Data Bootcamp Final

Introduction:

The goal of the final project is to look at what variables affect a student's final exam performance, and build a model to predict final performance based on the aforementioned variables. Our regression modelling results suggested that the top three most important variables in academic performance were first: hours spent studying, second: mental health rating, and third: either hours of social media use or hours of sleep, depending on the regression model. Meanwhile, our EDA data suggested that hours spent studying, exercise frequency, mental health rating, and participation in extracurricular activities were associated with positively influencing on final exam performance; hours spent on social media, hours spent on Netflix, and low attendance were associated with a downward trend in exam scores; and age, gender, part time job status, hours of sleep, diet quality, parental education level, and internet quality had no significant impact on exam scores. We expect this data will be useful to students, parents, educators, and anyone else that is interested in helping students perform better, especially as academic performance is related to future success in both professional and personal areas. Understanding what variables most strongly predict academic success can help schools design better support programs and set students up for better futures.

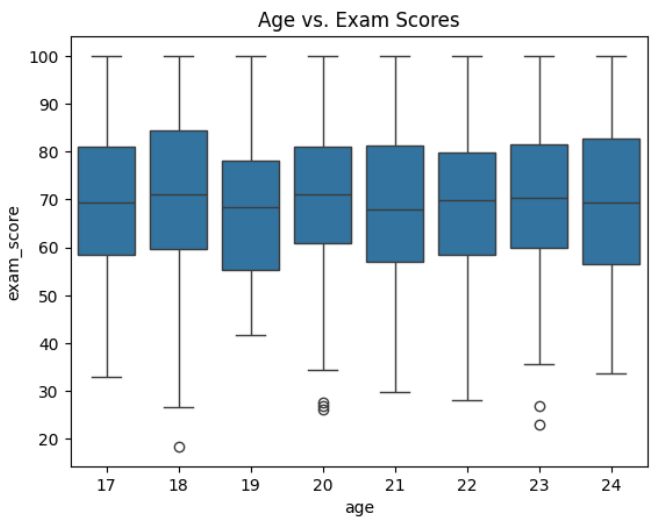
Data Description:

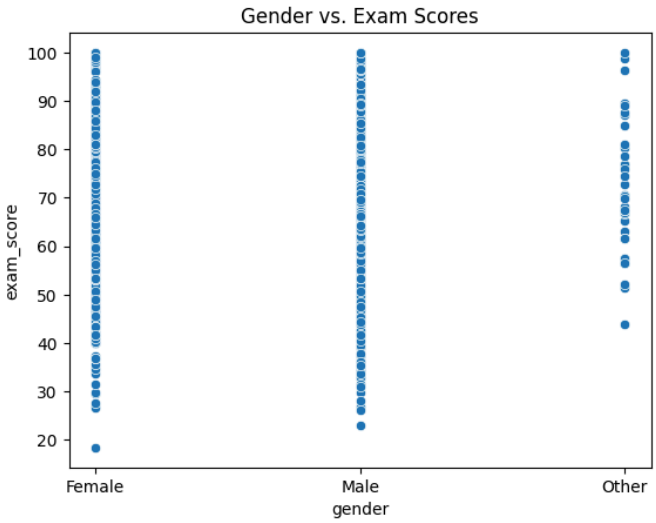
In a synthetic dataset from Kaggle consisting of 1000 students modeling academic performance, we looked at final exam performance through their grades, in relation to many different variables: age, gender, hours spent studying per day, hours spent on social media, hours spent on netflix, whether or not they held a part time job, attendance percentages, hours of sleep, quality of diet, frequency of exercise, their parent's education level, internet quality, mental health rating, and participation in extracurriculars.

