Predicting Students' GPA

with

Linear Regression

Research Report

By Kristian Rahnev

1. Introduction

1.1 Problem Description

Understanding the factors that influence academic performance is a critical area of focus in educational research. Grade Point Average (GPA) is widely recognized as a key indicator of student success, often influencing opportunities for further education and career prospects. Identifying how various factors such as study habits, attendance, and support systems affect GPA can help educators design targeted interventions to improve academic outcomes.

This study aims to predict students' GPA using linear regression analysis by examining the influence of selected academic and support related features. The goal is not only to develop a predictive model but also to analyze the strength and direction of the relationships between these features and GPA.

1.2 Description of the Chosen Model

The model chosen for this study is Linear Regression, a fundamental statistical technique used to predict the relationship between a dependent variable and one or more independent variables. Linear regression is appropriate for this analysis because the target variable, GPA, is continuous, and the goal is to quantify how changes in independent variables impact GPA.

The model operates under the assumption of a linear relationship between the features and the target variable. By calculating the coefficients of each feature, linear regression provides insights into the extent to which each factor contributes to GPA. Positive coefficients indicate a positive relationship, while negative coefficients suggest an inverse relationship.

2. Dataset Description

The dataset used in this study consists of 2,392 student records with 15 features that capture a range of academic and personal attributes. The dataset has a usability index of 10 on Kaggle, indicating its high quality and ease of use for analysis. Additionally, all variable fileds are complete, with no missing values, ensuring a clean dataset for modeling. Due to its well structured nature,

almost no preprocessing is needed, making it well suited for direct application in linear regression without additional encoding steps. The data includes both numerical and binary categorical features, ensuring compatibility with the chosen modeling approach.

Features Overview

Feature	Type	Description	Potential Impact
1. StudentID	int64	Unique identifier for each student.	Acts as an identifier; not useful for predictive modeling.
2. Age	int64	Age of the student (ranges from 15 to 18).	Could influence GPA, as maturity might correlate with performance.
3. Gender	int64 (Encoded; 0 = Female, 1 = Male)	Indicates the gender of the student.	Useful for demographic analysis; it may or may not affect GPA.
4. Ethnicity	int64 (Encoded categories)	Coded representation of the student's ethnicity.	Could be used to study performance across different groups, but handle with care to avoid biases.
5. ParentalEducation	int64 (Encoded from 0 to 4)	Educational level of the parents (higher values likely indicate higher education levels).	Parental education often correlates with student academic success.
6. StudyTimeWeekly	float64	Total hours spent studying per week.	Strong potential correlation with GPA; key predictor.
7. Absences	int64	Total number of classes missed.	More absences could negatively affect GPA.
8. Tutoring	int64 (0 = No, 1 = Yes)	Indicates if the student receives tutoring.	May have a positive effect on GPA depending on tutoring quality.
9. ParentalSupport	int64 (Scale 0-4)	Level of parental support the student receives (higher = more support).	Higher support might contribute to better academic performance.
10. Extracurricular	int64 (0 = No, 1 = Yes)	Participation in extracurricular activities.	Could either positively influence (through skill development) or negatively (if time conflicts with studies).
11. Sports	int64 (0 = No, 1 = Yes)	Indicates participation in sports.	Similar to extracurricular activities can foster discipline or create time management challenges.
12. Music	int64 (0 = No, 1 = Yes)	Participation in music-related activities.	Might enhance cognitive abilities or take time away from academics.
13. Volunteering	int64 (0 = No, 1 = Yes)	Indicates involvement in volunteer work.	Could show positive personality traits but may have minimal direct effect on GPA.

14. GPA	float64 (Range: 0.0 to 4.0)	Grade Point Average, a key measure of academic performance.	This is the variable we want to predict.
15. GradeClass	float64 (Encoded categories)	Categorized grade class (e.g., 1 = Freshman, 4 = Senior, depending on the encoding).	

3. Exploratory Data Analysis (EDA)

3.1 Understanding the Data

The GPA values range from 0.0 to 4.0, and features like StudyTimeWeekly and Absences display significant variation, suggesting potential influence on the target variable.

