# Methods in Data Science: Student Performance

What is the correlation between students' past test scores, patterns involving parents' level of education, lunch type, and test preparation courses. And how can learning how these factors predict student test scores and future performance? This study investigates how various factors, including academic behaviors, previous grades, and demographic characteristics, influence students' writing scores. By analyzing the student performance dataset, the objective is to uncover patterns, trends, and correlations that can provide insights into academic success.

To conduct this analysis, I utilized a student performance dataset obtained from Kaggle, comprising 1000 observations with attributes such previous test scores, and demographic factors. Before going into the analysis, data cleaning and preprocessing were essential steps. This involved addressing missing values and duplicated rows ensuring the test set is accurate. Following the cleaning process, I focused on preparing the data for analysis and visualization. I transformed categorical variables into numerical formats and standardized the data where necessary. This preprocessing step facilitated more effective analysis and allowed for clearer visual representation of trends.

A key aspect of data science is the visualization of findings. By creating various plots and graphs, I was able to find trends and relationships within the data. For instance, I created multiple boxplots displaying the average tests score for the different levels of education the students' parents have. Next, I examined the correlations between writing scores and other variables, specifically focusing on previous grades. I

plotted the relationships between the student's math scores as well as reading with the writing scores by doing simple linear regression and finding the line of best fit, indicating that a majority of students performed well, on the writing if they performed well on the math or reading. The analysis revealed a strong positive correlation between writing scores and previous grades.

In addition to these findings, I utilized logistic regression to classify students based on their likelihood of passing. This analysis highlighted that math and reading scores were strong predictors of whether a student would pass, while demographic factors such as lunch type had minimal impact. The confusion matrix from this model indicated that the classification was largely accurate, with a high number of true positives and true negatives.

Overall, the analysis of the student performance dataset revealed several important insights. The strong correlations between previous grades and writing scores suggest that targeted interventions in these areas could enhance student outcomes. Furthermore, the logistic regression analysis provided a framework for predicting student success based on measurable academic behaviors. These findings underscore the importance of understanding the factors influencing student performance. While the results offer valuable insights, further research is recommended to explore additional variables and potential influences, such as classroom environment and teaching methods. By continuing to analyze this dataset and similar ones, educators can develop more effective strategies to support student achievement and foster a positive learning environment.