Models and Methods/Results and Interpretation:

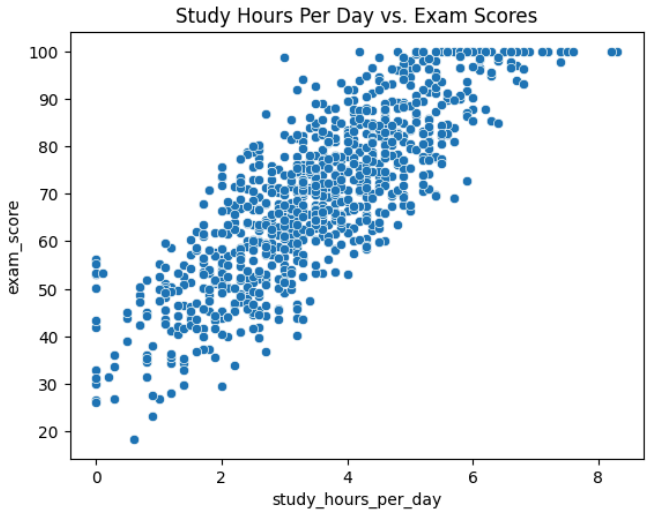
Exploratory Data Analysis:

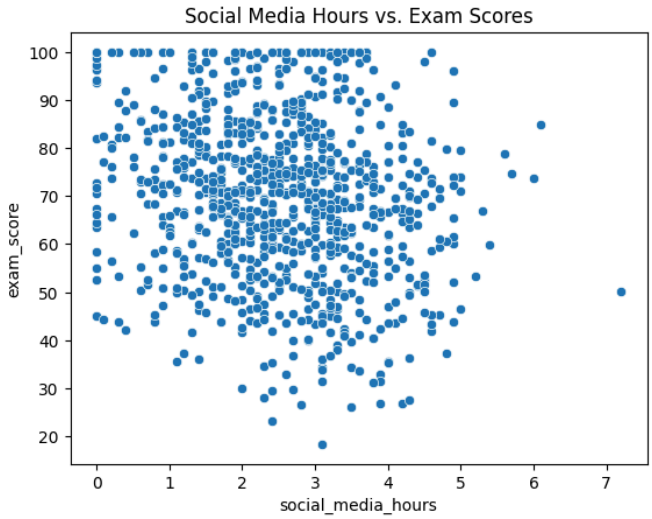
Firstly, we performed exploratory data analysis (EDA), to look at the relationship between different variables and exam scores. In our EDA, it was shown that our dataset consisted of male, female, and other students between the ages of 17 to 24, with a mean of 20. Looking at these variables –age and gender– we looked at if there was an association with final exam performance. The average for each age number was roughly the same, and was close to the overall exam average, showing no association between age and scores. Similarly, for gender the range of scores had similar spread for all genders and the range for males and female, however, the range for the other gender was generally higher than that for male and female, but there were fewer data points so it was insufficient to make an accurate conclusion that there is a difference between males/females and other.

