Table of contents

Predicting Student Final Grades Using Machine Learning

1. Introduction

Academic performance prediction is a crucial aspect of educational data analysis. This project aims to develop a machine learning model to predict students' final grades (G3) based on various academic and socio-economic factors. Using a dataset of student performance, we apply data preprocessing techniques and train a Random Forest Regressor to estimate final grades.

2. Dataset Description

The dataset used in this study is the "student-mat.csv" file, which contains student academic records with multiple categorical and numerical attributes. The key features include:

- Demographic information (e.g., age, sex, school)
- Academic performance indicators (e.g., past grades G1, G2)
- Family background (e.g., guardian, family support)
- Study habits and activities (e.g., study time, failures, extra activities)

The target variable for prediction is G3 (final grade).

3. Methodology

3.1 Data Preprocessing

- Loading the Dataset: The dataset is loaded using Pandas.
- Handling Categorical Data: Label Encoding is applied to convert categorical features into numerical values.
- Feature Selection: The feature set X consists of all columns except G3, which is the target variable y.

- Data Splitting: The dataset is split into 80% training and 20% testing sets.
- Feature Scaling: StandardScaler is applied to normalize numerical features.

3.2 Model Selection and Training

The Random Forest Regressor model is chosen due to its robustness in handling non-linear relationships and high-dimensional data. The model is trained using 100 decision trees with a fixed random state for reproducibility.

3.3 Model Evaluation Metrics

To assess model performance, the following metrics are calculated:
- Mean Absolute Error (MAE): Measures average absolute prediction error.
- Mean Squared Error (MSE): Evaluates prediction variance.
- R-squared (R^2) Score: Indicates the proportion of variance explained by the model.

4. Results and Analysis

4.1 Model Performance

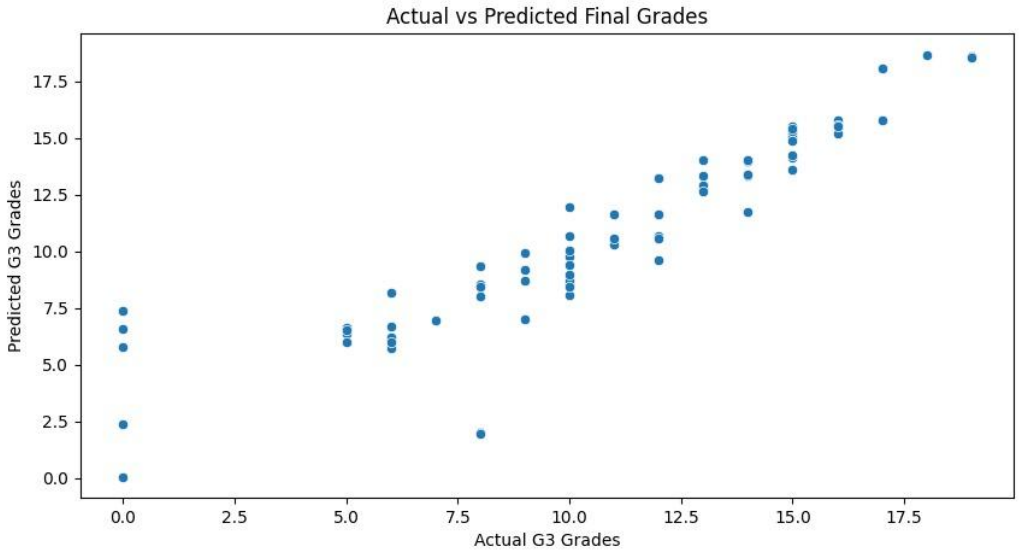
The trained model produced the following results:

- Mean Absolute Error (MAE): 1.1210126582278481
- Mean Squared Error (MSE): 3.5104506329113927
- R-squared Score (R^2): 0.8288006563935861

