Final Report DSA210 Term Project: Factors-Affecting Students Academic Performance

Berat Kumru-31906

DSA210-Fall2026-Term Project

1. Motivation

As a university student, I often wonder why some students get high grades with seemingly less effort, while others struggle despite having similar resources. I have friends who study all night but get lower grades than those who just attend classes regularly.

This observation led me to the main question of this project: "Is a student's success determined more by their daily habits (like going to class and studying) or by their background (like ethnicity and parents' education)?"

My goal was to use the data science skills I learned in class to see if I could predict a student's GPA based on these factors.

2. Datasets & Data Enrichment

To answer this question, I used a mix of public data and data I collected myself.

- Primary Dataset: I used the "Student Performance in Exams" dataset from Kaggle. This gave me information on 2,392 students, including their gender, ethnicity, parental education, absences, study time, and GPA.
- Enrichment Dataset (Self-Collected): The Kaggle dataset didn't have personal details like sleep or stress. So, to make the project more original, I collected data on Sleep Duration, Stress Level, and Motivation Level. I tracked this for a small group (N=20) using daily productivity apps.

3. Data Collection & Preparation

Before starting the analysis, I had to clean and prepare the data using Python.

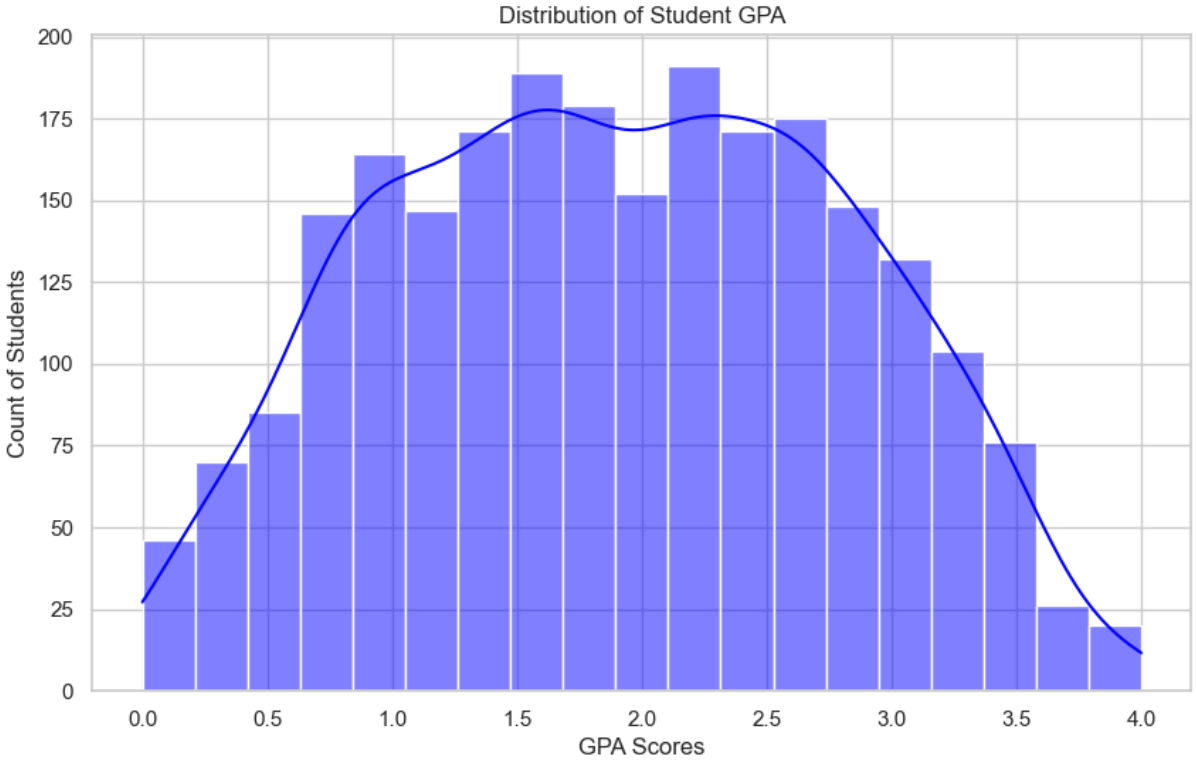
1. Cleaning: I checked the data for any missing values or weird errors. I also calculated the overall GPA by averaging the math, reading, and writing scores.
2. Merging: I merged my self-collected data with the main Kaggle dataset using StudentID so I could analyze everything together.
3. Encoding: Since computers can't understand words like "Male/Female" or "Group A/B", I converted these categorical variables into numbers so the machine learning models could use them.