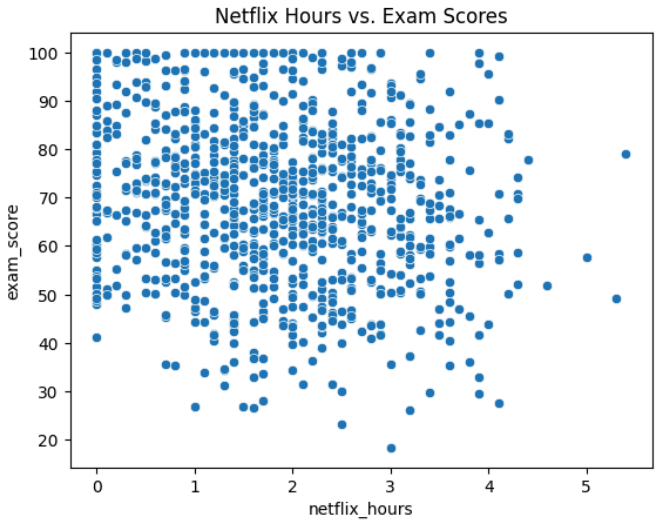


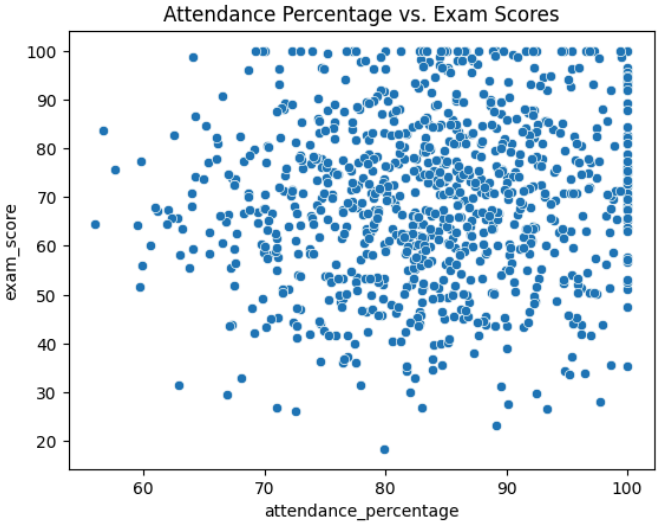


We also looked at behavioral variables: hours spent studying, hours spent on social media, hours spent on Netflix, and attendance percentages to see if they contributed to exam performance. Our data shows that hours spent studying played a significant role in exam performance, where the more hours spent on studying resulted in a better score. There was a strong, positive, linear relationship between the studying and scores. This makes sense as students who study more are generally more familiar with the content and understand it better. However, both hours spent on social media and hours spent on Netflix did not have a strong correlation to exam performance, but had a slight downward trend that suggested higher number of hours spent is associated with lower exam score performance, although the scatterplots were diffuse for both. This is probably attributed to the fact that time spent on social media or Netflix is time away from studying. On the other hand, attendance percentage was also weakly correlated to exam performance with a substantial amount of students with 60 to 80% attendance averages still performing well (90 and above), but there were also a cluster students with low attendance percentage (75% and below) that received low scores (around 70 and below), showing that low attendance may be correlated to low scores. One suggestion is that some students may be able to keep up through lecture recordings or textbooks, but these methods may not be as effective as attending class in person for everyone. Thus, more time spent studying is a significant predictor of scores, however social media, Netflix consumption, and attendance percentage may impact exam performance although they are just minor contributing variables and not the sole determining factor.



