

Predicting Student Final Grades Using Kaggle Data: A Comparison of Linear Regression and Random Forest Models

Ahmad Paturusi

December 19, 2025

Abstract

This study explores the prediction of student final grades (G3) in a math class by using socio-demographic as well as academic data from a Kaggle dataset. Two regression models (Linear Regression and Random Forest) were implemented with and without first and second period grades (G1 and G2) as features. Results show that models including G1 and G2 achieve significantly higher predictive accuracy, which indicates a strong continuity in academic performance.

1 Introduction

- **Problem statement:** Predicting student academic performance is important for early intervention and educational planning. This project aims to predict the final math grades (G3) of students using their background information and past period grades (G1 and G2).
- **Literature / References:** The dataset originates from educational data mining research by Cortez and Silva (2008), who explored factors influencing student performance. The data was later made available through the UCI Machine Learning Repository.

2 Model

- Two regression models were implemented:
 - **Linear Regression:** Assumes a linear relationship between features and target:

$$\hat{y} = \beta_0 + \sum_{i=1}^p \beta_i x_i \quad (1)$$

where \hat{y} is the predicted G3, x_i are features, and β_i are coefficients.

- **Random Forest Regressor:** An ensemble of decision trees that reduces overfitting and captures non-linear patterns:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(x) \quad (2)$$

where $T_b(x)$ are individual tree predictions and $B = 100$ trees were used.

- The target variable is the final grade G3 (0–20). Features include socio-demographic attributes and, in some experiments, prior grades G1 and G2.
- Data were split 80/20 into training and test sets, with categorical variables encoded via label encoding and one-hot encoding.

3 Methods

- The analysis was implemented in Python using several key libraries:
 - **Pandas and NumPy** for data loading, cleaning, and numerical operations
 - **scikit-learn** for preprocessing (LabelEncoder, train_test_split), model training (LinearRegression, RandomForestRegressor), and evaluation (R^2 , MAE, RMSE, cross-validation)
 - **Matplotlib** for generating visualizations and plots
- Two experimental settings were compared:
 - **With G1 and G2:** Including prior academic period grades.
 - **Without G1 and G2:** Using only socio-demographic and behavioral features.
- All code is modular (data_prep.py, train.py, visualizations.py) and outputs are saved for reproducibility in results.txt in my GitHub [clickhere](#).

4 Numerical Examples

- Table 1 summarizes model performance (from results.txt):

Model	R^2 Test	MAE Test	RMSE Test
Linear Regression (with G1 & G2)	0.7241	1.6467	2.3784
Random Forest (with G1 & G2)	0.8196	1.1664	1.9233
Linear Regression (without G1 & G2)	0.1415	3.3953	4.1957
Random Forest (without G1 & G2)	0.2733	3.0983	3.8601

- The results show that including G1 and G2 drastically improves prediction. It also shows Random Forest outperforming Linear Regression in both settings.
- In Figure 1, we see strong alignment between predicted and actual G3 when prior grades are included.
- Figure 2 shows that G1 and G2 are the most important features, followed by study time and absences.
- Without G1 and G2 (Figure 2), predictions are less accurate, indicating socio-demographic features alone are weak predictors.

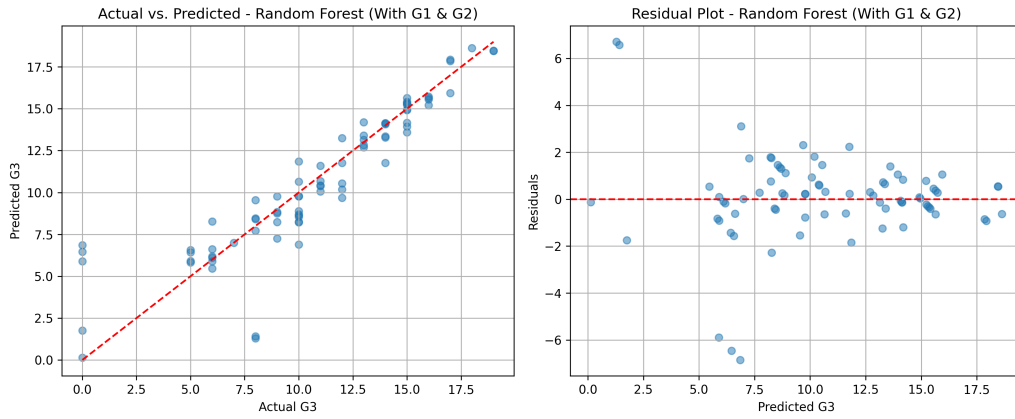


Figure 1: Random Forest predictions vs actual G3 (with G1 and G2). The model shows strong predictive capability.

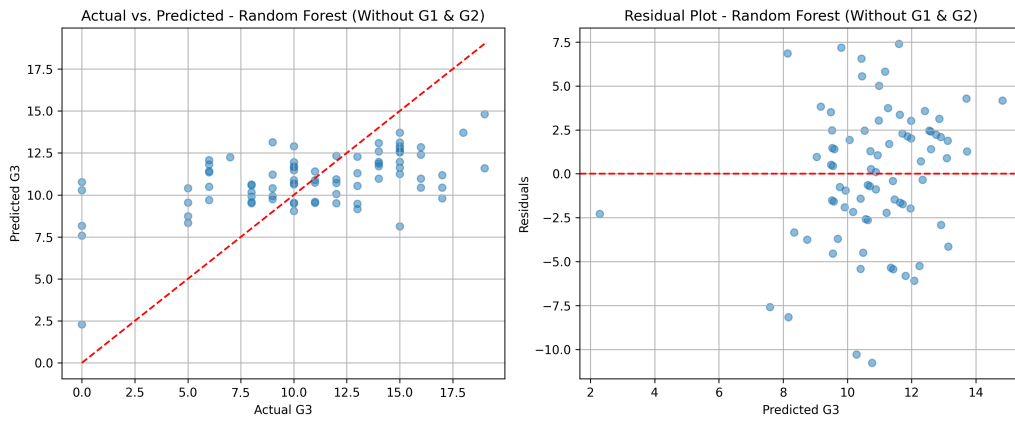


Figure 2: Random Forest predictions vs actual G3 (without G1 and G2). Predictions are scattered, indicating poor performance.

5 Conclusions

- **Usability of frameworks:** scikit-learn proved efficient and well-documented for this task. The modular Python scripts allowed reproducible experimentation.
- **Results obtained:** Prior academic performance in periods (G1, G2) are the strongest predictors of the final grade. Random Forest outperformed Linear Regression, especially when taking into account all the factors.
- **Future studies:** This approach could be extended further with neural networks or time-series models. The pipeline could also be applied to other educational datasets on Kaggle for broader insights into student success factors.

References

Cortez, P. and Silva, A. (2008). Using data mining to predict secondary school student performance. In Brito, A. and Teixeira, J., editors, *Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008)*, pages 5–12, Porto, Portugal. EUROSIS.