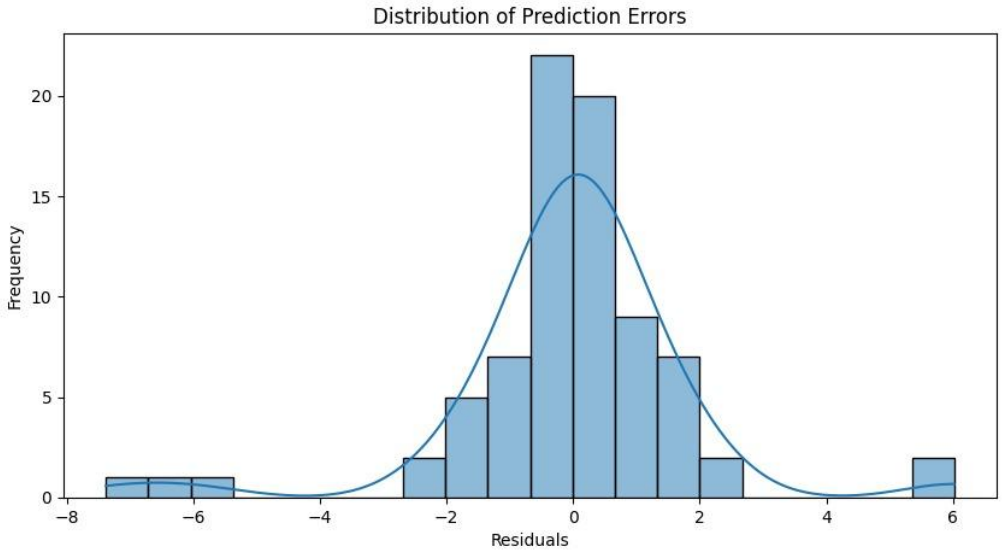
The R^2 score suggests the model's effectiveness in predicting final grades based on the available features.

