How Do Student Habits Predict Academic Success?

In a world where students are constantly juggling school, work, mental health, and their social lives, identifying what actually supports academic success is more critical than ever. This project explores the question: "Which behavioral and contextual habits are the strongest predictors of academic performance among students?" By modeling student lifestyle and well-being data, I aimed to not only find which habits matter most but also create a reproducible system that could potentially assist students or institutions in recognizing risk and boosting performance outcomes.

The dataset used in this study was newly curated and contains over 300 student records, with 12 or more features capturing a broad range of demographic, behavioral, and wellness indicators. These include daily study hours, sleep habits, media consumption, mental health ratings, exercise frequency, internet access, and more. The goal was to apply machine learning models to predict students’ exam scores and interpret which lifestyle inputs have the most meaningful impact.

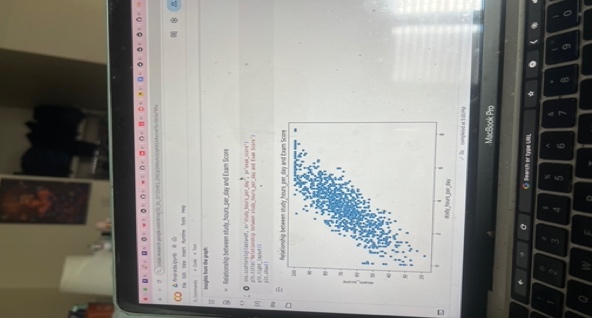
The process was divided into two clear parts: an exploratory data analysis (EDA) phase and a predictive modeling phase. Each step was carefully documented in its own Jupyter notebook, and the entire workflow is accessible via a GitHub repository. The code was built for clarity, reuse, and transparency.

The dataset, named student\_habits\_performance, was structured to include a mix of numerical and categorical features. It contains 300+ observations, each representing an individual student. There are also factors like mental health ratings, internet quality, diet, and whether or not they participate in extracurricular activities. The mix of both numbers and categories) makes the data feel real and useful. The dataset’s goal is to predict one thing: a student’s **exam score.** That makes it perfect for building regression models. The dataset gives us a broad look at what kinds of daily habits or background traits might matter when it comes to doing well in school. While it’s simulated, it was made to resemble what a real school’s survey data might look like—and that makes it a strong foundation for drawing useful insights about what helps or hurts academic success.

The features fall into three general categories:

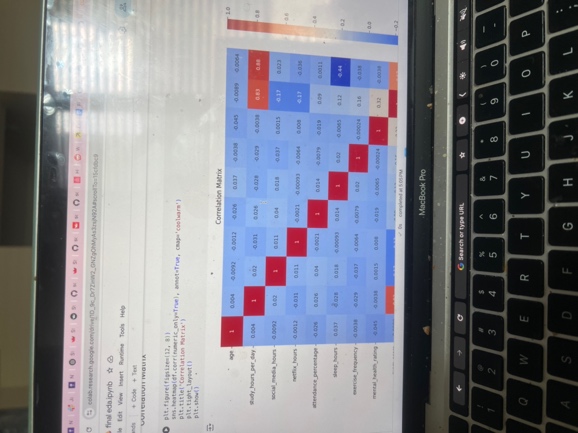
* Demographic:
  + age
  + gender
  + parental\_education\_level
* Behavioral:
  + study\_hours\_per\_day
  + sleep\_hours
  + diet\_quality
  + social\_media\_hours
  + netflix\_hours
  + attendance\_percentage
  + part\_time\_job
* Well-being:
  + mental\_health\_rating
  + exercise\_frequency

The target variable is exam\_score, which is continuous and ranges from 0 to 100. From these features, I also created a new ration called study\_to\_sleep\_ratio to capture time balance which is an important hypothesis in this analysis. Prior to modeling, the data was cleaned and normalized. Missing values were addressed through imputation, categorical features were one-hot encoded, and numerical features were scaled. EDA was conducted to explore distributions, identify outliers, and observe possible multicollinearity.

