, which may signal a small positive link between the two variables. The box plot with Internet_Access and Exam_Score naturally showed a higher median for students who did have access compared to those who didn't. The box plot with School_Type showed virtually identical median exam scores in both Public and Private schools which definitely undermines the case of School_Type as a predictor of the target. Now in the case of Motivation_Level, our box plot showed virtually identical median exam scores for all levels of motivation. However, we observed that students with a high motivation level had a higher upper quartile than students with medium and low motivation levels. In the box plot for Teacher_Quality, we observed even median scores across all tiers of Teacher_Quality. However, we did notice that students with high quality teachers had a higher upper quartile than their peers with medium and low quality teachers. In the box plot with Family_Income, we observed that students with high family income had a higher median exam score and a higher upper quartile than their peers with low and medium family income. A sample boxplot is shown below.



When we conducted the Chi-Square test with an alpha value of 0.05 on all the possible pairs of categorical variables in our data set, we found 4 pairs for which we rejected the null hypothesis of independence. We made contingency tables for each in order to better gauge how they might be connected. While the chi-square test mathematically asserted that there exists a degree of dependence between all of these pairs of variables, it was difficult to actually understand the nature of this dependence when looking at the values in the contingency table.

We designed two different linear regression models to predict the target variable of Exam_Score for this dataset. For the first model, we chose to use only the numerical features of the dataset as input variables to our model. After splitting our data into a training and testing split, training our model on the testing split of the data, and testing our model's true accuracy on testing split of the data, our model was able to achieve an 81% training and testing accuracy with a mean squared error of 2.01. The proximity in the values of training accuracy and testing accuracy confirms to us that our model hasn't been overfitting and has actually generalized a broader relationship. For the second model, we used both numerical and categorical features and one-hot encoded the categorical features so we could plug them into the regression model. When we split our data and completed our training and testing phases, we were able to achieve a training accuracy of 98.6% and a testing accuracy of 99% with a mean squared error of just 0.1. As our model performs analogously on our training and testing data, we can rule out the possibility that our model is overfitting. The differences in accuracy between our first model with just numerical features and our second model which includes categorical features is a testament to the predictive power of the categorical variables on the target in this dataset.

Conclusion:

Our analysis of student performance using cognitive and non-cognitive factors has provided valuable insights into the key drivers of academic success. The results confirm that while cognitive traits like hours studied, previous scores, and class attendance have a direct and measurable impact on exam performance, non-cognitive factors such as motivation level, teacher quality, and parental involvement also play a crucial role. The linear regression models demonstrated that including non-cognitive factors rather than just focusing on cognitive traits

improved predictive accuracy, highlighting the importance of a balanced approach to academic support.

Our findings align with existing research that underscores the importance of self-discipline and motivation in academic outcomes, reinforcing the conclusions drawn by Duckworth and Seligman (2005) and Heckman and Kautz (2012). The Chi-Square tests and correlation analysis further clarified the complex interplay between different factors, with motivation level and teacher quality emerging as significant contributors to student success.

In practical terms, these results suggest that interventions aimed at improving academic performance should extend beyond cognitive skill development. Enhancing teacher-student relationships, fostering intrinsic motivation, and increasing parental involvement could lead to more comprehensive improvements in student outcomes. Future research could explore more granular influences within these categories and test alternative machine learning models to further refine predictive accuracy.

Sources:

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, 16(12), 939–944. https://doi.org/10.1111/j.1467-9280.2005.01641.x

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, 19(4), 451–464. https://doi.org/10.1016/j.labeco.2012.05.014

Jeynes, W. H. (2007). The relationship between parental involvement and academic achievement. *Urban Education*, 42(1), 82–110. https://doi.org/10.1177/0042085906293818