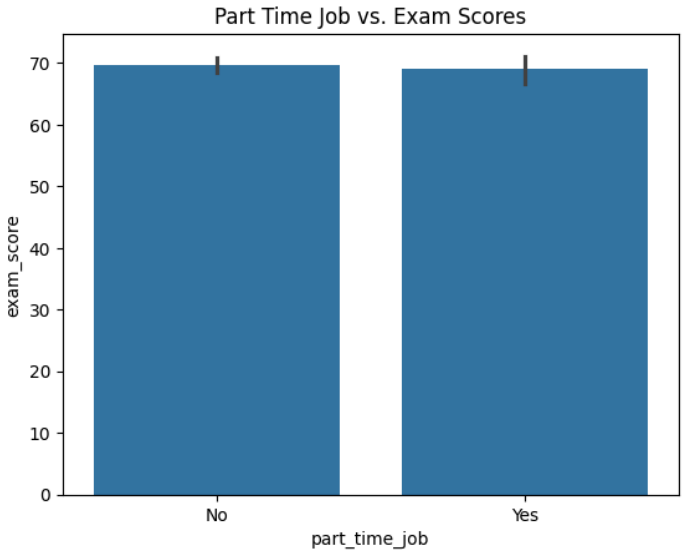


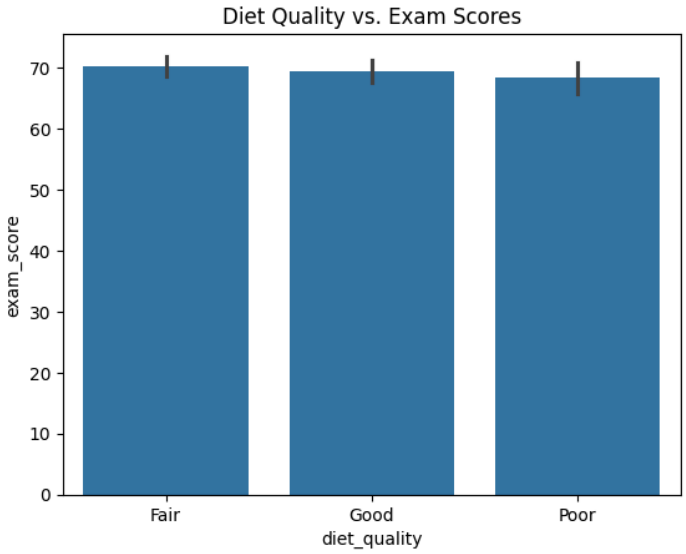


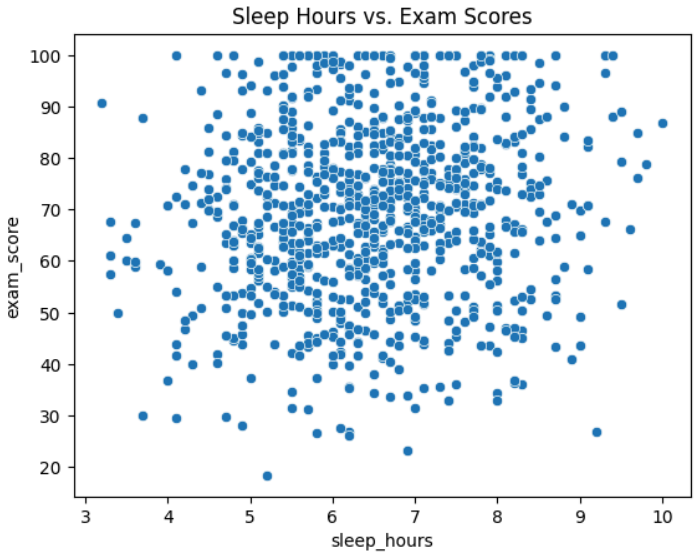


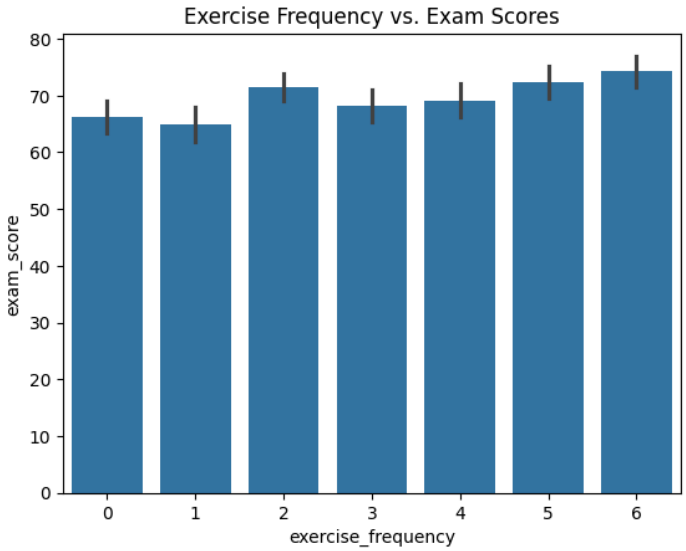
Additionally, we looked at lifestyle variables like whether a student has a part time job, hours of sleep, diet quality, and exercise frequency to see if there was a correlation between these variables with exam scores. Students who responded “no” performed slightly better than those who responded “yes”, but the difference was only by a few points and the results may not even be statistically significant. This could be attributed to students without a job having more time to study and not needing to balance work and school. Similarly, the hours of sleep a student received played a minimal role in explaining their score performance, with the majority of students reported sleeping between 5 to 8 hours performing between 40 to 100 on the exam. However, there were fewer students who slept between less than 5 hours and more than 9 hours and their scores were more scattered high and low, suggesting that it may not be ideal either. Thus, our data suggests that it is best for students to sleep between 5 to 8 hours, although more or less would not impact performance significantly as it is not strongly correlated with performance. This makes sense that as long as a student is receiving a healthy amount of