4. Exploratory Data Analysis (EDA)

I created several visualizations to understand the patterns in the data.

4.1 Summary Statistics

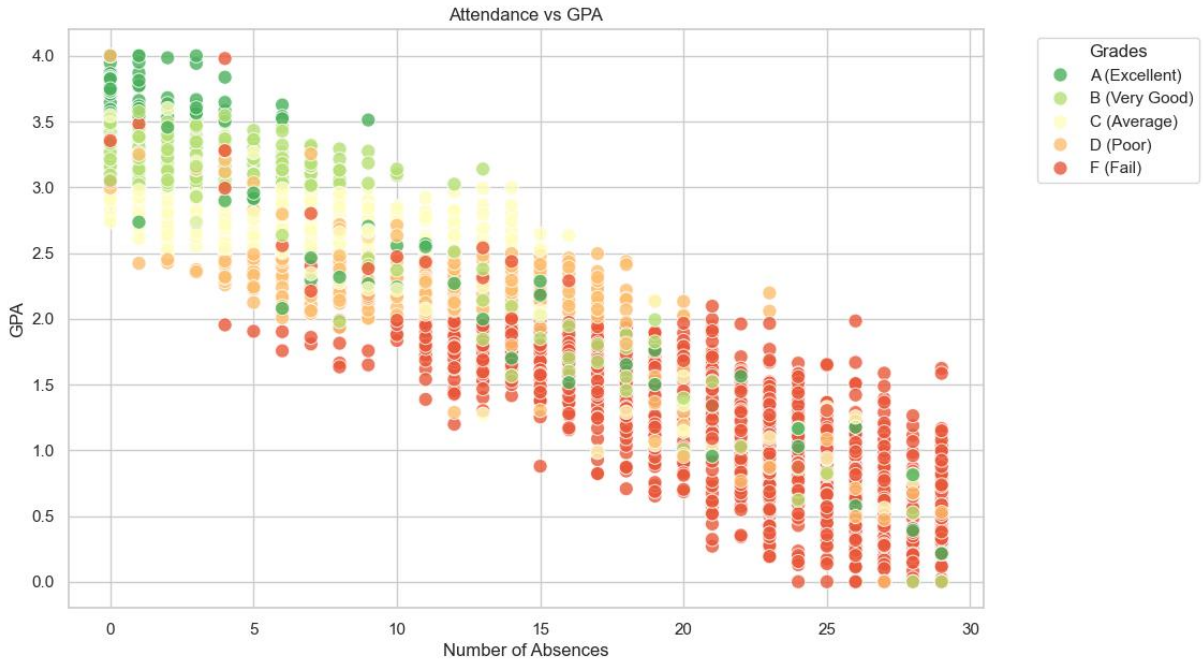
I plotted a histogram of the GPAs. It showed a "Bell Curve" (Normal Distribution), which means most students have average grades, and there are fewer students with very high or very low grades. This is a sign of a healthy dataset.



4.2 Attendance and Performance

I created a scatter plot to see the relationship between Absences and GPA.