The count plot on *Figure 1* of Ethnicity reveals a significant class imbalance, with Ethnicity 0 being the most represented group, while Ethnicities 1 and 2 have much lower but similar representation, and Ethnicity 3 is the least represented. This imbalance may introduce bias in predictive modeling, as the model could become skewed toward trends observed in Ethnicity 0, potentially underperforming for less-represented groups.

If the dataset is used for academic performance prediction, it is crucial to examine whether Ethnicity correlates with GPA or other performance indicators. Further analysis, such as boxplots and mean comparisons, can help explore these differences. To address the imbalance, resampling techniques such as undersampling Ethnicity 0 or oversampling Ethnicity 3 may be necessary to create a more balanced distribution.

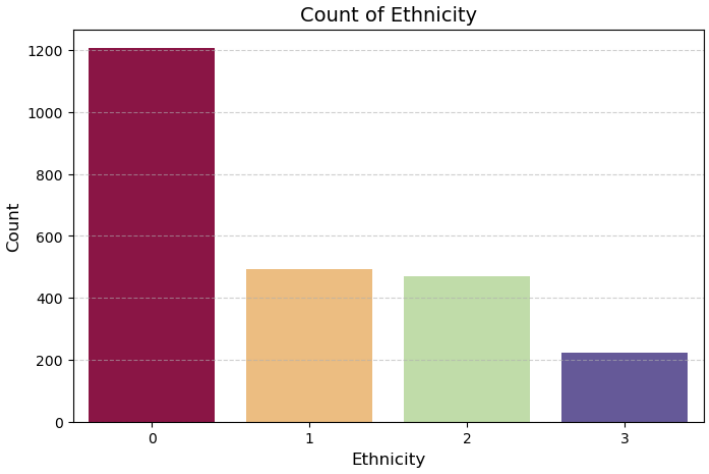

Figure 1: Ethnicity Distribution Count Plot
The figure illustrates the number of students in each ethnic group.

The count plot on *Figure 2* of Gender indicates a relatively balanced distribution, with both gender categories having nearly equal representation. This balance reduces the likelihood of gender based bias in predictive modeling, ensuring that the model does not favor one group over the other.

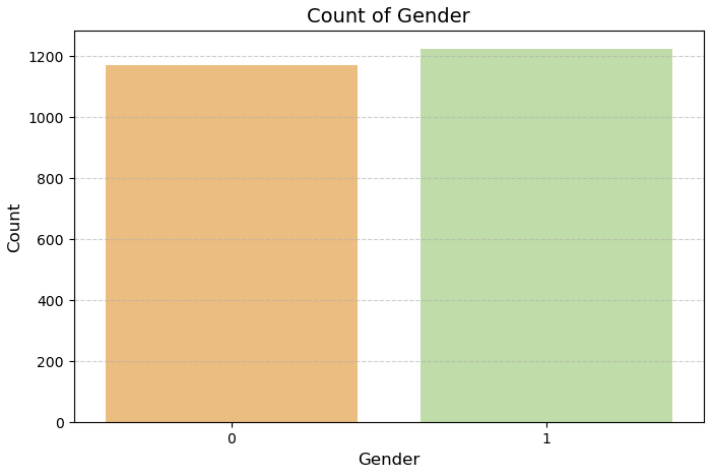

Figure 2: Gender Distribution Count Plot
The figure illustrates the number of students with each gender.

The boxplot on *Figure 3* illustrates the impact of tutoring on GPA, showing a slightly higher median GPA for students who receive tutoring compared to those who do not. The interquartile range (IQR) for both groups is similar, indicating comparable variability in GPA scores. However, the presence of overlapping distributions suggests that while tutoring may have a positive influence, its effect is not highly pronounced. Outliers are present in both groups, indicating some extreme GPA values.

