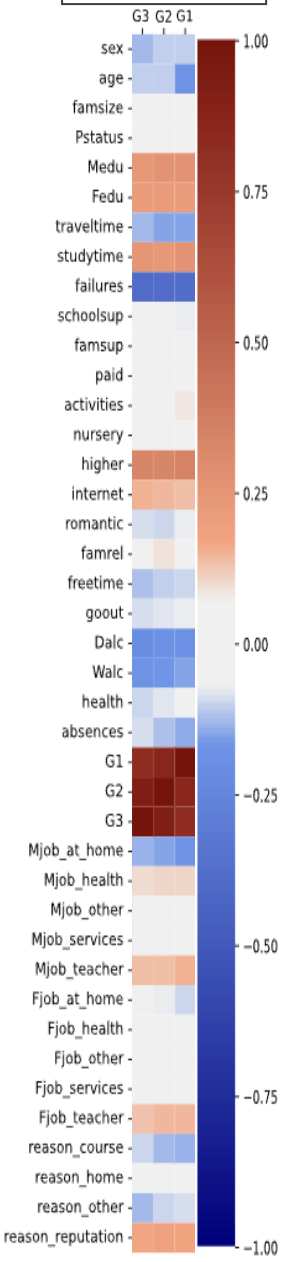
Capstone Project: Student Retention – “Who is at risk of dropping out?”

Kresna Laksono

The project aims to predict secondary school student performance based of a study conducted in Portugal by Paulo Cortez and Alice Silva (2008) academic paper. The original data includes two sets of 33 attributes that was gathered from official reports and supplementary student questionnaires covering mathematics grade perspective on one set and Portuguese language grade for the other. However, particularly for this project, only the latter dataset would be incorporated in due to it has more instances of 649 compared to 395 for the mathematics data; and also, time constraint reason.

Image 1. Heatmap of data attributes



The dataset comprises 33 attributes: 30 input features and 3 targets (G1, G2, G3). The input features include 11 ordinal integers, 2 general integers, 9 binary strings, and 8 categorical strings, while the targets are ordinal integers (0–20). G3, the primary target, determines passing status (≤10 is fail).

Upon gathering and cleaning the data, all the binary string attributes were transformed into binary integer data composed of 0 and 1, while a few of the 8 categorical strings were converted into binary integer as long as they consist of two components, and the others with more than 2 contents were being transformed utilizing one-hot encoding that broaden the values into binary integer for each of those components. In the exploratory data analysis phase, revealed that most features have low correlation with G3. However, study time, past class failures, and desire for higher education showed notable positive or negative correlations. G1 and G2 exhibited strong positive correlations with G3. Descriptively, pass-to-fail ratios were approximately 7:1 for females and 4:1 for males, with equal failure counts across genders.

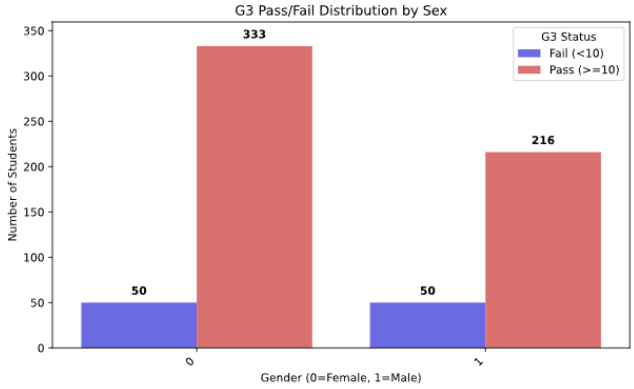


Image 2. Chart of Pass and Fail status based on gender

Initial Linear Regression using all 30 features yielded an R^2 score of 0.175. Feature selection, based on mean squared error (MSE), reduced the features to 16. Subsequent modeling improved R^2 to 0.298 (Linear Regression) and 0.258 (Random Forest Regressor). Despite these gains, the low R^2 scores suggest the models capture only a small portion of G3’s variance, indicating a need for better features or techniques. The project concludes that incorporating G1 and G2 is essential for predicting G3, as they represent prior grades and strongly correlate with the target. However, this high correlation may introduce multicollinearity, a challenge to address in future work.