

Executive Summary

The project focuses on understanding the factors influencing success in Mathematics and Portuguese subjects among secondary education students in two Portuguese schools. The analysis involves building predictive and classification models for G1.Math, addressing specific tasks and goals.

Tasks/Goals:

Task 1 involves constructing a predictive model for G1.Math without using other grade features, employing RandomForestRegressor. The model includes variables Dalc, famsup, absences.Port, higher, romantic, excluding address, sex, Medu.

Task 2 bins G1.Math into three categories, and a classification model is built using the same variables as Task 1.

Data Background:

The dataset encompasses student grades, demographic, social, and school-related features from two Portuguese schools, collected through school reports and questionnaires.

Approach/Methods Used:

For Task 1, RandomForestRegressor is employed with hyperparameter tuning using GridSearchCV. The model's performance is evaluated using RMSE.

For Task 2, G1.Math is categorized using qcut, and a RandomForestClassifier is used with hyperparameter tuning.

Results:

- Task 1 (Predictive Model for G1.Math) - RMSE: 0.9863
- Task 2 (Classification Model for Binned G1.Math) - Accuracy: 0.9316

Feature importance analysis used to understand the inner workings of the predictive and classification models, and to pinpoint the most influential factors that contribute to student success in secondary education.

Extras:

Additional models, Decision Tree and SVM, are also utilized for both tasks.

- Task 1 - Decision Tree RMSE: 1.1494
- Task 1 - SVM RMSE: 1.9214
- Task 2 - Decision Tree Accuracy: 0.9211
- Task 2 - SVM Accuracy: 0.8263

The RandomForest models show strong predictive and classification performance. Decision Tree and SVM models provide alternative perspectives, with Decision Tree exhibiting competitive accuracy. The project highlights the effectiveness of machine learning in understanding and predicting student success, offering insights for educational interventions. Further exploration and validation of the models on additional datasets could enhance their robustness and generalizability.