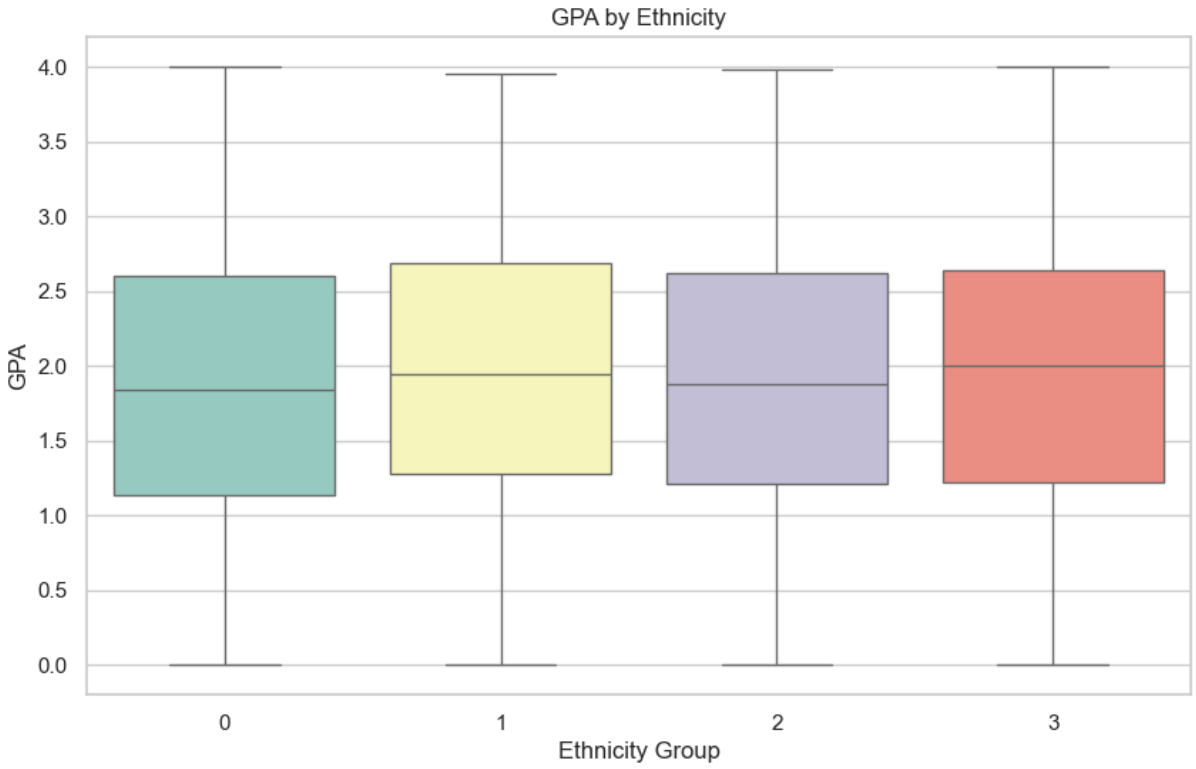
- What I saw: There is a very clear negative link. As students miss more classes, their GPA drops drastically.
- Cluster: Students with the highest grades are all clustered in the "low absence" area.



4.3 Demographic Analysis

I used boxplots to see if Ethnicity affects GPA.

