Jennifer Wang
Professor Jacob Koehler
Data Bootcamp
12 May 2025

Predicting Students' Exam Scores Through Lifestyle Habits

Introduction & Data Description:

Predictive Task & Objective:

In today's academic environment, exam scores directly impact a student's GPA. While GPA is not the sole determinant of future success, it, yet, remains a critical component considered by colleges, internship programs, and employers. Most students aim to maintain a high GPA as it can offer broader career opportunities, while a lower GPA can limit their career options. Given the importance of exam scores, it is essential to understand the factors that influence academic performance. Many of the lifestyle habits students develop can significantly impact the quality and amount of time they devote to studying, ultimately affecting their exam performance. For my final project, I aim to build a predictive **regression model** capable of accurately estimating **students' exam scores**, exploring how various **lifestyle habits** may impact academic performance.

By developing this model, the project seeks to help students make more informed choices about how they manage their time across various activities. Additionally, by analyzing the relationship between lifestyle habits and academic performance, the project can also offer valuable insights for educators and parents. Educators can use the findings to design educational programs that encourage productive habits, while parents can better support their children by fostering routines linked to academic success.

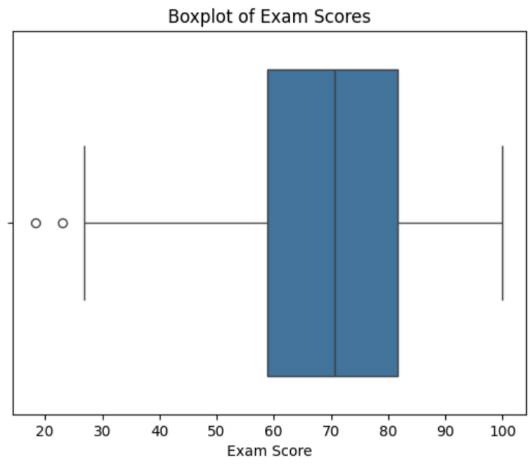
Dataset Overview:

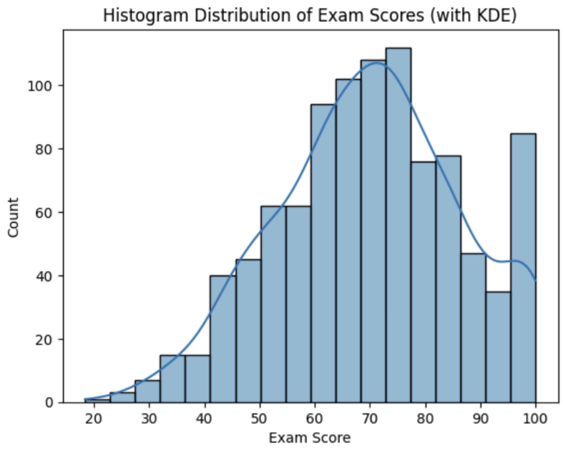
The dataset used for this project is sourced from Kaggle in CSV format. It contains 1,000 rows and 16 columns, including both numerical and categorical features related to a student's behavior and family background. Key features include daily study hours, social media hours, attendance percentage, and extracurricular participation, among others. The target variable is the student's exam score, a continuous numerical value. Overall, this dataset provides comprehensive information for unpacking how various factors impact academic performance.

Exploratory Data Analysis

After adding new features and removing rows containing “infinity” values, the final dataset consists of 987 rows and 19 columns.