sleep, their performance would be attributed to other variables instead. We also evaluated diet quality, with students rating between poor, fair, and good. Our data suggests that students who rated fair had the highest exam score average at slightly above 70, while good and poor had averages slightly below 70, although the differences may not even be statistically significant since all of the error bars overlapped with each other, which implies the diet quality is not a strong predictor of exam performance. This makes sense because as long as a student is eating appropriately, there are other variables that are more important. On the other hand, we looked at exercise frequency that showed a clear upwards trend that those who exercised more frequency had on average a higher exam score. This suggests a possible positive correlation between exercise frequency and exam score, which may be attributed to the positive benefits of exercise contributing to better student wellbeing.

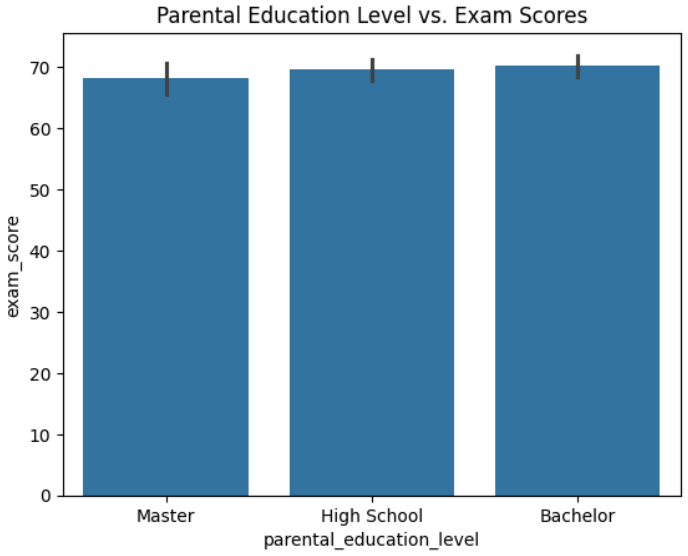


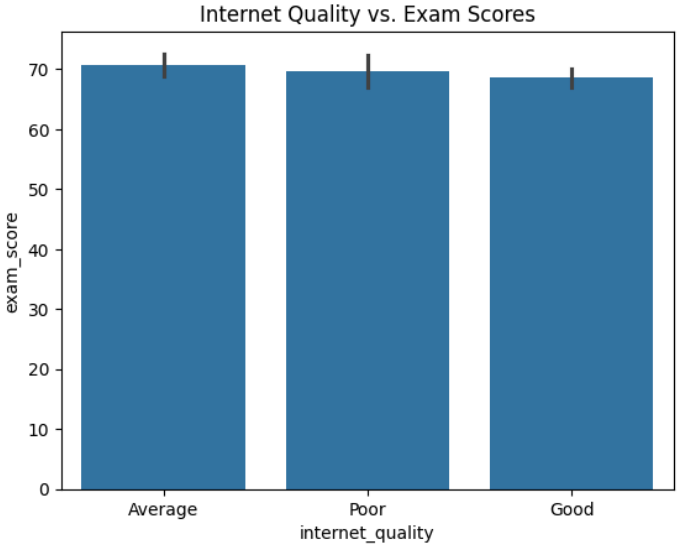


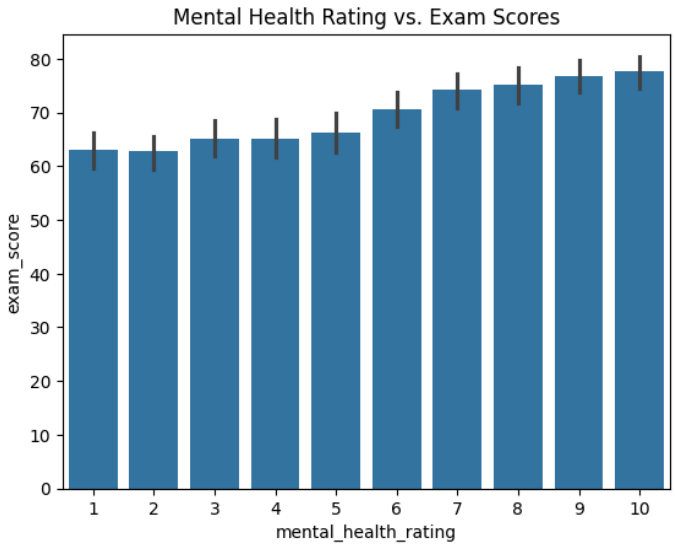


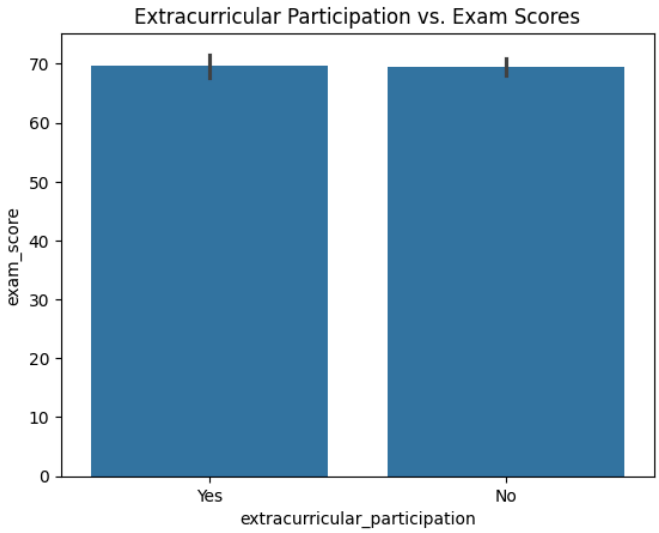


We are also looking at environmental variables that could affect a student's performance, which are parental education levels, internet quality, mental health rating, and extracurricular activity participation. We looked at parental education level, whether someone's parents who had a master's degree, bachelor's degree, or high school degree played a role in a student's exam scores. The results showed that students whose parents had bachelor's degrees had the highest scores, followed by high school, then master's degrees, suggesting that parental education may influence the academic environment but not to a significant degree as it is not the case where students who performed best had parents with the highest degrees. However, the difference between them was small and the error bars overlap each other, suggesting that