Predicting Students’ Exam Scores

Introduction & Data Description

Problem Statement & Objective:

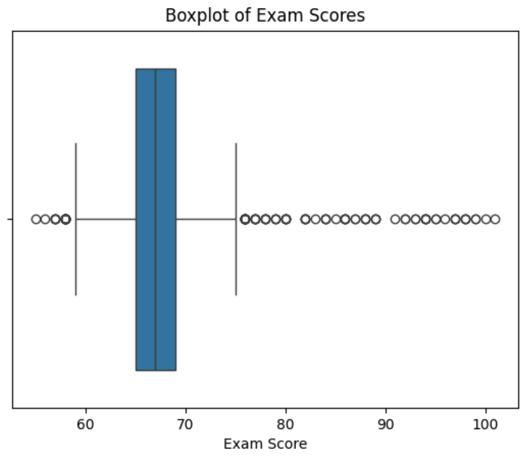
Academic performance is a key concern for students, educators, and policymakers, with exam scores serving as a widely used metric to assess success. High scores can open doors to scholarships and competitive programs, while low scores highlight areas for improvement and intervention. Exam performance also has broader societal implications, influencing long-term outcomes like employability and earning potential. This project aims to predict students' **exam scores** using **regression**, exploring how various factors like study habits, attendance, and extracurricular activities impact performance. By analyzing these relationships, the project seeks to provide actionable insights to help students optimize their academic strategies and inform educational support initiatives.

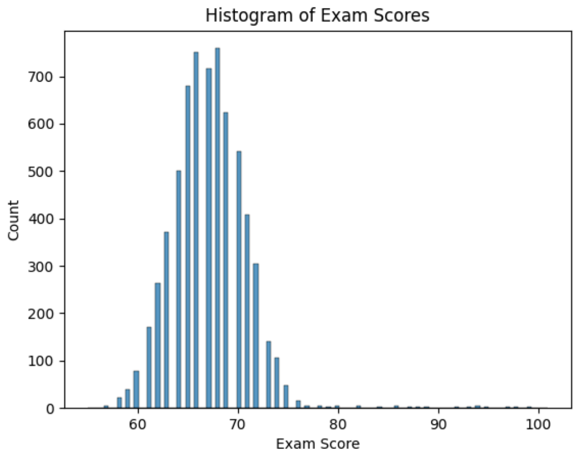
Dataset Overview:

The dataset used for this project is sourced from Kaggle in CSV format and contains 6,607 rows and 20 columns. It will require some cleaning, given the presence of potentially unnecessary columns and null values within the dataset. The dataset includes both numeric and categorical features that capture various aspects of student behavior and background. Key features include hours studied, attendance, sleep hours, and participation in extracurricular activities, among others. The target variable is exam scores, which are continuous numerical values. This dataset provides a comprehensive foundation for exploring how these factors interact to influence academic performance.

Exploratory Data Analysis (EDA)

After cleaning the dataset for null values by dropping rows with missing values, there are 6,378 rows. The exam scores range from a minimum of 55 to a maximum of 101. The mean exam score is 67.25, and the median score is 67.