4.2 Visualizations

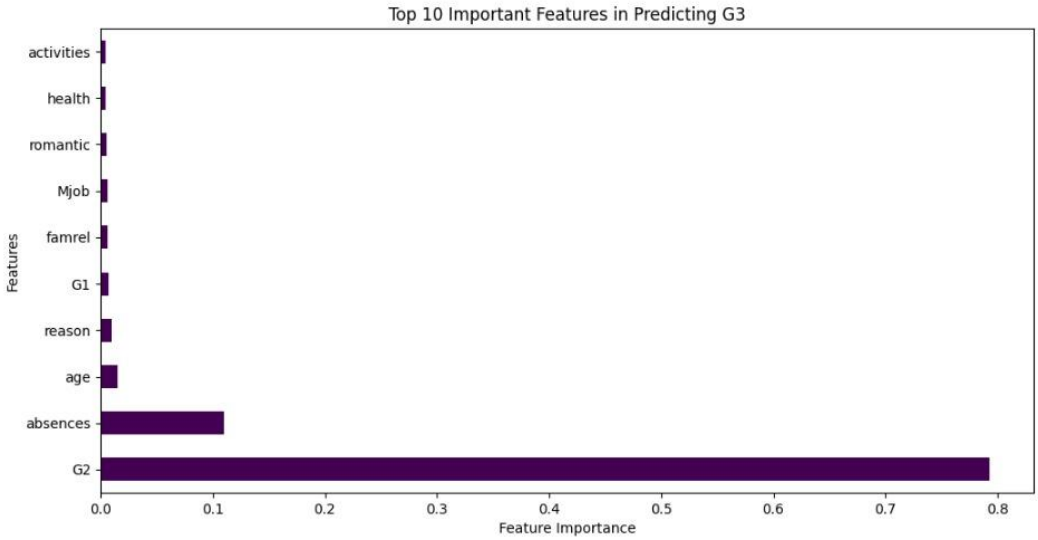
4.2.1 Actual vs. Predicted Grades



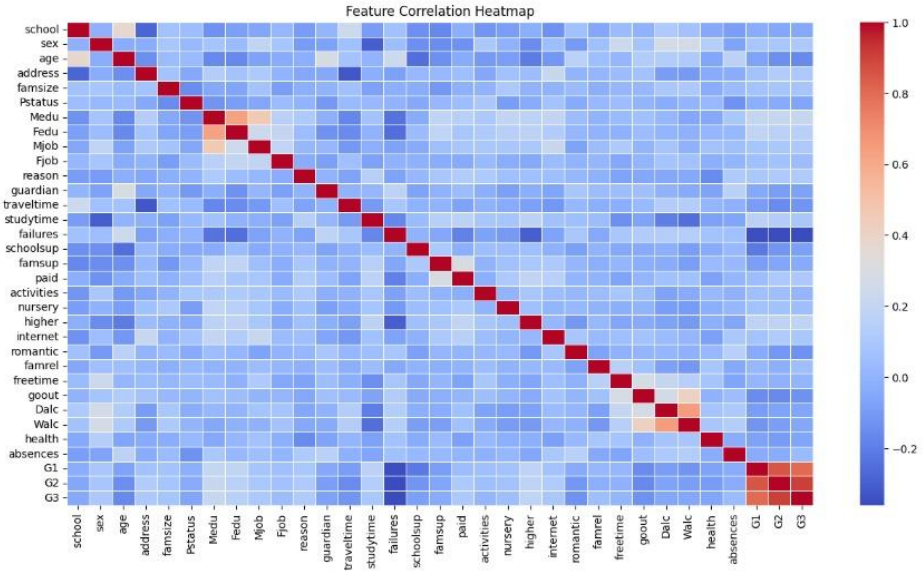
4.2.2 Residual Analysis



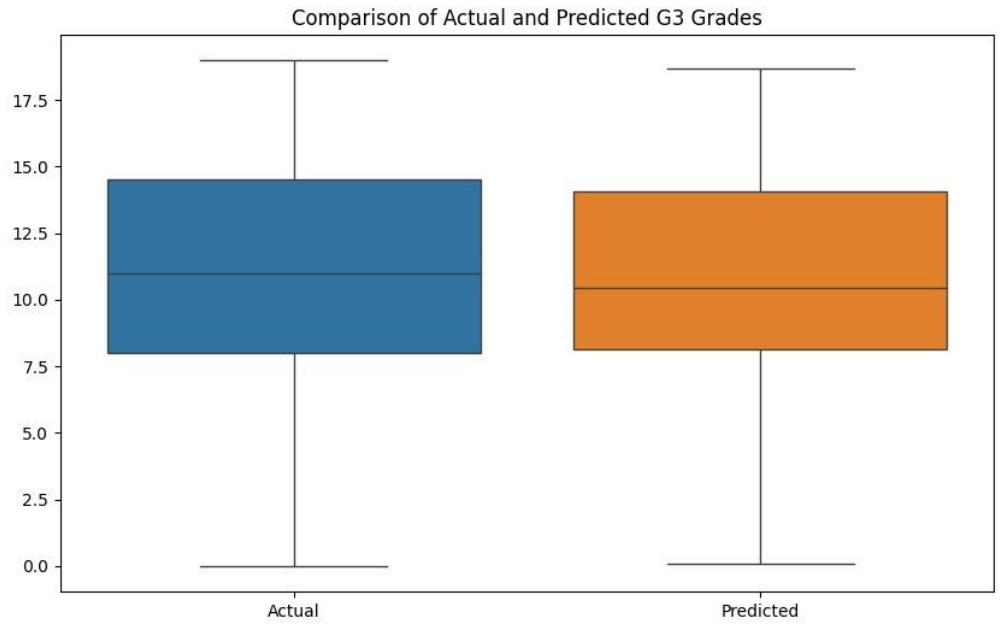
4.2.3 Feature Importance



4.2.4 Feature Correlation Heatmap



4.2.5 Boxplot of Actual vs. Predicted Grades



5. Conclusion

This project successfully developed a machine learning model for predicting student final grades. Key findings include:
- Past academic performance (G1, G2) significantly influences final grades.
- The Random Forest model demonstrated strong predictive capabilities with reasonable error margins.
- Further improvements could involve testing additional algorithms and hyperparameter tuning.

Future work could explore more advanced feature engineering techniques, additional datasets, or deep learning approaches for enhanced accuracy.

6. References

- Dataset Source: Student Performance Dataset
- Scikit-Learn Documentation
- Seaborn and Matplotlib for Visualization