Predicting Student Performance Using Linear Regression

Aryan Nagaraj

August 2025

Abstract

This report explores predicting final student grades using Linear Regression. We employ three methods (Sklearn, Gradient Descent, Normal Equation) and conduct hypothesis tests (Pearson, t-test, ANOVA). Results indicate that previous grades are the strongest predictors of final academic performance.

1 Introduction

Predicting student academic outcomes is crucial for educational planning. We use the UCI "Student Performance" dataset to predict final grades (G3) from demographic, family, and academic features.

2 Methodology

2.1 Linear Regression Hypothesis

The hypothesis function is:

$$h_\theta(x) = \theta^T x$$

2.2 Cost Function

We minimize the mean squared error (MSE):

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

2.3 Gradient Descent

Parameters are updated iteratively:

$$\theta := \theta - \alpha \frac{\partial J(\theta)}{\partial \theta}$$

2.4 Normal Equation

Alternatively, we can directly compute:

$$\theta = (X^T X)^{-1} X^T y$$

3 Experiments

The dataset was preprocessed by encoding categorical variables and scaling numerical features. We trained models on an 80/20 train-test split.

3.1 Model Performance

Method	R^2	RMSE
Sklearn Linear Regression	0.780	2.122
Gradient Descent	0.780	2.122
Normal Equation	0.780	2.122

Table 1: Performance of Linear Regression models.

3.2 Top Features by Coefficients

Feature	Coefficient
G2	+0.978
failures	-0.416
famrel	+0.335
age	-0.198
Fedu	-0.188
G1	+0.161
goout	+0.138

Table 2: Top features ranked by coefficient size.

3.3 Figures

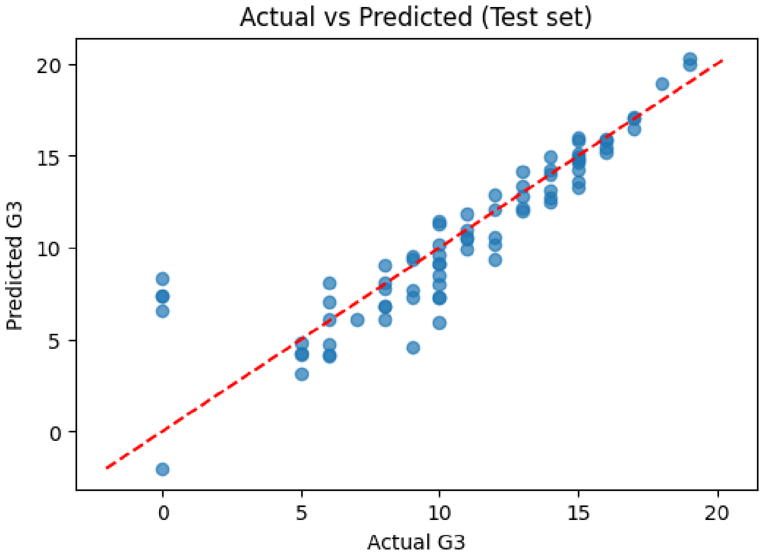

Figure 1: Predicted vs Actual student grades on the test set. The red dashed line represents the ideal case where prediction = actual.

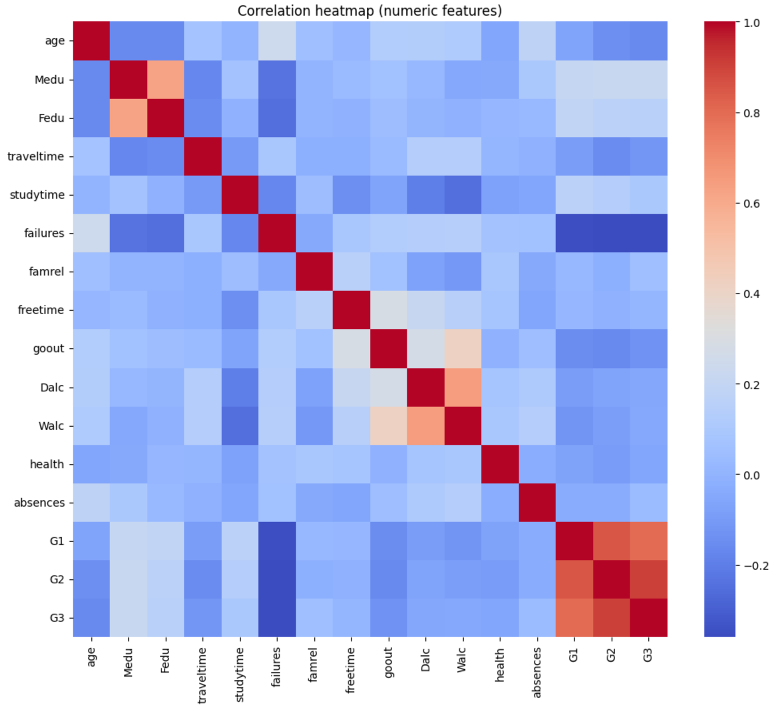

Figure 2: Correlation heatmap of numeric features. Strong correlation is observed between previous grades (G1, G2) and final grade (G3).

4 Hypothesis Testing

- **Pearson Correlation:** G2 strongly correlated with G3 ($r \approx 0.98$).
- **t-test:** No significant difference between male and female students.
- **ANOVA:** Studytime has a moderate effect on grades.

5 Conclusion

Linear Regression predicts student performance with good accuracy ($\mathrm{R}^2 = 0.78$). The strongest predictor of final grade is the previous grade (G2). Social and family factors also play a role.