Descriptive Statistics for the Target Variable: Exam Score:

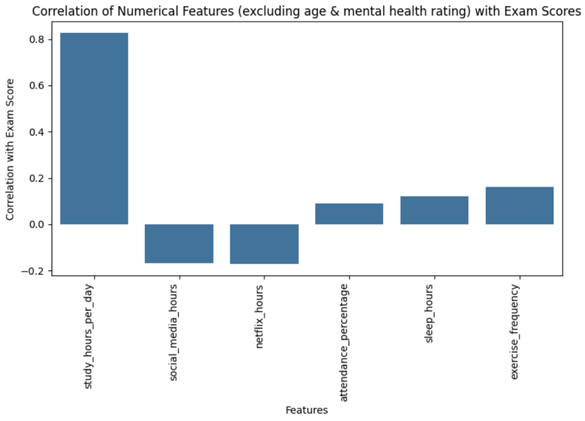




The exam scores range from 18.4 to 100. The mean exam score is 69.98, and the median exam score is 70.6. The boxplot and histogram above show that exam scores are approximately normally distributed, with a slight left skew. Interestingly, while most students cluster around the 70–75 score range, there is also a secondary peak around 95 - 100 score range. Given the mild skewness, the close alignment of mean and median, and the overall normality of data distribution, a log transformation of exam scores was not necessary for this project.

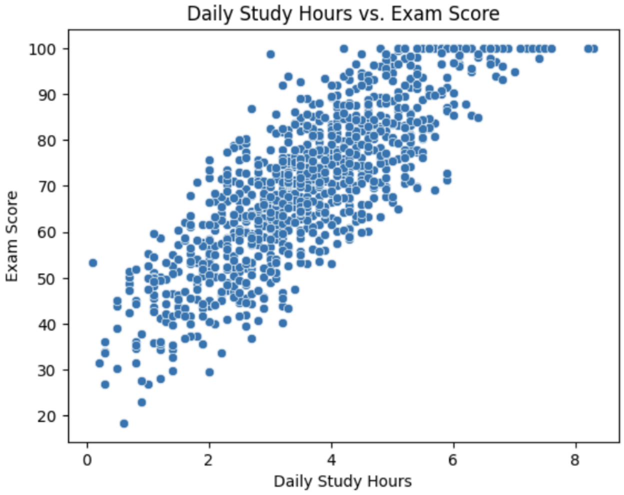
Initial Visualizations Between Independent Variables and the Target Variable:

Numerical Features:

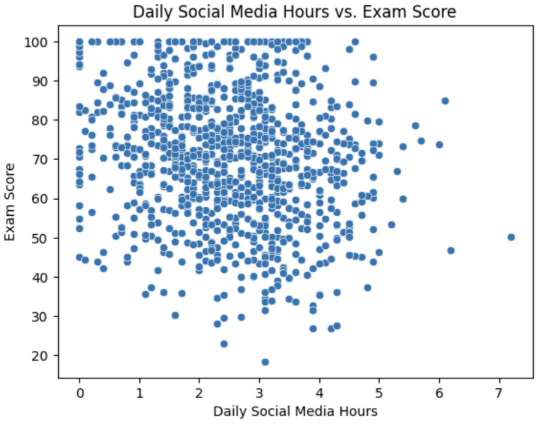


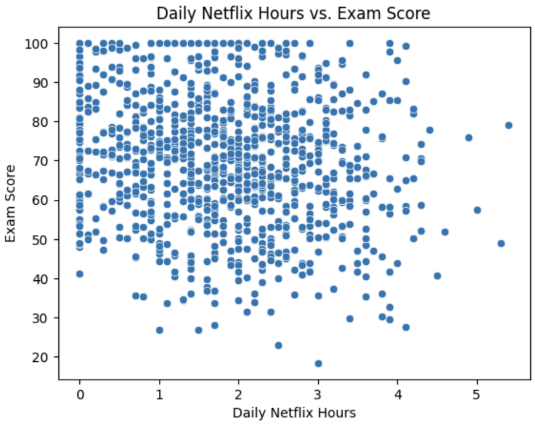
This barplot illustrates the correlation between each numerical feature and exam scores. Among all features, daily study hours has the strongest positive correlation with exam scores (0.825), indicating that more time studying leads to higher exam scores. Weekly exercise frequency (0.16) and daily sleep hours (0.12) also show modest positive correlations. Although their correlations are significantly lower than that of daily study hours, these features may still enhance the model's predictive accuracy. In contrast, daily social media hours (-0.17) and daily Netflix hours (-0.17) are negatively correlated with exam scores, implying, on a superficial level, that increased screen time leads to lower exam scores.