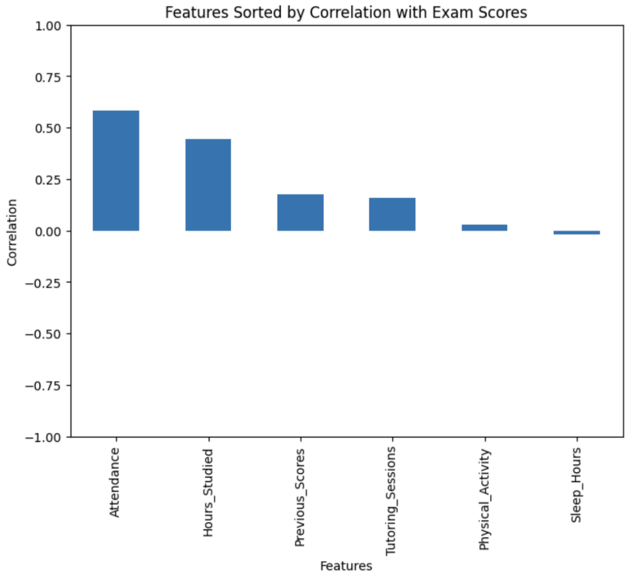




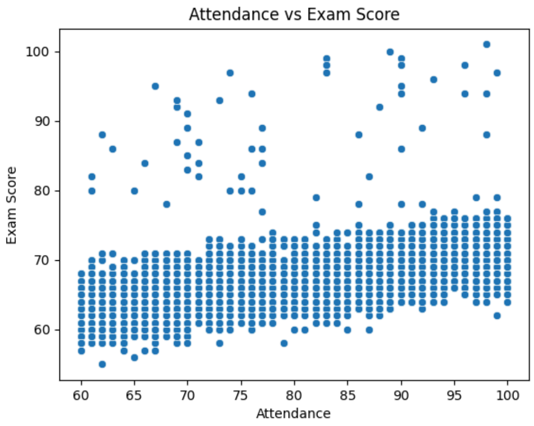
Based on the boxplot and histogram of exam scores, the distribution of exam scores is roughly normally distributed with a slight right skew. Given the mild nature of the skewness, the mean not being that much greater than the media, the majority of the data following a normal distribution, and the use of non-linear models on top of linear regression, a log transformation was deemed not absolutely necessary for the scope of this project.

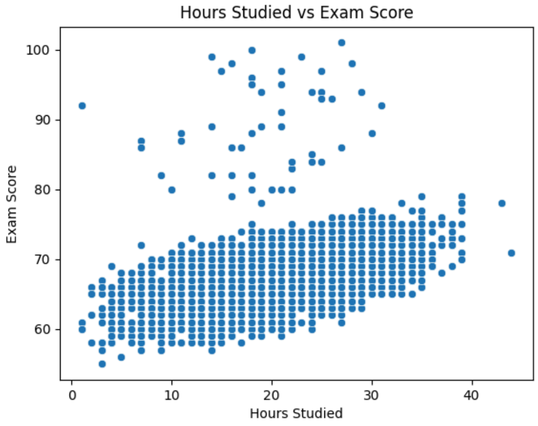
Numerical Features:

The graph below illustrates the correlation of numerical features (i.e. hours studied, attendance, sleep hours, previous scores, tutoring sessions, and physical activity) and exam scores.

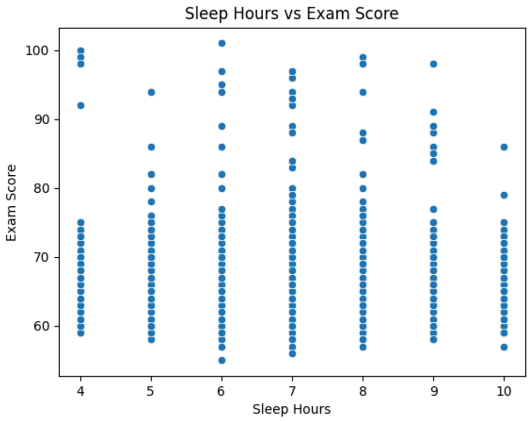


Attendance and hours studied have the highest correlation with exam scores. Attendance has a positive correlation of 0.58 meaning a higher attendance rate leads to higher exam scores, and hours studied has a positive correlation of 0.45 meaning higher hours studied leads to higher exam scores. However, these are still moderate positive correlations, not extremely strong. This suggests that while attendance and hours studied are important, they are not the only factors influencing exam scores. The correlations of previous scores and tutoring sessions are 0.18 and 0.16, respectively, and although these are low positive correlations, they may still contribute to the model in combination with other features. However, the remaining features of physical activity and sleep hours are 0.028 and -0.017, respectively, which are both close to zero. If these do not have a non-linear relationship with exam scores either, it does not make sense to include them in my models as there is no suggestion more or less of these features contribute to higher or lower exam scores. Thus I will make scatterplots of each of these numerical features to see if there may be any non-linear relationships between the features and exam scores.





