Exploring Student Habits and Academic Performance

Introduction

For my final project, I chose to work with the "Student Habits and Performance" dataset. It includes information on 1,000 students, things like their daily routines, personal backgrounds, and exam scores. I picked this dataset because I've always been curious about how everyday habits, like how much someone studies, sleep, or uses social media, can affect their academic performance. I thought this could be useful for students and teachers who want to better understand what influences success in school.

Data Preprocessing and Cleaning

The dataset was already pretty clean, which made things easier. There were no missing values, and it included both numerical and categorical variables like gender, whether the student had a part-time job, parental education level, and diet quality.

Here's what I did to get it ready for analysis:

- Changed all the categorical columns (like gender and job status) into a format that's easier for models to understand.

- Scaled the numerical data so that features like study hours and sleep were in a similar range, which helps with modeling.

These steps helped make sure the data was ready for analysis and machine learning.

Data Analysis

While exploring the data, a few patterns stood out:

- **Exam scores** were mostly high (between 60 and 100), but there were still some students who scored much lower.

- Students who **studied more slept more**, and had better **attendance** generally did better on exams.

- On the flip side, students who spent more time on **social media** or **watching Netflix** tended to score lower.

These findings make sense—good habits like studying and sleep help performance, while too much screen time can be a distraction. I used charts like histograms and heatmaps to visualize these patterns and figure out which features mattered most for prediction.

Machine Learning Modeling

To test how well we could predict exam scores based on these features, I built two models:

1. Linear Regression

- **RMSE**: 5.15
- **R^2**: 0.897

This model did a really good job. It explained about 90% of the variation in exam scores and showed that the relationships between study habits and scores are mostly linear.

2. Random Forest Regressor

- **RMSE**: 6.21
- **R^2**: 0.850

The Random Forest model also worked well but wasn't quite as accurate. It may have been overfitting a bit or just wasn't as suited for this dataset since the patterns weren't too complex.

Overall, linear regression ended up being the better fit.

Conclusion

This project helped me apply everything we learned in the course: from cleaning and preparing data to exploring it and building models. I learned that even simple models can be powerful when the data is clean and the patterns are clear.

What I learned:

- Cleaning and prepping data is just as important as the modeling part.
- Looking at correlations early on helps you know what to focus on.
- Simple models like linear regression are easier to understand and often work really well.

For the future, it might be interesting to add features like past grades or motivation level or to look at how student performance changes over time. I'd also like to explore which features are most important using more advanced model techniques.