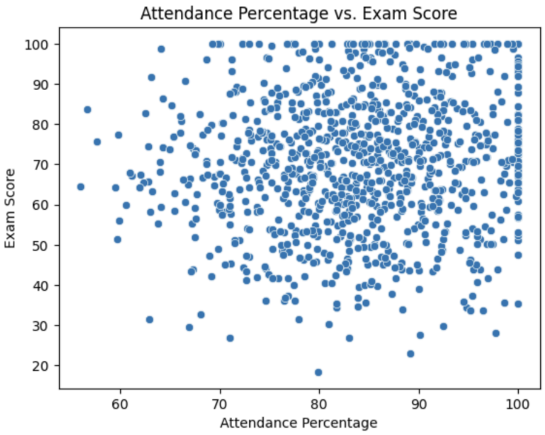
Since correlation only evaluates the degree of linear association between two variables, I created scatterplots between these features and exam scores to explore potential nonlinear relationships.

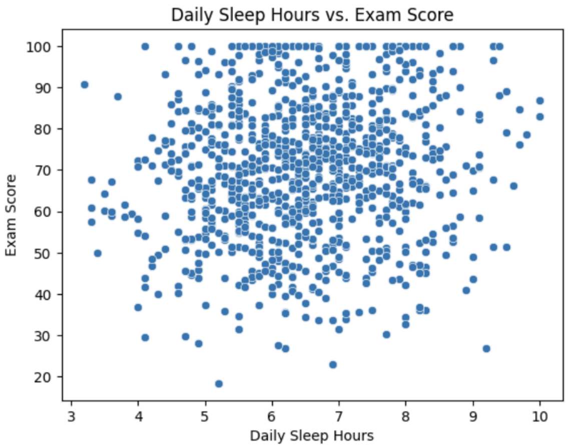

Daily Study Hours vs. Exam Score

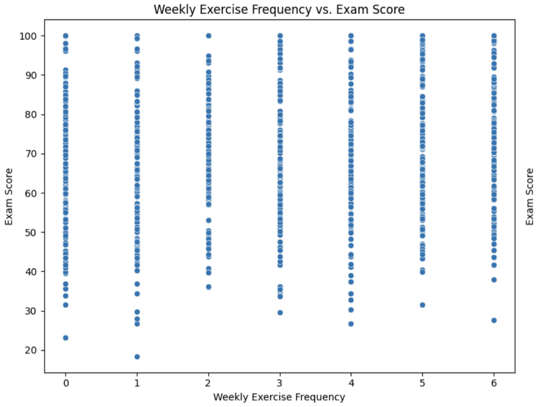
This scatterplot shows a strong positive linear relationship between daily study hours and exam scores – students who studied more tend to achieve a higher exam score. This observation aligns with daily study hours' high correlation value with exam scores.