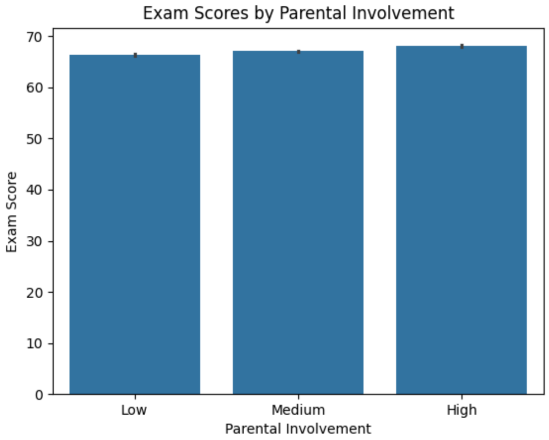
These scatterplots illustrate that attendance and hours studied do indeed have a linear relationship with exam scores.

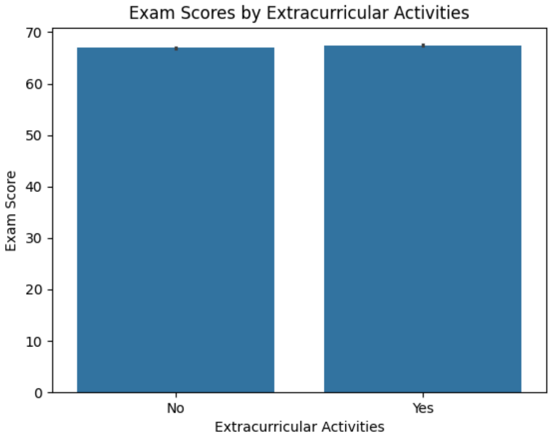


However, like the scatterplot on sleep hours on the left, the rest of the numerical features did not show obvious signs of any non-linear type of relationship with exam scores. Since the amount of physical activity and number of sleep hours don't suggest significant linear or non-linear relationships with exam scores, they will not be included in the model to reduce redundancy and the risk of overfitting, allowing the model to focus on features with stronger predictive power.

Categorical Features:

For the categorical features, I made bar plots to assess if any of those features seem to contribute to higher or lower exam scores. Differences in family income, teacher quality, school type, parental education level, and gender do not show any impact on exam scores.



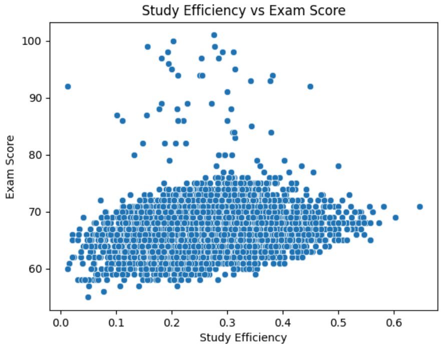


Like these bar plots, the rest of the features (i.e. parental involvement, access to resources, extracurricular activities, motivation level, internet access, peer influence, learning disabilities, distance from home) exhibit slight trends where one would expect; higher levels of parental involvement, access to resources, motivation, doing extracurricular activities, having internet access, not having learning disabilities, living closer to school, and positive peer influence are related to slightly better scores, but just by a few points. Although these features do not show significant differences in exam scores overall, they will be included in the initial model to explore their potential combined effects and interactions with other variables.

