National University
of computer and emerging sciences

Project Report

Predictive Modeling of Student Academic Performance Using Socioeconomic and Behavioral Indicators: A Linear Regression Approach

Submitted by:

Abdulmanan Arshad

22i-2391

Department of Computer Science BS(CS)

FAST-NUCES Islamabad

Abstract:

Predicting student academic performance is a critical task in educational research, enabling early interventions and informed policy decisions. This study applies linear regression models to forecast final grades in Mathematics and Portuguese subjects using the UCI Student Performance dataset. The dataset incorporates socioeconomic, academic, and psychological factors of students. Separate models were developed for each subject, and feature importance was analyzed to identify the most influential predictors. The Mathematics model achieved a R^2 score of 0.84, indicating a strong ability to explain variance in student performance, whereas the Portuguese model scored 0.46, suggesting moderate predictability. Results highlight that prior academic performance (G1 and G2) significantly influences final grades, while other variables such as study time, failures, and absences contribute marginally. The findings emphasize the relevance of academic continuity and suggest that Portuguese grades may be affected by additional non-linear or unobserved factors. This work demonstrates the utility of interpretable machine learning techniques in educational analytics and opens avenues for further exploration using advanced models.

Introduction:

In an era where data-driven insights are shaping education systems worldwide, predicting student academic performance has emerged as a critical area of research. Timely and accurate performance forecasting enables educators, institutions, and policymakers to design personalized learning plans, provide early interventions, and reduce dropout rates. Traditional assessments often fail to account for the complex interplay of socioeconomic, psychological, and behavioral factors that influence a student's academic journey.

Machine learning, particularly regression-based models, offers an opportunity to uncover hidden patterns within educational data. Linear regression is one of the most interpretable models, making it ideal for understanding how specific variables impact academic outcomes. While deep learning models may achieve higher predictive accuracy, their black-box nature limits transparency. In contrast, linear regression provides clear insights into feature importance, which is essential in educational settings where explainability is a priority.

This study uses the well-known UCI Student Performance dataset to explore and model student grades in Mathematics and Portuguese. It focuses on understanding the relative importance of various factors—such as parental education, study time, alcohol consumption, and prior grades—in predicting final academic success.

Project Overview:

Objective

The goal of this project is to:

- Predict the final grade (G3) of students in Mathematics and Portuguese using linear regression.
- Identify and compare the key factors that influence student performance in each subject.
- Evaluate and interpret the predictive power of the model using performance metrics such as R^2 score and mean squared error (MSE).
- Visualize feature importance to provide interpretable insights for educators and researchers.

Dataset

The project uses the UCI Student Performance dataset, which includes two subsets:

- `student-mat.csv` – academic records of students in a Mathematics course
- `student-por.csv` – records of students in a Portuguese course

Each subset contains 33 attributes covering:

- Demographic information (age, sex, family size)
- Academic background (G1, G2, study time, failures)
- Socioeconomic conditions (parental job, education, internet access)
- Behavioral patterns (alcohol consumption, extracurricular activities)

A third merged dataset was also created to examine students who took both subjects.

Tools and Methods

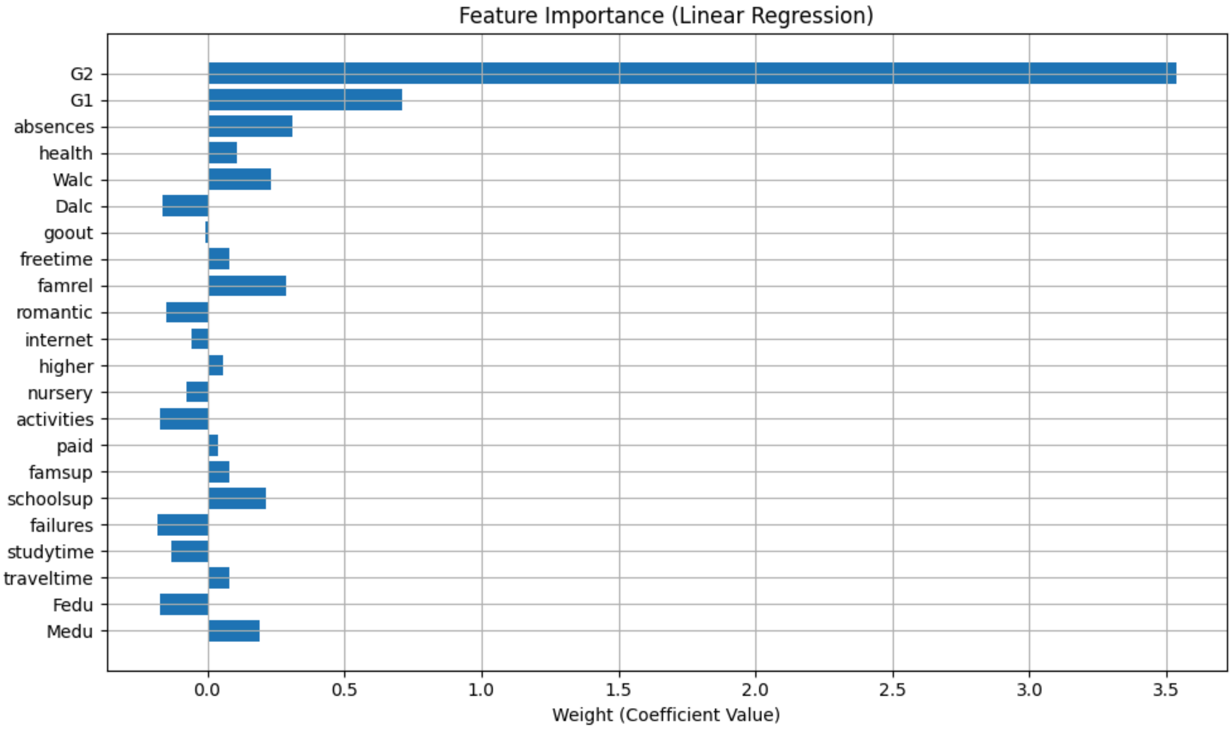
- **Programming Language**: Python
- **Libraries**: pandas, NumPy, scikit-learn, matplotlib
- **Models Used**: Custom implementation of Linear Regression using Gradient Descent and comparison with scikit-learn's LinearRegression
- **Preprocessing**: Feature encoding, normalization/standardization
- **Evaluation Metrics**: R^2 Score, Mean Squared Error (MSE)
- **Visualization**: Scatter plots, bar charts for feature importance, cost convergence plots