Daily Social Media Hours vs. Exam Score


Daily Netflix Hours vs. Exam Score

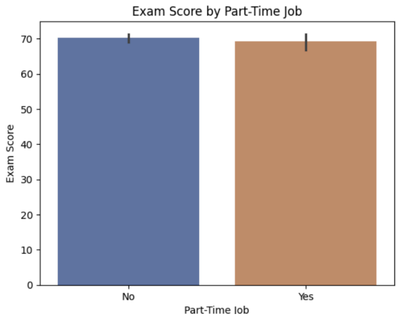


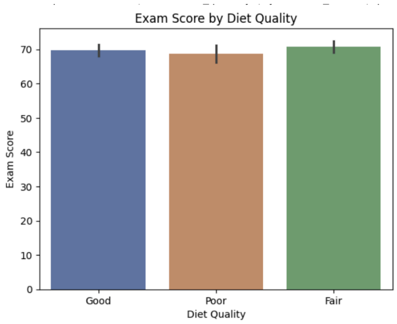


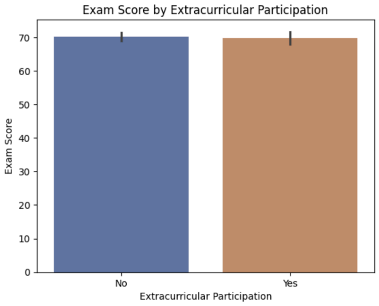


Based on the scatterplots above, the rest of the numerical features showed no clear linear relationship with exam scores. Rather, daily social media hours, daily netflix hours, attendance percentage, and daily sleep hours all exhibited clustered patterns. Despite their comparatively low correlation values, these features will be included, as regression models like K-Nearest Neighbors and Decision Tree can capture nonlinear and localized patterns, possibly enhancing predictive accuracy and uncovering deeper insights into exam performance. Although weekly exercise frequency showed no clear linear or localized pattern with exam scores, the feature will be retained for modeling as it has the second highest positive correlation.

Categorical Features:





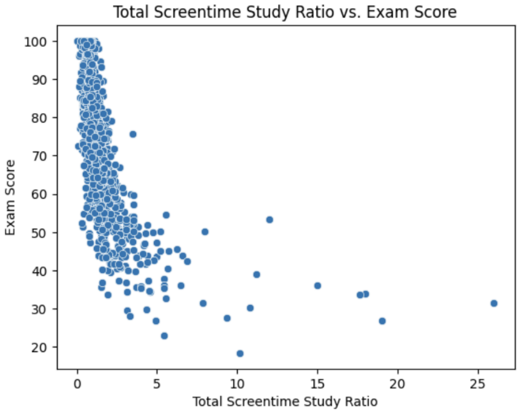


The barplots above show average exam scores by part-time job status, diet quality, and extracurricular participation. Across all categorical features and categories within them, differences in average exam scores are minimal. Although these lifestyle-related categorical variables show limited impact on exam scores individually, they will be included in the models to explore potential variable interactions and their combined predictive value.

Feature Engineering & Visualization:

To more comprehensively assess how lifestyle habits influence exam scores, I performed feature engineering using existing features from the raw dataset.

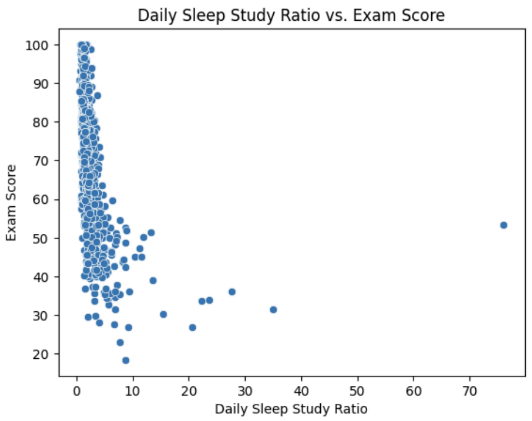
1. *Screentime - Study Ratio*

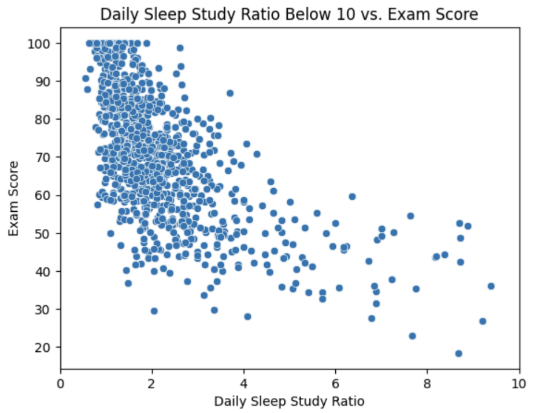


Screentime-study ratio is derived from dividing total screentime (social media hours + netflix hours) by daily study hours. A lower ratio suggests the student dedicates relatively more hours daily to studying than to screen-based entertainment. A higher ratio indicates the student spends relatively more time daily on screen-based entertainment than on studying.