parental education level does not play a significant role in scores. Additionally, we looked at internet quality, where students reported the quality as good, average, or poor. Students with average internet quality performed best, followed by poor, then good. The differences were small and had overlapping error bars, which suggests that it does not significantly affect performance, suggesting that as long as a student has access to the internet, there was no difference in scores. However, mental health rating seemed to have a positive correlation to scores, such that the better a student's mental health rating is, the better their scores are. This can be explained by the fact that students who are rating their mental health higher are probably not dealing with stress, anxiety, etc. that could affect behavior in school. Lastly, we looked at extracurricular activity participation. Our results suggest that students who participate in extracurricular activities do slightly better than students who do not participate. This could possibly be explained by the fact that students involved in extracurriculars may be developing better time management or benefiting from social connection, both of which can support academic success. However, similar to the prior variables, there is a lot of overlap between students who participate and those who do not, which seems to suggest that the significance is minimal and other variables play a bigger role.









Our exploratory data analysis therefore suggested that hours spent studying, exercise frequency, mental health rating, and participation in extracurricular activities were associated with positively influencing on final exam performance; hours spent on social media, hours spent on Netflix, and low attendance were associated with a downward trend in exam scores; and age, gender, part time job status, hours of sleep, diet quality, parental education level, and internet quality had no significant impact on exam scores.