Results:

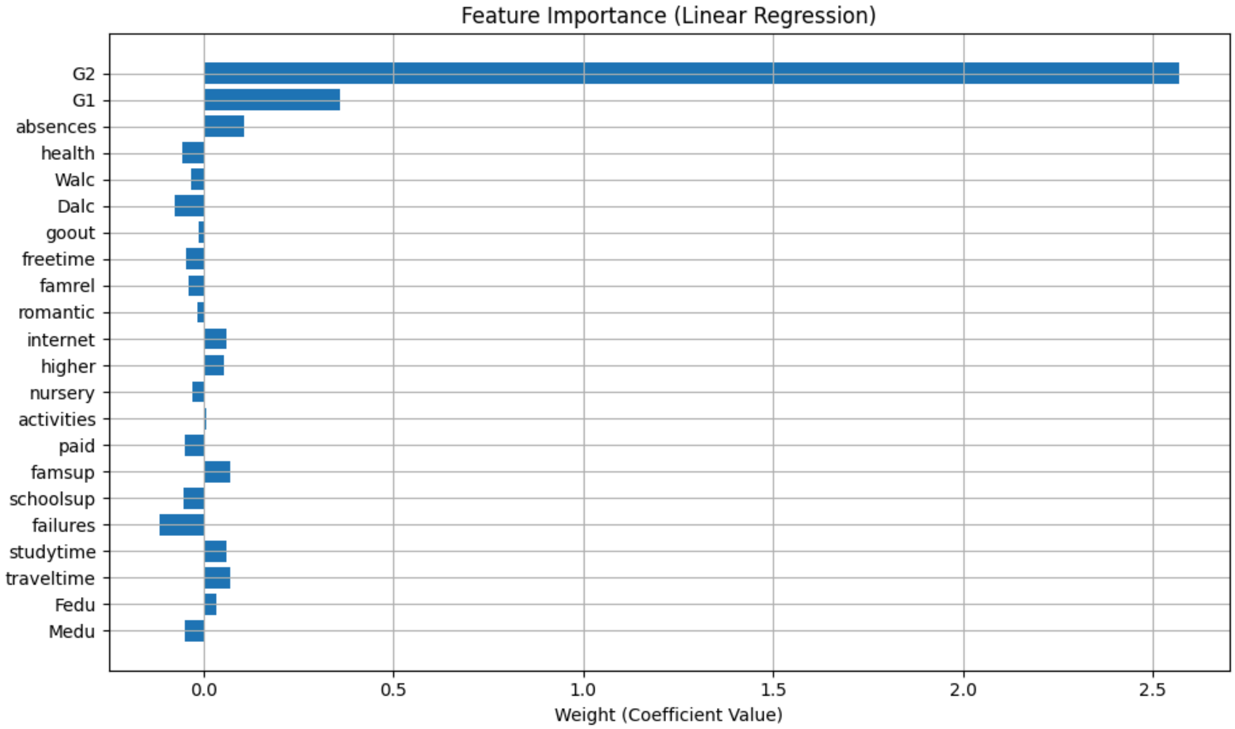
	R² Score	Interpretation
Maths	0.84	Excellent – explains 84% of variance in grades
Portuguese	0.46	Moderate – explains 46% of variance in grades

The Math model significantly outperforms the Portuguese model, suggesting that student performance in Mathematics is more strongly correlated with the available features.

Feature Importance in predicting maths grade:



Feature Importance in predicting portuguese grade:



Conclusion:

This study explored the use of linear regression models to predict students' final academic performance in Mathematics and Portuguese subjects using the UCI Student Performance dataset. By leveraging a combination of academic history, behavioral indicators, and socioeconomic features, we built interpretable models capable of identifying the key factors that influence student success.

The results show that the Mathematics model performed significantly better than the Portuguese model, achieving an R^2 score of 0.84 compared to 0.46. This suggests that performance in Mathematics is more directly influenced by the available features, particularly prior grades (G1 and G2), absences, failures, and study time. In contrast, Portuguese performance appears to be affected by more complex or unobserved variables that are not fully captured by a linear model.

Feature importance analysis reinforced the idea that early academic performance is the strongest predictor of final grades, emphasizing the value of monitoring student progress throughout the semester. Additionally, the successful convergence of the custom gradient descent implementation validated its correctness and alignment with scikit-learn’s regression outcomes.

In conclusion, this project demonstrates the feasibility and utility of linear regression as a transparent, interpretable, and effective tool for academic performance prediction. It also highlights the limitations of linear models in capturing more subtle or nonlinear relationships in certain subjects, paving the way for future work using more advanced machine learning algorithms.