This scatterplot shows a strong negative correlation between screentime-study ratio and exam scores, suggesting students who spend more time on screen-based entertainment than studying tend to achieve a lower exam score. Most high-performing students (scores above 75) exhibit low screentime–study ratios.

2. *Sleep - Study Ratio*

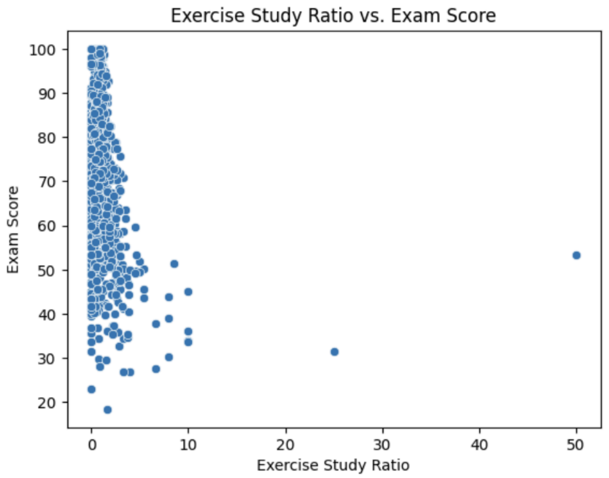




Sleep-study ratio is derived from dividing daily sleep hours by daily study hours. A lower ratio may indicate that the student spends more time daily studying relative to sleeping, while a higher ratio may suggest an overemphasis on sleep.

The scatterplots above show a strong negative correlation between sleep-study ratio and exam scores, suggesting students who sleep significantly more than they study tend to achieve a lower exam score. This relationship indicates that excessive sleep may reduce study time, negatively impacting academic performance.

3. *Exercise - Study Ratio*



Exercise-study ratio is derived from dividing weekly exercise frequency by daily study hours. A lower ratio indicates that the student is dedicating relatively more time to studying than exercising, while a higher ratio may suggest an overemphasis on exercise.

This scatterplot shows that lower exercise-study ratios are the most common. However, they correspond to a wide range of exam scores, suggesting that spending more time studying relative to exercising is not a strong predictor of exam performance on its own.

EDA Conclusion:

Based on exploratory data analysis, I selected the following features for modeling:

- Study hours per day
- Social media hours per day
- Netflix hours per day
- Whether the student has part part-time job
- Attendance percentage
- Sleep hours per day
- Diet quality
- Weekly exercise frequency
- Extracurricular Participation
- Screen Time study Ratio
- Sleep study ratio
- Exercise study ratio

I believe these independent variables effectively capture key aspects of students' lifestyle habits. In contrast, features like parental education level, internet quality, and mental health rating generally fall outside the scope of personal habit-related behaviors and were therefore dropped from modeling.

Models and Methods Overview

To predict exam scores, I decided to use multiple different regression models and examine which one performs the best in predicting students' exam scores and accounting for variations in my dataset. For each of these models, I used an 80-20 train-test split, training my model on 80% of the data and testing it on the remaining 20%. I evaluated the performance of each of my models by comparing its mean squared error to that of the baseline. My baseline value, which is 275.97, is the mean squared error of the mean exam score of my dataset. Mean squared error measures the average squared difference between observed and predicted values, efficiently penalizing larger errors more significantly. This characteristic makes MSE effective for identifying models that demonstrate consistent prediction accuracy while revealing those prone to larger errors.

Additionally, I used permutation importance to assess the impact of each feature on the model's predictive performance. Permutation importance measures the degree to which the model's performance declines when the values of a feature are randomly shuffled, thereby identifying the predictor that most strongly influences exam scores. Since this technique is applicable to any model, it provided a consistent method to evaluate feature relevance across all models used in this project.

