Group 4

Data Mining

Interim Project Report

For our data mining project, we're focusing on the Student Alcohol Consumption dataset, available on Kaggle [link]. This dataset contains information about Portuguese secondary school students enrolled in Math and Portuguese courses. It captures a wide range of variables covering academic performance and behavioural patterns, with a particular emphasis on students' drinking habits. The dataset includes 33 variables across 649 student records, offering insights into demographics, family background, study habits, and academic performance, making it a great candidate for in-depth exploratory data analysis (EDA). Among the key variables are G1, G2, and G3, which reflect the students' grades for the first, second, and final periods of the school year, respectively. These numeric grades range from 0 to 20 and indicate academic progress, with G3 being especially important since it shows overall performance by the end of the year. The dataset also tracks weekday (Dalc) and weekend (Walc) alcohol consumption, rated on a scale from 1 (very low) to 5 (very high). These variables allow us to explore possible connections between drinking habits and academic outcomes, as measured by G3.

The primary motivation behind this study is to explore whether students' alcohol consumption patterns—during the week (Dalc) and on weekends (Walc)—affect their academic performance, specifically their final grades (G3). Given the ongoing concerns about youth alcohol abuse and its potential to disrupt learning, gaining a deeper understanding of this relationship is important. Our analysis may yield insights that can inform educational policies or interventions aimed at promoting student health and academic success. We chose this dataset because of its rich blend of academic and behavioural data. In addition to alcohol consumption, it provides context through various other variables, such as age, sex, address type (urban/rural), family size (famsize), study time (studytime), absences, and family support (famsup), all of which can also influence academic outcomes.

Our analytical approach involves using Multiple Linear Regression, with G3 as the dependent variable. Independent variables will include alcohol consumption measures (Dalc and Walc), study habits (studytime), and demographic factors (age, sex, and family background). This method allows us to determine if a statistically significant link exists between alcohol use and academic performance while accounting for other important factors. We'll carefully select features to prevent multicollinearity and make sure the assumptions for regression analysis hold. We’ll evaluate our model’s overall fit using R-squared, which will show how much of the variance in G3 can be explained by our chosen variables. To measure prediction accuracy, we’ll rely on Mean Absolute Error (MAE) due to its ease of interpretation. To strengthen our analysis and avoid overfitting, we’ll use cross-validation techniques such as k-fold cross-validation.

In summary, our project aims to understand how lifestyle choices, particularly alcohol consumption, impact students’ academic performance. We intend our findings to shed more light on how these factors intersect, offering evidence that could help shape educational policies or interventions aimed at student well-being. By examining the connection between alcohol use and academic achievement, we aim to contribute to the broader conversation on how social and behavioural factors influence educational success.