These results suggest that the variables students can actually control, specifically the amount of time spent studying, exercise frequency, mental health ratings, and participation in extracurriculars, were actually the most relevant when looking towards ways to improve exam scores. This means that students, schools, parents, and anyone else invested in a student's performance should focus on supporting those areas if they want to help improve performance. On the other hand, spending more time on social media or Netflix and skipping class seems to worsen exam performance, even by just a little, so it could be recommended to avoid those variables if a student is already performing poorly. Meanwhile, variables like age, gender, sleep, diet, parental education, and internet quality did not appear to make much of a difference. This shows that improving academic performance is mostly about behavior.

Regression Modeling:
The goal of our project is to predict the impact various behavioral, lifestyle, and environmental variables have on a student's academic performance through their final exam scores. After performing EDA, we performed modeling with several regression models: multiple linear regression model, KNN regression model, decision tree regression model, and random forest regression model. We also created a baseline model which predicted the mean score for all students to use as a comparison of our regression models. The goal of our regression models is to identify any relationships between predictors and outcomes and form connections between multiple variables and identify the most significant variables that influence student performance.

Out of all four models that we ran, the best to worst was multiple linear regression, random forest, decision tree, and lastly k-nearest neighbor. In general, our EDA results were the same as our regression model results, however some variables showed inconsistencies between EDA and modeling, such as extracurricular participation and diet quality appearing weakly positive in EDA but had minimal or negative effects in regression, which suggests any benefits are likely indirect. Another exam is sleep ranked higher in tree-based models than in linear regression, possibly due to non-linear effects. Attendance showed a weak EDA trend but no model significance after study time and mental health were considered.

In our baseline model, we used it to predict an average exam score for all of the students. The MSE (mean squared error) of this was 286.29, which is high but shows variability in the data due to all of the different variables.

In our multiple linear regression model, we wanted to understand how multiple variables (different lifestyle, behavioral, and environmental variables) interact and contribute to a student's overall final exam score. The model identified hours spent studying per day as the most significant predictor, with each additional hour associated with a 9.61-point increase in exam scores. This was followed by hours of sleep (+2.02), mental health rating (+1.98), exercise frequency (+1.53), and not having a part-time job (+0.41). The strongest negative predictors were hours spent on social media (-2.65), hours spent watching Netflix (-2.15), reporting a good diet over other options (-0.67), extracurricular participation (-0.39), and internet quality (-0.35). The variables identified to have no effect on scores are gender, age, attendance, and parental education. Similarly, our feature importance scores suggest likewise with hours studied per day to be most significant (1.33), followed by mental health rating (0.237), and social media use (0.07). Meanwhile, the training MSE was 28.21 while the MSE for testing data is 30.01, suggesting our model generalizes well and is stable across different subsets of data, and supports our EDA data overall.

In our random forest model, which reduces overfitting by averaging multiple trees, the data is similar to MLR. The most important features were identified as study hours per day (1.34), mental health rating (1.93), and sleep hours (0.04). Interestingly, MLR put social media as the third most important factor, while random forest puts sleep hours. Variables such as age, gender, internet quality, parental education played a minimal/no impact on performance. The training MSE is 6.69 while the testing MSE was 32.43 which suggests the training data was fit extremely well but the testing data had overfitting and poor generalization, and did not captures the data patterns well probably because it memorized the training data pattern and is applying it to the testing data.

Additionally, a K-Nearest Neighbors model was performed to see if students with similar characteristics performed similarly to each other. In our regression model, it showed that the most important features were hours spent studying (0.85), mental health rating (0.14), and hours of social media (0.04), which is similar to the results from MLR. Like multiple linear regression, the features that played a very minimal role are variables like age, gender, diet quality, internet quality, and extracurricular participation. However, the MSE for training data was

62.16 and MSE for testing data was 78.74, which are quite high and shows a moderate amount of error and a drop in performance between training and testing, probably due to overfitting or sensitivity to noise in the data, and it did not generalize data patterns as effectively as the other models.