Multiple Linear Regression Model:

First, I chose to build a Multiple Linear Regression model to analyze the relationship between a dependent variable and multiple independent variables simultaneously. This model served as an appropriate starting point, as I believed the selected predictors may collectively influence exam scores. Ultimately, the multiple linear regression model allowed me to assess both the individual and combined effects of each independent variable on exam scores through an interpretable framework.

Lasso Regression Model:

Next, I chose to try a Lasso Regression model, as it is capable of shrinking the coefficients of less important features to exactly zero, effectively removing less relevant features that may have distracted the performance of my Multiple Linear Regression model previously. This helps reduce model complexity and the risk of overfitting, especially with many independent variables. Based on the feature importance from the previous model, some features appeared to have minimal influence on exam scores. Lasso Regression offers a way to account for these less relevant features, allowing me to focus on the predictors that contribute most to exam scores.

K-Nearest Neighbors Regression Model:

Then, I chose to use a K-Nearest Neighbors (KNN) Regression model. Unlike linear regression models, KNN makes predictions based on the similarity of data points. My exploratory data analysis revealed several non-linear relationships and localized clusters between the independent variables and exam scores. As such, KNN would be effective in capturing these complex patterns in my data that linear regression models might overlook.

Decision Tree Regression Model:

Then, I chose to implement a Decision Tree Regression model. Like KNN, decision trees can capture non-linear relationships, but they do so by recursively splitting the data based on feature thresholds rather than relying on distance alone. They are also robust at handling irrelevant features, which can help reduce the risk of overfitting. Lastly, they offer a clear visual representation of the decision-making process, making it easier for me to understand how predictions are made and which features most influence exam scores.

Random Forest Regression Model:

Lastly, I decided to build a Random Forest Regression Model. As an ensemble method, random forest combines predictions from multiple decision trees to produce more stable outputs compared to a single decision tree. This makes it effective in addressing non-linear relationships and handling complex interactions among features, ultimately offering a more comprehensive picture of how various predictors collectively impact exam scores.

Results and Interpretation

Multiple Linear Regression Model:

My multiple regression model performed better than my baseline, with an MSE of 56.50 on the training data and 63.90 on the testing data, which are both lower than the baseline MSE. Overall, the model made more nuanced predictions than the baseline by accounting for the combined effects of multiple predictors. However, the disparity between the training and testing MSE suggests the model may be slightly overfitting. The higher testing MSE could be attributed to noise or outliers in the training set that the model failed to generalize; it is plausible given the frequent presence of nonlinear patterns observed in my EDA, which multiple linear regression model does not capture well. This prompted me to use a Lasso Regression model, which can more efficiently address overfitting through L1 regularization.

By evaluating the permutation importance, the input that was most important in the multiple regression model was a student's daily study hours (1.145), followed by weekly exercise frequency (0.067) and daily social media hours (0.049). In contrast, extracurricular participation, part-time job status, diet quality, and sleep-study ratio contributed the least to predicting a student's exam score in this model.

Lasso Regression Model:

My Lasso Regression model outperformed the baseline, with a training MSE of 56.62 and a testing MSE of 63.79. However, it showed only minimal improvement over the multiple regression model despite eliminating less important features using L1 regularization. I believe this limited improvement is because Lasso Regression, like Multiple Regression, assumes linear relationships between independent variables and the dependent variable. However, as revealed in my EDA, many relationships present in the data appear to be non-linear – variations that linear models may fail to address. This prompted me to build a K-Nearest Neighbors Regression model, which does not assume linearity and is well-suited for capturing non-linear relationships or localized patterns.

Once again, the permutation importance analysis showed that a student's daily study hours (1.143) is the most important feature in predicting exam scores, followed by weekly exercise frequency (0.072) and daily social media hours (0.042). Extracurricular participation, part-time job status, diet quality, and sleep-study ratio all had a permutation importance of 0, reinforcing the results from the multiple regression model and suggesting they add minimal value to the model's predictive ability.