Feature Engineering:

To explore other possibly more important features not included in the raw data, I did some feature engineering:

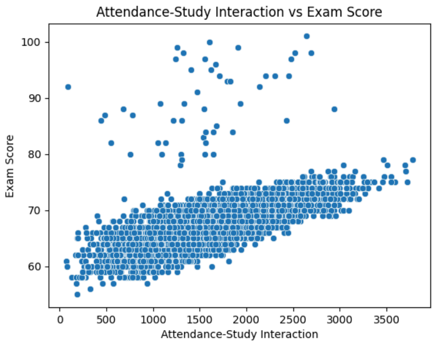
Study Efficiency (divides hours studied by attendance):



A higher study efficiency indicates that the student studies more hours relative to their attendance (e.g. reliance on self-studying).

A lower study efficiency indicates that the student studies fewer hours relative to their attendance (e.g. dependence on attendance without sufficient self-study).

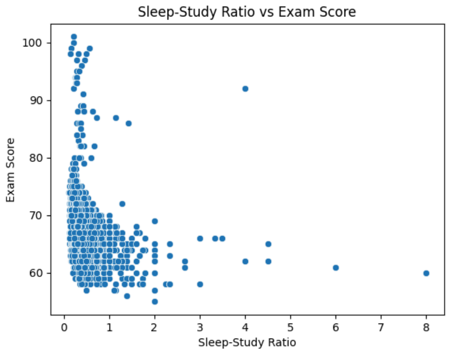
Attendance-Study Interaction (multiplies attendance and hours studied instead):



A higher Attendance-Study Interaction suggests that a student regularly attends classes while also dedicating significant time to studying outside of class.

A lower Attendance-Study Interaction suggests that both attendance and study hours are low (the student may be less academically engaged).

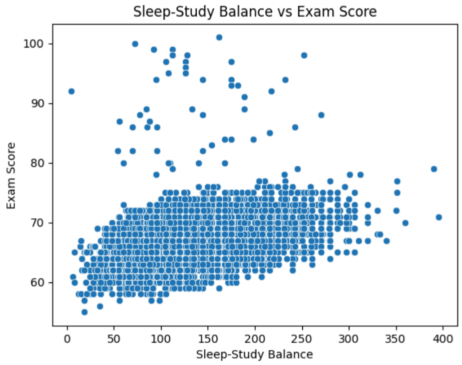
Sleep-Study Ratio (divides sleep hours by hours studied):



A high ratio may suggest an overemphasis on sleep, potentially leading to insufficient study time.

A low ratio may suggest an overemphasis on studying, potentially leading to sleep deprivation and reduced cognitive function.

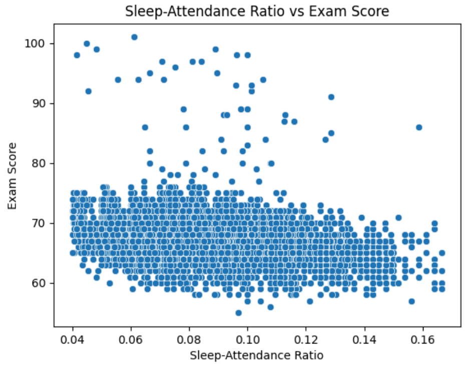
Sleep-Study Balance (multiplies sleep hours and hours studied instead):



A higher Sleep-Study Balance indicates that the student is both getting enough rest and dedicating sufficient time to studying.

A lower Sleep-Study Balance could result from lower habits for both.