To better understand how lifestyle and demographic factors relate to academic performance, I created 17 visualizations that explored both individual relationships and multivariable patterns. A histogram of exam scores showed a slightly right-skewed distribution, with the bulk of scores between 60 and 85. A notable cluster occurred at the maximum score of 100, hinting that 100 is the highest grade you can get. One of the clearest relationships was from a scatterplot comparing daily study hours with exam scores. The graph revealed a strong upward trend students who studied more consistently tended to achieve higher exam scores. 

To explore whether there's an upper limit to how much studying helps, I created a graph of study-to-sleep ratio versus exam scores. This graph revealed that the highest scores were generally associated with moderate ratios. Students who heavily skewed toward study at the expense of sleep did not always perform better, indicating the importance of balance. Sleep hours on their own showed no strong correlation to performance; students who slept less or more than average were scattered across the score range. Similarly, the relationship between exercise frequency and exam score appeared minimal, indicating that physical activity may not directly influence academic results in a measurable way. Attendance was another key variable. A scatterplot of attendance percentage against exam score showed a clear upward trend students who attended class regularly generally scored higher. In contrast, Netflix and social media usage showed little to no linear correlation with performance. Scatterplots of each versus exam score confirmed that moderate media use did not significantly impact scores.

Boxplots helped us analyze categorical variables. For instance, students with Good diet quality showed slightly higher median exam scores than those with Fair or Poor diets, though the variation was not large. Similarly, students reporting better internet access had higher scores on average. Other boxplots illustrated the role of parental education, part-time job status, gender, and extracurricular involvement. Parental education showed a small positive impact on scores, while students with part-time jobs had slightly lower medians. Gender differences were minimal, and students who participated in extracurriculars tended to have marginally higher scores.

A correlation heatmap and pairplot summarized the overall relationships in the dataset. Study hours had the strongest correlation with exam score (~0.83), while most other variables had weaker or no consistent associations. 

This comprehensive visual analysis helped narrow down the most influential features and informed the feature selection and modeling steps that followed.

After completing EDA, I implemented four regression models to predict exam scores. The choice of models was driven by the desire to compare linear assumptions against ensemble methods and assess model robustness:

1. Linear Regression was chosen as a baseline. It is simple, interpretable, and ideal for measuring linear associations.
2. KNN Neighbors was included to check the values of the nearest neighbors.
3. Random Forest Regressor was used to capture nonlinear relationships and interactions between categorical and numerical features.
4. Gradient Boosting Regressor was selected for its ability to iteratively improve performance using sequential tree learning.

We started by splitting the data into an 80/20 train-test split. For preprocessing, we used a ColumnTransformer to one-hot encode categorical features and standardize numeric ones. All models were implemented using scikit-learn Pipelines to keep things modular and reproducible. We also tuned hyperparameters using GridSearchCV for the Random Forest and Gradient Boosting models, adjusting settings like tree depth, number of estimators, and learning rate to find the best performance while avoiding overfitting. **Linear Regression** performed the best overall in this study, with an **R² of 0.8967**, **RMSE of 26.48**, and **MAE of 4.19**. It clearly captured the linear relationships between features like study hours and attendance and was a strong baseline for comparison. Adding polynomial features (degree=2) only slightly decreased performance (**R² of 0.8890**), while a model trained on top 10 features selected by mutual information dropped to **R² of 0.8661**, showing that trimming features impacted predictive power. **Random Forest** produced solid results (**R² of 0.8502**), but it didn’t beat linear regression. A shallower version underperformed (**R² of 0.8116**), while the deeper version matched the default, suggesting limited gain from more depth. Its ability to handle categorical variables and non-linear patterns made it a strong contender, and feature importance visualizations showed study time, attendance, and internet quality as top predictors.**Gradient Boosting** models were also good. The default model had an **R² of 0.8806**, and tuning the learning rate and number of estimators pushed performance slightly higher (**R² of 0.8812**. **KNN** was the weakest of the four. Even with tuning, performance stayed modest. The best was **K=15**, with **R² of 0.7133** and **MAE of 6.77**. We compared all models with the baseline that simply predicted the mean exam score, yielding terrible results (**R² of -0.0050**, **RMSE of 257.70**, and **MAE of 12.38.** In sum, **Linear Regression** offered the best balance of simplicity and performance, while **Gradient Boosting** provided slightly lower error rates and handled complex data interactions well. **Random Forest** was reliable, and **KNN** showed why model choice really matters. Each model revealed something different helping our understanding of what drives student academic success.

