DSC530 Final Project

Jack Huffman

Master of Data Science, Bellevue University

DSC500-T301: Data Exploration and Analysis

Professor Matthew Metzger

February 26th, 2025

Final project

This project uses statistical analysis and machine learning techniques to investigate how external factors influence student exam performance. The StudentPerformanceFactors dataset includes five key variables that were selected: Hours studied, attendance, parental involvement, previous scores, and tutoring sessions, which were selected using F-score and mutual information (MI) scores to ensure the most meaningful variables were chosen. These variables were analyzed to determine their relationship with exam scores and their potential role in improving predictive models for educators.

Exploratory Data Analysis revealed that Hours Studied and Tutoring Sessions were right-skewed, while other variables followed a more uniform distribution. Outliers were identified but left in thed dataset due to time constraints. The Shapiro-Wilk test (p < 0.05) confirmed that exam scores are not normally distributed, influencing the proceeding statistical choices.

A PMF analysis compared high- and low-performing students based on study hours, revealing that higher-performing students’ study, on average, 10 hours more than lower-performing peers. The CDF showed that 75% of students study between 10 and 30 hours, highlighting central tendencies in study habits. ANOVA testing (p < 0.05) indicated that external factors significantly impact exam scores, leading to rejecting the null hypothesis.

Two predictive models, Linear Regression and LGBM—were trained to forecast exam performance. The Linear Regression model achieved an R-squared of 0.64 and an accuracy of 98.14%, demonstrating the strong predictive power of the selected features.

Future work will include developing a student recommendation system with the trained LGBM models, allowing the students to hyperparamter-tune their habits to achieve the best score. Also, expanding the dataset with additional variables such as socioeconomic data, and a larger student sample could further refine predictive accuracy and improve interventions for academic success.

Some challenges faced was the time constraints due to obligations at work I wasn’t able to explore this data as much as I would have liked. That said. I do look forward to continuing this project on my own time.