Lastly, we performed a decision tree model in order to look at any non-linear relationships in the variables with performance. Similar to other models, the decision tree model suggested that the most important features were hours spent studying (1.40), mental health rating (0.23), and hours of sleep (0.04). The insignificant variables were gender, parental education, diet quality, part time job status, internet quality, extracurricular activities, which all had 0 or near 0 values. The training MSE is 46.45 while the testing MSE is 61.84, which suggests that the model performed reasonably well with low overfitting since the MSE values are not significantly further from each other, and generalizes moderately well to testing data, however the high numbers itself shows that it does not capture data patterns as well as other models did.

The differences in how each model performed can be explained by how they process relationships between variables. Multiple linear regression did well because it captures linear, additive effects, which matched our data's structure, especially for variables in the dataset like hours studied and mental health rating. Random forest was able to pick up more complex, non-linear relationships, but it also showed signs of overfitting, likely because it memorized patterns in the training data that didn't carry over to the test data. Decision trees performed moderately well but didn't capture the data patterns as effectively as linear regression or random forest, possibly because they focus on sharp split points rather than broader trends. Lastly, K-nearest neighbors had the worst performance, which is likely due to how it struggles with high-dimensional data such that when too many features are involved, it becomes harder for the model to identify meaningful patterns based on similarity between students. These differences also explain why some models ranked sleep hours higher while others emphasized social media use, due to how each model detects patterns, so some models may have picked up subtle interactions while others prioritized broader trends. However, even our best model had a test MSE of 30.01, indicating moderate prediction error—suggesting that while these variables are important, academic performance also involves elements that are harder to quantify, such as motivation or instructor quality.

Nonetheless, our models generally show the same results: hours spent studying is most significant, followed by mental health rating, then either hours of sleep or hours of social media use, depending on the model (our best model suggested social media use to be the third most significant model). This aligns with our EDA findings: students who studied more tended to perform better, likely due to stronger content familiarity. Additionally, with a higher mental health rating, the student probably feels less anxious or stressed about academics and the exam, allowing them to take it with a better mindset and perform better without any distractions. Lastly, a lower number of hours on social media would allow for more hours studying and less distractions while studying allowing them to focus completely and get more out of studying. Meanwhile, the more hours spent sleeping within a healthy range allows the brain to rest and is more beneficial to health, allowing students to perform optimally when well rested. Given how

consistent these findings were across all our models, and how well they aligned with our EDA, these results can inform practical recommendations for students, parents, and educators.

Insight and Recommendations:

It would be recommend that students, educators, parents, and anyone else to find ways to support the students in the areas they can change, such as giving the opportunity to allow student to dedicate time to study in a good environment where they can focus their attention on studying, supporting mental health initiatives like offering therapy and counseling to students who could benefit from it, offering programs for students to participate in programs like yoga, meditation, amongst other options, to reduce stress and anxiety while encouraging extracurricular participation, offering more extracurricular activities (a positive influence on grades) that provide social interactions for less social media usage, and getting enough sleep. Students can be supported in working on variables that can improve their performance like studying effectively and improving their mental health, parents can help by providing a calm environment for students to study at and minimizing distractions by placing screen time limits, TV limits, etc., and educators can help by offering structured assignments that keep students on track, offer mental health check ins, host extracurricular events, and offer good study spaces. By implementing these strategies, our model suggests that students will perform better academically, which can set them up for greater success in the future.

Conclusions and Next Steps:

While our models identified several significant predictors of academic performance, there are limitations to consider. Our dataset was synthetic, which may not reflect real-world complexity. Real life variables like motivation, learning disabilities, instructor effectiveness, or peer influence, which are not involved here may also play a significant role. These variables can greatly influence student performance despite someone having possibly the same stats as a student in our dataset, and could have interacted with some of the variables that went into our modelling. Additionally, other environmental variables could have impacted our modeling like course load, academic standing, stress and anxiety, which also could explain any scores that deviated from the general trend. In future work, it could also be helpful to use real data that takes into account more complex variables to get a more comprehensive idea of the extent to which different variables affect academic performance. Although these limitations exist, our data was still able to show a clear correlation between different variables showing both strongly and weakly correlated variables with academic performance.