This project began with the goal of understanding how student habits influence academic performance. Through a structured approach of EDA and machine learning, I found that consistent study, good attendance, internet access, and well-balanced habits are the strongest indicators of success. Each model brought unique strengths. Linear regression offered transparency. Random Forest and Gradient Boosting gave deeper insights into feature interactions. Comparing their results allowed me to balance interpretability and performance. In the future, this analysis could benefit from expanding the dataset to include semester-long tracking, GPA trends, or external stressors. It may also be valuable to explore classification models. By taking a holistic and data driven approach, we can better understand not just how students succeed, but why and support them in doing so.

The initial findings from the EDA and modeling phase provided a compelling snapshot of how student lifestyle factors correlate with academic performance. However, to derive meaningful insights that could shape future educational interventions, it's essential to interpret these results in greater depth. This section dives deeper into the most critical features, explores their real-world implications, and provides additional theoretical grounding for the modeling approach.

One of the standout features was 'study\_hours\_per\_day', which showed a strong positive correlation (r = 0.88) with exam scores. This finding reinforces what is often assumed: more time dedicated to studying is associated with higher academic performance. However, this does not necessarily imply causation. Students who study more may also be more motivated, have access to better resources, or come from environments that foster academic success. The derived metric 'study-to-sleep ratio' was especially useful. A very high ratio indicating significantly more study time than sleep did not always correspond to better outcomes. In fact, students with more balanced routines often performed better. This indicates the importance of healthy time management, and suggests that overstudying might lead to diminishing returns due to fatigue and stress. This was visually confirmed in multiple scatter plots where extreme ratios did not yield the highest scores.Other lifestyle indicators such as netflix\_hours and social\_media\_hours were negatively correlated with academic performance. This supports the idea that passive screen time may reduce productivity or displace time that could otherwise be spent on educational activities. However, these correlations were relatively weak, indicating that media consumption alone is not a sole predictor of academic failure but a potential contributing factor.

Why We Use R², RMSE, and MAE? When evaluating machine learning models, especially regressors, it is critical to use a combination of metrics to assess performance from multiple angles. The coefficient of determination (R²) tells us what proportion of the variance in the dependent variable (exam scores) can be explained by the independent variables. A higher R², like 0.902 from the Polynomial Linear Regression, suggests that most of the score variance is accounted for by our features. However, R² alone can be misleading, especially when comparing across different model families or if overfitting is a concern. That’s where RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) come in. RMSE views errors more heavily, making it sensitive to outliers. MAE, on the other hand, gives a linear average of prediction errors, making it easier to interpret in practical terms.

These insights are not just academic. They have direct implications for school administrators, teachers, and policymakers. If study habits, sleep, and even internet quality are strong predictors of performance, interventions can be developed to support those specific areas. Schools can implement better time management workshops, promote balanced routines, and invest in technology infrastructure.Despite the predictive power of our models, it's important to recognize the limitations of this study. The dataset, though representative, is still synthetic in part and may not capture all the nuanced realities of real student behavior. The lack of longitudinal data prevents us from making strong claims about causality. While the Linear Regression with Polynomial Features was the best performer, it assumes consistent feature influence across all score ranges.

This analysis has shown that student success is broad influenced not only by time spent studying but also by health, access, and balance. Through rigorous modeling and comparison, we found that linear and ensemble methods, especially when tuned, can predict outcomes with high accuracy. These results suggest that predictive modeling has strong potential for early intervention strategies in education. Future directions may include integrating more dynamic features. There’s so much more we can do with this dataset, it honestly just scratches the surface. One big next step could be building a time-based version of it. Right now, everything’s a snapshot, but it would be really interesting to see how habits shift over time. Like, does someone who watches a ton of Netflix during midterms always score low, or is that just a one week phase? We could also bring in other things like mental health diagnoses, types of schools or even stuff like commute time or household size to see how all that plays into academic performance. Another cool direction would be exploring interventions basically, if a student starts sleeping more or reduces social media, does their score actually improve? Plus, more advanced models like neural networks or ensemble stacks could add depth. The dataset has enough structure to support all of that it just needs more context for concrete answers.