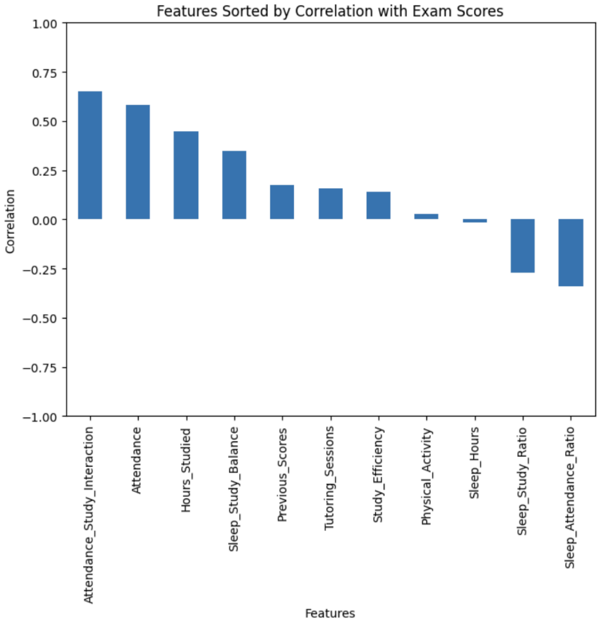
Sleep-Attendance Ratio (divides sleep hours by attendance):



A high ratio may indicate an overemphasis on sleep at the expense of attendance, which might suggest lower engagement in structured academic activities.

A low ratio may indicate an overemphasis on attendance at the expense of sleep, which might lead to fatigue and reduced effectiveness in learning.

The scatterplots for the new features all show linear relationships between the features and exam scores, so I used correlations to evaluate their significance.



Attendance-study interaction, sleep-study balance, and sleep-attendance ratio all had moderately high correlations of 0.65, 0.35, and -0.34, respectively.

Sleep-study ratio and study efficiency had lower correlations of -0.27 and 0.14, respectively. Since these correlations all have some significance and are not that close to zero, they may still contribute to the model.

Conclusion of EDA:

Overall, I will include these features to begin modeling: attendance, hours studied, previous scores, tutoring sessions, parental involvement, access to resources, extracurricular activities, motivation level, internet access, peer influence, learning disabilities, distance from home, attendance-study interaction, sleep-study balance, sleep-attendance ratio, sleep-study ratio, and study efficiency. In other words, I will drop these features: physical activity, sleep hours, family income, teacher quality, school type, parental education level, and gender.

Models and Methods

To predict exam scores, I decided to use multiple different regression models and see which one performs the best for predicting the scores. My “X” variable will include the features mentioned in the conclusion of my EDA, and my “y” variable will be the target, exam scores. Because my data is moderately sized, I decided to use an 80-20 train test split (training my model on 80% of the data and then testing it on the remaining 20%) for each of these models.

First, I chose to build a **multiple linear regression model** to predict exam scores using multiple independent variables. This method allows me to examine how each predictor individually influences exam scores while also considering their combined effects. By assuming a linear relationship between the predictors and the target variable, multiple linear regression provides a straightforward and interpretable framework to understand how various factors contribute to academic performance.

Next, I chose to use a **K-Nearest Neighbors (KNN) model** to explore whether localized relationships in the data could better predict exam scores. Unlike linear regression, KNN makes predictions based on the similarity of data points, which may help capture patterns in small clusters of students with similar characteristics, such as study habits or attendance levels. Since KNN does not assume a linear relationship between predictors and the target variable, it can potentially reveal insights that a linear model might miss. For this model, I applied feature scaling since KNN relies on distance metrics, ensuring that all features contributed equally to the predictions.

After that, I used a **Random Forest model** to leverage its ability to handle complex interactions between features and capture non-linear relationships. Random Forest is an ensemble method that combines multiple decision trees, which makes it robust to overfitting and effective in identifying the most important predictors. This model is particularly valuable in understanding the role of interactions between features like attendance and hours studied, or the cumulative effect of less significant predictors, in determining exam scores.

Lastly, I chose to implement **Lasso Regression** to focus on feature selection and regularization. Lasso applies an L1 penalty to shrink less important feature coefficients to zero, effectively selecting only the most impactful predictors. This helps simplify the model, especially when dealing with a dataset that includes many features, and reduces the risk of overfitting. By using Lasso, I aim to identify which variables contribute most significantly to predicting exam scores while discarding irrelevant or redundant information.

