**Student Performance Prediction**

**Summary**

The goal of this project is to make a prediction about the last grade (G3) of the students based on academic, demographic, and behavioral attributes of the UCI Student Performance dataset. Two data sets were given: one math data set and one Portuguese data set. The two data sets were combined based on 13 identifying columns in order to identify overlapping students. Overlapping records' grades were averaged, and unique students from both data sets were included.

**Data Preprocessing**

Preprocessing included

* Mapping binary categorical values (yes/no, M/F, U/R, etc.) to 0/1.
* One-hot encoding nominal categorical features (e.g., school, Mjob, Fjob, reason, guardian).
* Standardizing numeric features to have zero mean and unit variance.
* Splitting the data into 80% training and 20% testing sets.

**Model Building**

Three regression models were evaluated:

1. Linear Regression
2. Decision Tree Regressor
3. Random Forest Regressor

Baseline performance was measured using 5-fold cross-validation on the training set. Hyperparameter tuning was performed using GridSearchCV for Decision Tree and Random Forest models.

**Model Evaluation**

Performance on the held-out test set:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **R2** | **MAE** | **RMSE** |
| **LinearRegression** | 0.9151 | 0.7133 | 0.9713 |
| **DecisionTree** | 0.9150 | 0.7526 | 0.9717 |
| **RandomForest** | 0.9217 | 0.6687 | 0.9328 |
|  |  |  |  |

**Feature Importance (Random Forest)**

Top 10 most important features influencing the prediction:

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| G2 | 0.8202 |
| absences | 0.0340 |
| G1 | 0.0187 |
| freetime | 0.0084 |
| goout | 0.0084 |
| failures | 0.0078 |
| age | 0.0077 |
| famrel | 0.0059 |
| Dalc | 0.0057 |
| Walc | 0.0054 |

**Conclusion**

The Random Forest model had the best R² value on the test set, meaning it accounts for more than 92% of the variance in final grades. The feature with the greatest predictive power was G2 (second period grade), followed by absences and G1 (first period grade). This implies that previous performance is the best predictor of final exam performance. Also, behavioral factors such as free time, social activity, and previous failures had lesser but significant contributions.