Week 1 Student Performance Classification Report

Yuxuan Zhang

Westcliff University

AIT 506

Instructor: Noha Bhairy

Date: Nov 1, 2025

Student Performance Classification Report

The dataset has 649 rows and 33 columns. No rows have missing values so all records are kept. A target variable is created that is 1 if G3 is greater than or equal to 80% of the full score and 0 otherwise. In preprocessing, all categorical variables are one hot encoded. A 80-20 train test split is used. Fitting models with this setup and default parameter reported two issues. First, all models are overfitting to some degree. The random forest model overfit most severely. Second, due to the positive class being underrepresented, all models seem to favor the negative class. To address the first issue, we add regularization or/and tune parameters. To address the second issue, we oversample the positive class using SMOTE. Cross validation is used to find the best hyperparameters for each mode. Among all models, XGBoost achieved the best result, giving a 0.36 cross validation F1 score. The test set was used to evaluate the models, but due to the limited size, it is not used to measure generalizability. To that end, the cross validation result is favored.