STUDENT PERFORMANCE PREDICTION – FINAL SUMMARY REPORT

Project Overview

This project was focused on predicting the final exam grades (G3) of students using their academic performance and personal background. The main objective was to analyze patterns and identify key factors that influenced student performance using machine learning.

Goals

- To analyze student performance data using data science techniques
- To identify features that affected final grades
- To train regression models to predict G3 accurately
- To use cross-validation and parameter tuning for improving model performance

Objectives

- The dataset (student-mat.csv) was collected from the UCI Machine Learning Repository
- The data was preprocessed by encoding categorical values and checking for nulls
- Exploratory Data Analysis (EDA) was performed to discover relationships between features and final grades
- Three regression models were trained: Linear Regression, Decision Tree, and Random Forest
- Models were evaluated using R^2, MAE, and RMSE
- GridSearchCV was used to tune the Random Forest model for better accuracy
- Feature importance was calculated from the best model to interpret impactful features

Approach

- The project followed a data science pipeline starting with data cleaning and preprocessing
- Label encoding was used to convert categorical variables into numeric form
- Visualization tools like seaborn and matplotlib were used to generate plots and heatmaps
- Linear, Decision Tree, and Random Forest regressors were trained and tested
- GridSearchCV with cross-validation was applied to the Random Forest model to tune n_estimators, max_depth, and min_samples_split
- The best Random Forest model had:
- n_estimators=100, max_depth=5, min_samples_split=10
- It achieved a cross-validated R^2 score of 0.854

Findings

- Previous grades (G1 and G2) showed a strong positive correlation with the final grade (G3).
- More study time was generally linked to better final exam scores.
- Students with past failures or higher absences tended to perform poorly.
- Weekend and weekday alcohol use showed a weak negative effect on grades.
- Health, internet access, and other background features had minimal impact on G3.
- G2, G1, studytime, and absences were identified as the most important predictors by the Random Forest model.