I will evaluate the success of each of my models with the evaluation metric of mean squared error (MSE). I will compare each model’s MSE with the baseline MSE value of 16.86. MSE quantifies the average squared difference between the actual and predicted values, penalizing larger errors more heavily than smaller ones; this makes MSE particularly useful for identifying models that consistently produce accurate predictions, while highlighting those that may struggle with larger deviations.

Additionally, I will use permutation importance to assess the contribution of each feature to the model's predictive power, identifying which predictors most strongly influence exam scores. Permutation importance measures the decrease in model performance when the values of a specific feature are randomly shuffled, thus isolating the effect of that feature. Permutation importance can be applied to any model, making it an ideal choice for understanding which predictors most strongly influence exam scores across all models.

Results and Interpretation

Multiple Regression Model:

My multiple regression model performed better than my baseline overall, as both the training data MSE (4.34) and the testing data MSE (5.44) were lower than the baseline MSE (16.86). The results suggest that a linear relationship captures some of the variation in exam scores, but there may still be non-linear relationships or localized patterns that a linear model cannot identify. This prompted me to explore a K-Nearest Neighbors (KNN) regression model, which does not assume linearity.

From the permutation importance analysis, the two most important features were Attendance (0.574) and Hours Studied (0.380). These features had significantly higher importance scores compared to the other predictors, indicating their strong influence on exam scores. Attendance had the highest importance score, suggesting that consistent class attendance has a substantial impact on exam performance. This aligns with expectations, as students who attend class regularly likely benefit from greater exposure to instruction and structured learning environments. Similarly, Hours Studied was the second most important predictor, confirming that dedicated study time contributes to better academic outcomes. However, its importance score was lower than Attendance, which may indicate that quality of instruction and classroom participation play a more critical role than self-study alone. Features like Sleep-Study Balance and Study Efficiency had near-zero or negative importance scores, indicating that these derived features did not contribute meaningfully to predicting exam scores.

K-Nearest Neighbors Regression Model:

My KNN regression model performed better than my baseline model, with the training data MSE (5.37) and the testing data MSE (7.20) both lower than the baseline MSE. KNN’s ability to make predictions based on the similarity of data points allowed it to capture localized patterns. However, the slightly higher training MSE compared to testing MSE could be due to noise or outliers in the training set that don’t generalize to the test data. Additionally, KNN struggles with high-dimensional data and does not naturally handle feature interactions. To address these limitations, I decided to try a Random Forest regression model, which excels at capturing feature interactions and non-linear relationships.

The permutation importance analysis revealed that Attendance (0.244) and Attendance-Study Interaction (0.121) were the most influential predictors of exam scores, emphasizing the importance of consistent classroom engagement and a balanced approach to studying. Features like Internet Access and Sleep-Study Ratio, with negative importance scores, appeared to detract from the model's performance.

Random Forest Regression Model:

My random forest regression model performed better than my baseline model, with the training data MSE (1.62) and the testing data MSE (7.21) both lower than the baseline MSE. I optimized the model using cross-validation, selecting a maximum depth of 10 and 200 estimators. These hyperparameters provided a balance between model flexibility and overfitting. The large disparity between the training and testing MSE indicates that the model may be slightly overfitting the training data, which is expected for Random Forest due to its ability to fit training data almost perfectly. While Random Forest effectively handles feature interactions and ranks feature importance, its complexity can make it less interpretable. To address these concerns and simplify the model, I used Lasso Regression.

The permutation importance analysis revealed that the Attendance-Study Interaction (0.551) was the most influential predictor of exam scores, indicating that the combined effect of high attendance and effective study habits is critical for academic success. Attendance alone (0.301) was the second most important feature, further emphasizing the importance of consistent classroom engagement. Interestingly, Hours Studied (-0.0009) and Sleep-Attendance Ratio (-0.002) showed slightly negative importance, suggesting they added noise rather than meaningful predictive value. This highlights the need to focus on features with clear, interpretable relationships to academic performance.

