Executive Summary

Goal:

This data science project aimed to explore and understand the factors influencing student success in the Portuguese subject. The primary goals were twofold: first, to build a predictive model for initial Portuguese grades ('G1.Port') using linear regression, and second, to develop a classification model categorizing students into five performance levels based on their 'G1.Port' grades.

Data Background:

The dataset encompassed a diverse set of features, including demographic, social, and school-related attributes collected through school reports and questionnaires. Noteworthy features included student grades, family background, parental education, and lifestyle choices. The dataset provided a rich foundation for investigating the intricate dynamics influencing academic outcomes in the Portuguese subject.

Approach Used:

The project followed a systematic approach, starting with comprehensive data preprocessing to handle values, drop unnecessary columns, and encode non-numeric features. Exploratory data analysis involved visualizing grade distributions, examining correlations, and creating a new categorical variable for classification. Linear regression was employed for grade prediction, and logistic regression was utilized for multi-level classification. Linear regression was applied for grade prediction, and logistic regression was employed for multi-level classification. Additionally, subsequent implementations of the Decision Tree and Random Forest models were investigated.

Results:

The limited predictive capabilities of the linear regression model, indicated by a mean squared error (MSE) of 0.798, led to the exploration of alternative approaches. Subsequent implementations of the Decision Tree and Random Forest models demonstrated substantial improvements. The Decision Tree model, after feature selection, achieved an accuracy of 77.48%, while the Random Forest model exhibited an impressive initial accuracy of 84.77%. These results underscore the effectiveness of decision tree-based methods in enhancing predictive performance.