K-Nearest Neighbors Regression Model:

My KNN model performed better than my baseline, with a training MSE of 72.48 and a testing MSE of 91.51. However, it performed significantly worse than my multiple and lasso regression models. While KNN model can capture non-linear relationships, its reliance on the proximity of data points in feature space may have caused it to overfit, capturing noise and overly specific local patterns that do not generalize well to unseen data. Additionally, KNN often struggles with high-dimensional data, which could be the case in my model, and does not naturally account for feature interactions, leading to its weaker performance. To overcome these limitations, I decided to build a Decision Tree Regression model, which can capture non-linear relationships and incorporate feature interactions without depending solely on data proximity.

By evaluating permutation importance, daily study hours (0.551), once again, emerged as the most important predictor of exam scores, followed by screentime-study ratio (0.097) and weekly exercise frequency (0.065). Notably, the permutation importance of daily study hours decreased significantly compared to previous models (1.14), suggesting that the KNN model, likely due to its localized approach, relied less on this feature than my previous models. Extracurricular participation, part-time job status, and diet quality all showed negative importance, suggesting they added noise rather than meaningful predictive value to the KNN model.

Decision Tree Regression Model:

My decision tree regression model outperformed the baseline, with a training MSE of 64.45 and a testing MSE of 101.06. However, it performed worse than my previous models, particularly on the testing data. The larger difference between the testing and training MSE suggests the model is overfitting, which may be partly due to the model’s shallow maximum depth of 4. As seen in the visualization, the tree relied heavily on daily study hours and screentime-study ratio while completely discounting the majority of other independent variables. Although some independent variables have shown lower permutation importance, they may still have contributed to the model; disregarding them entirely could have weakened the model's generalizability and contributed to its weaker performance on the testing data. To enhance Decision Tree Regression's performance, I decided to try Random Forest Regression model, which produces predictions from a collection of decision trees.

Similar to previous models, daily study hours (0.930) remained the most influential variable in predicting exam scores, followed by screentime-study ratio (0.214) and weekly exercise frequency (0.031). Part-time job status, diet quality, extracurricular participation, and sleep-study ratio all had a permutation importance of 0, consistent with previous models. However, unlike other models, daily social media hours, netflix hours, and attendance percentage also had a permutation importance of 0, indicating they added no value to the decision tree model despite showing some relevance to previous models.

Random Forest Regression Model:

My random forest model outperformed the baseline, with a training MSE of 17.76 and a testing MSE of 78.88. I optimized it using cross-validation and grid search, fine-tuning hyperparameters with a maximum depth of 8 and 150 estimators. Compared to previous models, it performed better than my KNN and Decision Tree models but not as well as my Multiple and Lasso Regression models. I believe Random Forest achieved higher predictive accuracy than my KNN and Decision Tree models because, in addition to capturing non-linear relationships, it was also able to reduce variance and improve stability by averaging predictions across multiple decision trees. While Random Forest's testing MSE was much lower than that of KNN and Decision Trees, this model had the largest disparity between the testing and training MSE. This difference suggests that the model is overfitting the training data. Its complexity, despite its strength in handling non-linear patterns and feature interactions, may have caused it to not generalize well to unseen data, resulting in weaker performance compared to Multiple and Lasso Regression models that have a simpler framework.

Through permutation importance analysis, daily study hours (0.846) is, once again, identified as the most influential predictor of exam scores, followed by screentime-study ratio (0.211) and daily sleep hours (0.028). However, two notable differences emerged in this model: daily netflix hours (-0.002) showed a negative permutation importance, indicating it may have added noise, while weekly exercise frequency (0.014) was assigned much less importance than in previous models.