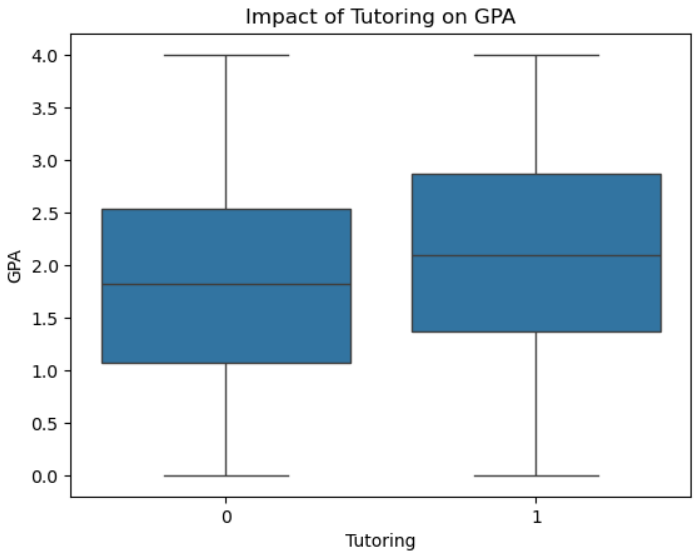

Figure 3: Impact of Tutoring on GPA
The boxplot compares GPA distributions between students who receive tutoring (1) and those who do not (0).

3.2 Correlation Matrix

To assess the relationships between features and GPA, a correlation matrix on *Figure 4* was generated. The Pearson correlation coefficient was chosen as it measures the linear relationship between continuous variables and is well suited for this dataset, which primarily consists of numerical features. This matrix highlights the strength and direction of linear relationships between variables

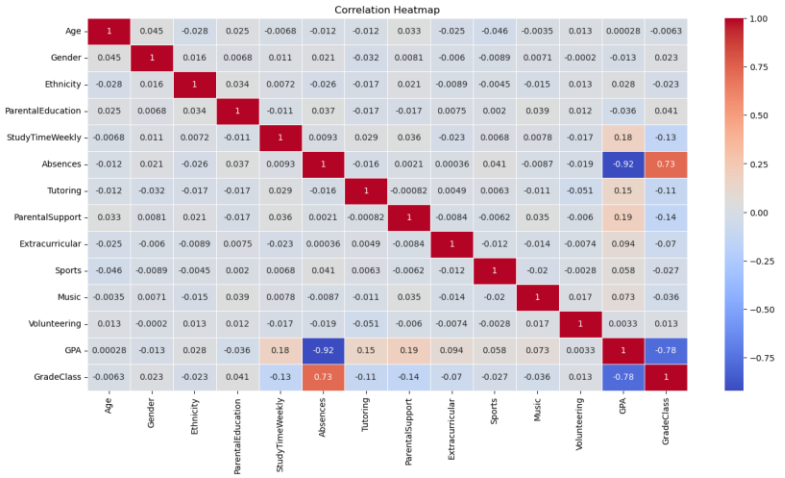

Figure 4: Correlation Heatmap of Dataset Features
This heatmap visualizes correlation coefficients between features in the dataset. Darker red shades indicate strong positive correlations, while darker blue shades represent.

Key observations from the correlation matrix include:

- Absences show a strong negative correlation with GPA (-0.92), suggesting that higher absenteeism significantly lowers academic performance.
- StudyTimeWeekly and ParentalSupport have weak positive correlations with GPA (0.18 and 0.19, respectively), hinting a mild improvement in GPA with increased study time and parental involvement.
- Tutoring and Extracurricular activities exhibit moderate positive correlations with GPA, indicating that structured support systems may enhance academic outcomes.

3.3 Correlation Coefficient Interpretation

The correlation coefficients reveal that while some factors like Absences have a substantial impact on GPA, others such as StudyTimeWeekly and ParentalSupport contribute positively but less significantly. These insights guide the feature selection process and inform expectations about each variable's influence in the linear regression model.

4. Model Implementation

4.1 Model Parameters and Training

The linear regression model was implemented using scikit-learn and the following features were chosen as independent variables:

- **StudyTimeWeekly**
- **Absences**
- **Tutoring**
- **Extracurricular**
- **ParentalSupport**

These features were selected based on their correlation with GPA and their relevance to academic performance.

