Data Analysis Report

Introduction

This report presents the key findings of a machine learning project to predict student final grades. The data included past grade performance and information about the student's family and personal life.

Data

Total students: 316
Target variable: G3 - final grade in math course in Portugal
Target grades are integers from 0 to 20, but handled as regression problem
3. Data Cleaning
Data set had no missing values and was clean upon import.
Scaling: Standard scaler applied to numerical values
Categorical encoding: One-hot encoding for categorical variables.

Exploratory Data Analysis (EDA)

Distributions of all numerical features and target were created.
Correlation was investigated using heatmap.

Feature Engineering

Created 5 new feature:

- Education - combined parents' individual education values
- Alcohol - combined reported daytime and weekend alcohol use
- Party - combine weekend alcohol and 'goout'
- Grades - combined previous two term scores

Model Building Strategy

Models used: Linear Regression, Linear Support Vector Regressor, Support Vector Regressor, Random Forest, and Gradient Boosting.
Training process: 75% training, 25% test.
Evaluation metrics: mean squared error, root mean squared error, mean absolute error, and R^2 score.

Model Performance

Random Forest and Gradient Boost were most reliable and generally performed best across all metrics. Models seemed to generalize better, but perform generally worse after parameter tuning.

Conclusions

Random Forest models performed best, followed by Gradient Boosted models. Only numerical features provided important signal, specifically G1, G2.
Recommendations: improve Random Forest and Gradient Boost parameter tuning, continued exploration of feature engineering