Comparison of Model Performance:

All of the models I developed demonstrated improved performance over my baseline prediction, suggesting their overall usefulness in predicting exam scores. Among them, Lasso Regression model proved to be the most effective, achieving the lowest testing MSE (63.79). This slight improvement over my Multiple Regression model (63.90) is likely due to Lasso Regression's ability to simplify the model through L1 regularization, selecting only the most important features. Random Forest Regression model achieved the lowest training MSE (17.76) but had a much higher testing MSE (78.88). Its ability to handle non-linearity and high-dimensionality allowed it to outperform KNN and Decision Tree; however, it did not generalize as well as the linear models. The Decision Tree Regression model performed the worst overall, with a testing MSE of 101.06, which is likely due to its overreliance on a few variables.

Looking at feature importance, daily study hours consistently emerged as the most important variable in predicting exam scores across all models. In the Lasso Regression model, daily study hours had the highest permutation importance (1.143), far exceeding the next most influential variable, weekly exercise frequency (0.072). This suggests that sufficient daily study hours has a substantial impact on exam performance, highlighting the importance for students to prioritize consistent study routines while complementing them with moderate physical activities. While weekly exercise frequency ranked second in importance in Multiple and Lasso Regression models, screetime-study ratio emerged as the second most important feature in predicting exam scores in the rest of the models. Notably, extracurricular participation, part-time job status, diet quality, and sleep-study ratio consistently showed negligible importance across all models, suggesting they have minimal influence on exam scores.

Conclusion and Next Steps

Summary of Findings:

Through this project, I applied various regression models to predict exam scores and identified critical features impacting a student's academic performance. Among all models used, the Lasso Regression model performed the best, achieving the lowest testing MSE (63.79). Its ability to perform variable selection and regularization by eliminating irrelevant or redundant features proved particularly effective in managing high dimensionality and complex patterns within the data. Daily study hours is shown to be pivotal in predicting exam scores, highlighting a strong relationship between study time and academic performance. Features, such as diet quality and extracurricular participation, had minimal impact on exam scores. In conclusion, this project offers meaningful insights into the effectiveness of various regression models and highlights the relative importance of lifestyle-related factors in predicting academic performance, providing a useful foundation for future data analysis and predictive modeling in the educational sphere.

Implications:

The findings emphasize the importance of dedicating time daily to studying as a key factor in improving academic performance. Students could benefit from consistently allocating time each day to studying, building a study schedule that could support long-term academic growth. Educators and schools could focus on offering structured study sessions, extended office hours, or guided study programs. Parents, too, could contribute by setting aside time daily to study with their children, fostering a supportive home environment that encourages motivation and educational engagement. Overall, these insights could inform the development of educational materials and programs that cultivate productive study habits across student populations.

Next Steps & Future Improvements:

To improve the predictive capabilities of my models and better understand factors influencing academic performance, several improvements can be made:

- Testing Additional Models:
    - I think it would be beneficial to use other regression models, such as Elastic Net Regression and Gradient Boosting, to improve accuracy by handling feature selection, possible multicollinearity, and non-linear relationships within the data.
- Addressing Overfitting:
    - Since all the models I tested showed signs of overfitting, further hyperparameter tuning and the use of stronger regularization techniques could help enhance generalizability.
- Feature Engineering:
    - I think incorporating more detailed features, like sleep and study qualities or specific extracurricular activities, could offer deeper insights. Additionally, refining the current dataset by removing less relevant or redundant features may enhance model performance.

I would also want to incorporate these additional features:

- Sources of External Support:
    - While my current analysis focuses on personal lifestyle habits, integrating variables such as tutoring frequency, study group participation, and access to mentoring programs could provide valuable context. These features may help explain academic outcomes beyond individual behaviors.
- Family Background:
    - I think it would be interesting to explore how a student's family background, an uncontrollable factor, impacts their academic performance, as it can affect the resources and support available to the student. Relevant features could include parental education levels (included in the original dataset), family income, and number of siblings.

By adding these features, I would be able to refine my models and gain a more holistic view of the factors influencing exam scores. Future research could also build upon these findings, potentially offering more targeted, data-driven insights for educators, schools, and parents to better support student success.