Executive Summary: Predictive Modeling of Student Achievement in Secondary Education

Task/Goals: The primary goal of this project was to analyze and predict student achievement in secondary education, specifically focusing on Mathematics and Portuguese subjects in two Portuguese schools. The project aimed to understand the factors influencing success in these subjects and, subsequently, build predictive models for the final grades (G3) in Mathematics.

Data Background: The dataset used in this project consists of student grades, demographic features, social and school-related attributes from two Portuguese secondary schools. Collected through school reports and questionnaires, it encompasses various aspects of students' academic performance.

Approach/Methods Used: The project employed an exploratory data analysis (EDA) approach to understand the dataset's features, identify missing values, and examine variable types, and also the visualization part. The modeling tasks were divided into two parts:

1. Predictive Modeling for G3.Math: Here specific variables (health, failures.Math, absences.Port) were mandated in the predictive model. The model excluded certain variables (Medu, Mjob, Fedu). In this task I used Linear Regression, Lasso Regression, Ridge Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, and Gradient Boosting Regressor. Evaluation metrics (R2, MAE, MSE) were computed, and Random Forest Regression yielded the best results.

2. Multiclass Classification for G3.Math Category: Target variable G3.Math was binned into 5 categories. Similar variable restrictions as in Task 1 applied. Classification models (Logistic Regression, Gradient Boosting Classifier, Decision Tree Classifier, Random Forest Classifier, Gaussian Process Classifier) were employed. Evaluation focused on accuracy score, confusion matrices, AUC Score, and ROC plots. Decision Tree Classifier exhibited the highest accuracy (70%), and AUC scores ranged from 80%-83%.

Feature Selection: SelectKBest from sklearn was employed for feature selection, focusing on the top 10 variables. This streamlined models, emphasizing the most influential features for both predictive modeling and multiclass classification.

Hyperparameter Tuning: GridSearchCV was utilized to fine-tune hyperparameters in regression and classification models. This iterative process optimized model configurations, enhancing accuracy and performance.

Cross-Validation: K-fold cross-validation ensured robust model evaluation and mitigated overfitting. Implemented during hyperparameter tuning, it provided a reliable estimate of model performance, fostering generalization to unseen data and contributing to trustworthy results.

Results:

Task 1: Random Forest Regression performed the best with a 41.7% accuracy in predicting G3.Math.

Task 2: Decision Tree Classifier achieved the highest accuracy of 70% in classifying G3.Math categories. AUC scores indicated good model performance.

Conclusion: Despite a relatively small dataset and potential imbalance, the project successfully explored and modeled student achievement in secondary education. Noteworthy findings include the importance of specific variables in predictive modeling and the efficacy of Decision Tree Classifier for multiclass classification. Challenges were observed in feature significance, possibly attributed to limited data points. Overall, the project provides valuable insights into factors impacting student success and serves as a foundation for future data analysis projects.