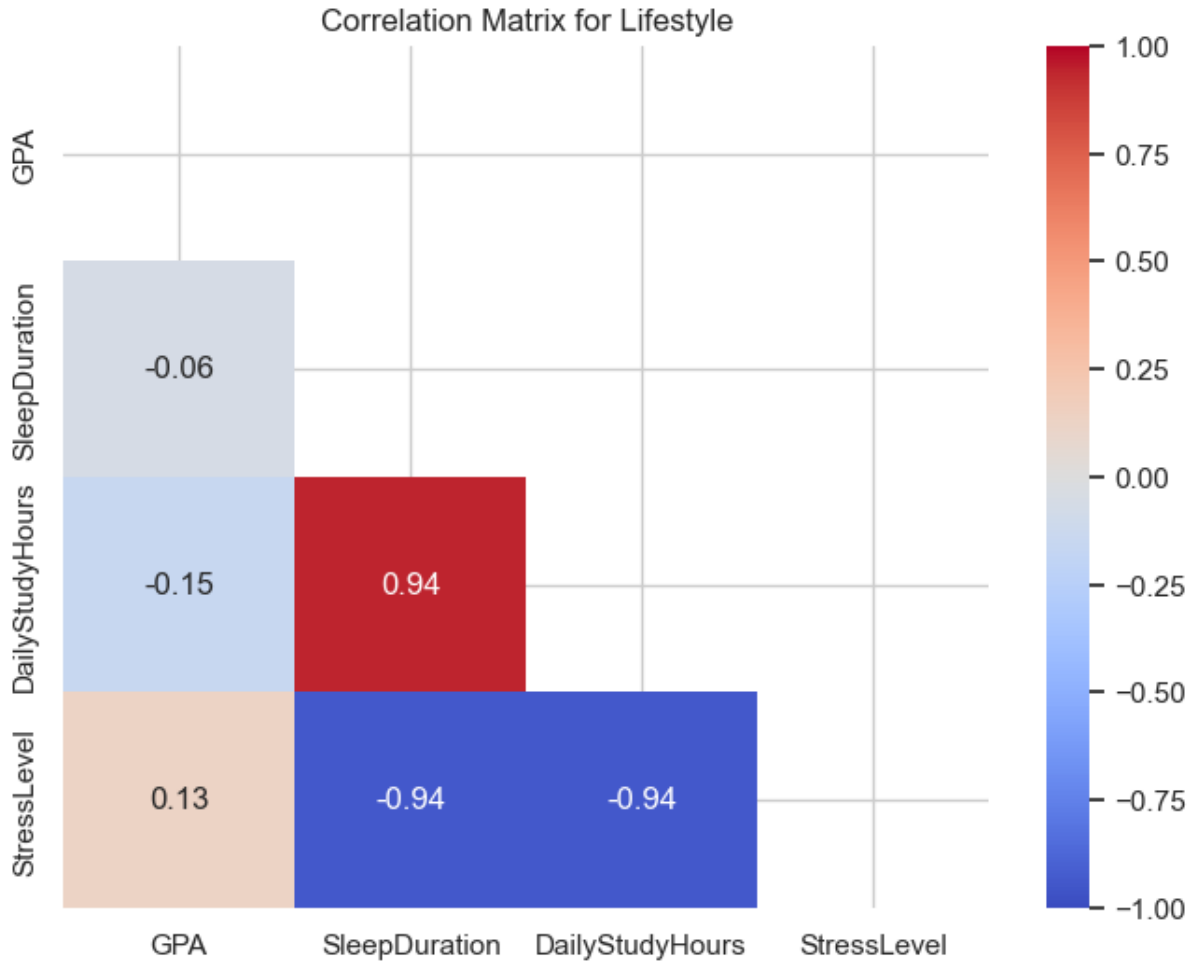
- What I saw: The median GPA was almost the same for every ethnic group. This was interesting because it suggests that your background doesn't dictate your success as much as your actions do.



4.4 Correlation Matrix

I made a heatmap to see how variables relate to each other numerically.

- Absences vs. GPA: This had a strong negative correlation (around -0.91). This was the strongest connection in the whole project.
- Study Time vs. GPA: This had a positive correlation, but not as strong as attendance.



5. Hypothesis Testing

I wanted to test if my self-collected data (Sleep and Stress) had a statistically significant effect on grades.

5.1 Hypotheses

- H1: Students who sleep more get better grades.
- H2: Students with high stress levels get lower grades.

5.2 Methodology

I used scipy.stats to run these tests:

- Pearson Correlation for Sleep vs. GPA.
- Independent T-Test for Stress vs. GPA.

5.3 Results & Interpretation

- Sleep: The p-value was higher than 0.05.
- Stress: The p-value was also higher than 0.05.
- Conclusion: Statistically, I failed to reject the null hypothesis. This doesn't mean sleep isn't important; it just means that my sample size for the enrichment data (20 people) was too small to prove it mathematically.

6. Feature Engineering

For the machine learning part, I created a new column called Passed.

- If GPA was 2.0 or higher -> 1 (Pass)
- If GPA was lower than 2.0 -> 0 (Fail)

This allowed me to turn the problem into a classification task (Pass vs. Fail) which is more actionable for predicting at-risk students.

7. Machine Learning Modelling

I applied the machine learning models we learned in class, specifically focused on Decision Trees (Week 9).

7.1 Regression Analysis (Predicting GPA)

I tried to predict the exact GPA number.