The dataset was split into 80% training data and 20% test data, ensuring that the model generalizes well to unseen data.

4.2 Model Evaluation

The model's performance was assessed using the following metrics:

- Mean Squared Error (MSE): A value of 0.0506 suggests minimal difference between actual and prediced GPA.
- R-squared (R^2) Score: Indicates the proportion of variance in GPA explained by the model. A value of 0.9387 suggests a strong fit.

Additionally, K-fold cross-validation was employed to ensure a more robust evaluation of the model. The number of folds was set to **5** as a balanced choice to provide sufficient validation while maintaining computational efficiency. The results from a 5-fold cross-validation are as follows:

- R^2 Scores for each fold: [0.9387, 0.9419, 0.9359, 0.9409, 0.9440]
- Mean R^2 Score from Cross-Validation: 0.9403

A scatter plot on *Figure 5* of actual vs. predicted GPA was generated to visualize prediction accuracy, with a reference line indicating perfect predictions. The results confirm that the model effectively captures trends in academic performance, with Absences showing the strongest negative impact on GPA.

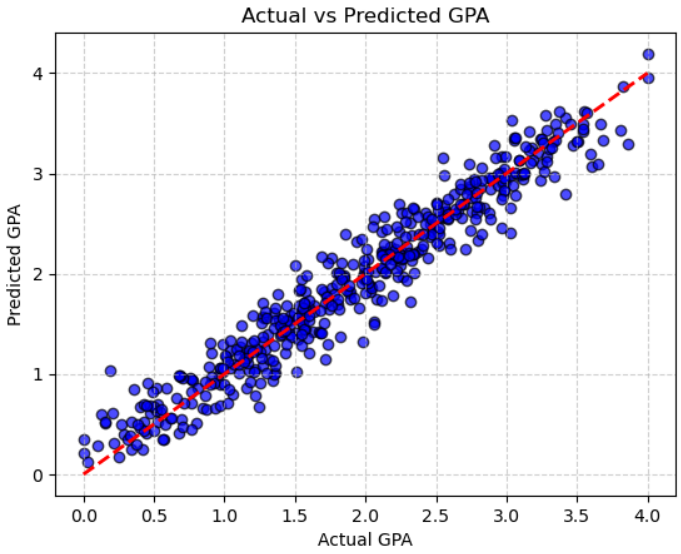

Figure 5: Actual vs. Predicted GPA
This scatter plot visualizes the relationship between actual and predicted GPA values from the linear regression model.

5. Conclusion

The linear regression model demonstrated strong predictive performance.

The findings confirm that Absences have the most significant negative impact on GPA, while StudyTimeWeekly, Tutoring, Extracurricular Activities, and ParentalSupport contribute positively.

Nevertheless, the study has certain limitations. The assumption of linear relationships may not fully capture complex interactions affecting GPA. Future research could explore non-linear models such as decision trees, random forests, or deep learning techniques to improve predictive accuracy. Additionally, integrating more diverse features, such as socioeconomic background, motivation levels, or teacher influence, could enhance the model's robustness.

6. References

IBM (n.d.) 'What is Linear Regression?', available at: https://www.ibm.com/think/topics/linear-regression (Accessed: [10.01.2025]).

StatQuest with Josh Starmer (2020) *Linear Regression, Clearly Explained!!!* [YouTube video], available at: https://www.youtube.com/watch?v=nk2CQITm_eo (Accessed: [20.12.2024]).

Analytics Vidhya (2021) 'Linear Regression: A Comprehensive Guide', available at: https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/ (Accessed: [26.01.2025]).

Kaggle (2024) 'Student Performance Dataset', available at: https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset/data (Accessed: [13.02.2025]).

Scikit-learn what 'Cross-validation: evaluating estimator performance', available at: https://scikit-learn.org/stable/modules/cross_validation.html (Accessed: [14.02.2025]).

Pure Storage (n.d.) 'What Are Machine Learning Performance Metrics?', available at: https://www.purestorage.com/knowledge/machine-learning-performance-metrics.html (Accessed: [02.01.2025]).