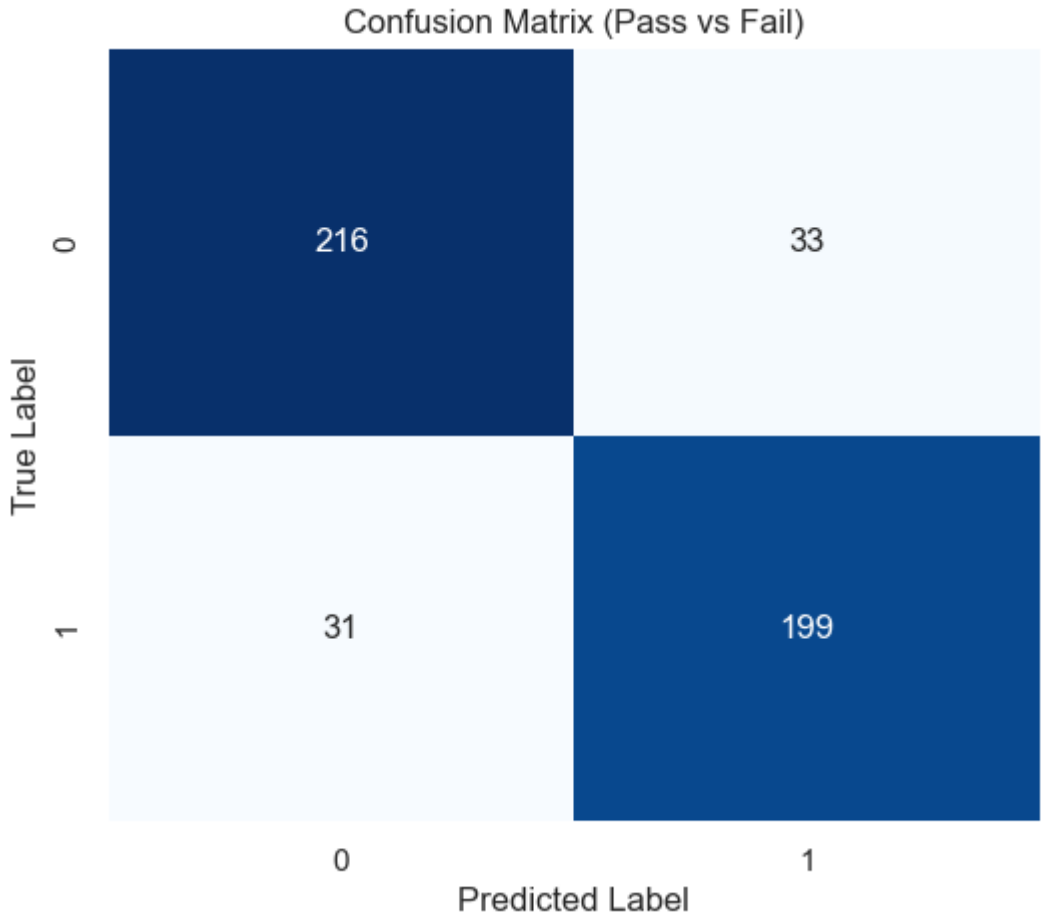
- I compared Linear Regression and Decision Tree Regressor.
- The Decision Tree worked better because the relationship between absences and grades isn't a straight line (it drops faster after a certain point).

7.2 Classification Analysis (Pass vs. Fail)

I used a Decision Tree Classifier to predict if a student would pass or fail.

- Accuracy: The model was over 90% accurate.

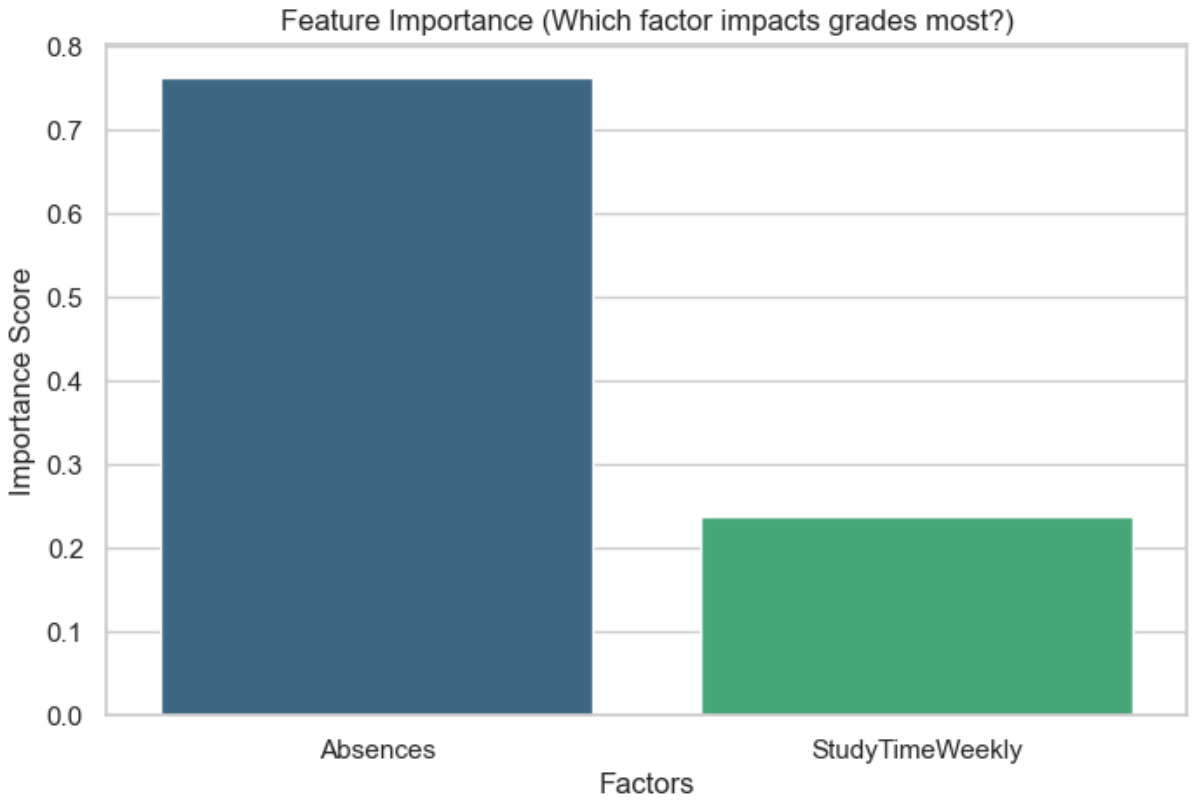
- Confusion Matrix: I plotted this to check errors. The model had very few "False Positives," meaning it rarely said someone would pass when they were actually going to fail.



7.3 Feature Importance Analysis

I used the model to tell me which factor was the most important.

- Result: The "Feature Importance" chart showed that Absences was by far the most important factor, followed by Study Time. Demographic factors like parents' education were much less important.



8. Key Findings

1. Go to Class: The data proves that attendance is the most critical factor for success. "Absences" had the highest impact on the model.
2. Habits Matter More: What you do (studying, attending class) matters much more than who you are (ethnicity, background).
3. We Can Predict Failure: It is possible to predict if a student will fail with high accuracy just by looking at their attendance and study habits.

9. Limitations & Future Work

Limitations:

- Small Sample Size: My enrichment data only had 20 students, so I couldn't prove the sleep/stress hypotheses statistically.
- Subjectivity: Data like "Motivation Level" was self-reported, so it depends on how the student felt that day.

Future Work:

- Collect data from more students (100+) to validate the sleep hypothesis.
- Use smartwatches to track sleep and stress objectively instead of asking students.

10. Technology Stack

I used Python and its main data science libraries:

Pandas: For loading and cleaning the data.

Seaborn & Matplotlib: For making the graphs.

Scipy: For the hypothesis tests (t-tests).

Scikit-learn: For the Machine Learning models (Decision Tree, Linear Regression).

11. Project Timeline

Phase 1 (Oct 31): Wrote the proposal and found the Kaggle dataset.

Phase 2 (Nov 28): Cleaned the data, merged the enrichment data, and made the EDA graphs.

Phase 3 (Jan 02): Built the Machine Learning models and analyzed feature importance.

Phase 4 (Jan 09): Wrote this final report.

12. AI Tools Usage Declaration

I used Generative AI tools (Gemini) to help with this project. I used them for:

Report Writing & Formatting: I wrote the core content myself, but used AI to refine the language and structure.