Lasso Regression Model:

My lasso regression model performed better than my baseline model, with the training data MSE (4.34) and the testing data MSE (5.43) both lower than the baseline MSE. By applying L1 regularization, Lasso simplifies the model by selecting only the most important features, which helps reduce overfitting. The slight difference between training and testing MSE reflects the trade-off between simplicity and predictive accuracy inherent in regularization. Compared to Random Forest, Lasso provides a simpler and more interpretable model that highlights the key predictors influencing exam scores.

The permutation importance analysis for lasso regression highlighted Attendance (0.550) as the most influential predictor of exam scores, with a wide margin over the next most important feature, Hours Studied (0.200). This emphasizes the critical role of classroom engagement and consistent independent study in academic success. Study Efficiency (0.000) and Sleep-Attendance Ratio (-0.0012) had negligible or negative importance scores, indicating they added little to no value or even detracted from the model's predictive power.

Comparison of Results:

When comparing the models, Lasso Regression emerged as the most effective overall. It achieved the lowest testing MSE (5.43), nearly matching the performance of the Multiple Regression Model (5.44) while simplifying the feature set through regularization. The Random Forest Regression Model, while achieving the lowest training MSE (1.62), had a higher testing MSE (7.21), indicating overfitting despite hyperparameter tuning. Similarly, the KNN Regression Model had a testing MSE of 7.20, which, while lower than the baseline, struggled with noisy or high-dimensional data, resulting in a less robust performance compared to Lasso and Multiple Regression.

In terms of feature importance, all models consistently identified Attendance as the most critical predictor of exam scores, though the importance score varied across models. Lasso Regression ranked Attendance highest (0.550), followed by Hours Studied (0.200), highlighting the importance of classroom engagement and independent study. The Random Forest Model also prioritized Attendance-Study Interaction (0.551) and Attendance (0.301), reflecting the combined influence of attendance and study time. In contrast, Multiple Regression and KNN Regression placed greater emphasis on Attendance and Hours Studied but did not highlight interactions as strongly. Notably, features like Study Efficiency and Sleep-Attendance Ratio consistently showed negligible or negative importance across models, underscoring their limited predictive value. Ultimately, while Lasso Regression provided the most balanced combination of predictive accuracy and interpretability, Random Forest excelled at identifying feature interactions.

Conclusion and Next Steps

Summary of Findings:

This study explored the prediction of exam scores using regression models and identified key factors influencing academic performance. Among the models tested, Lasso Regression demonstrated the best balance of predictive accuracy and interpretability, achieving the lowest testing MSE (5.43) and effectively identifying the most important predictors while simplifying the feature set. Attendance consistently emerged as the most critical predictor across all models, with its direct impact on exam scores reinforced by its interaction with Hours Studied in models like Random Forest. While Hours Studied was also a significant contributor, its importance was consistently ranked lower than attendance, suggesting that quality classroom engagement may outweigh self-study alone. Features like Study Efficiency and Sleep-Attendance Ratio showed minimal or negative importance, indicating they may not meaningfully contribute to predicting exam scores.

Implications:

The findings emphasize the importance of promoting consistent classroom attendance as a critical factor for academic success. Schools and educators could focus on strategies to improve attendance rates, such as tracking and supporting students with frequent absences.

Additionally, encouraging structured and effective study habits could complement classroom engagement, providing students with a holistic approach to improving performance. The insights from this analysis could also be used by parents and educators to allocate resources and design interventions, such as targeted tutoring programs or workshops on time management, aimed at fostering both attendance and independent study.

Next Steps:

To further enhance the analysis and its applicability, several steps could be undertaken:

- Testing Additional Models:
    - Explore other regression models, such as Gradient Boosting or Support Vector Regression, to further evaluate non-linear relationships and improve predictive performance.
- Feature Engineering:
    - Refine or create new features based on domain knowledge, such as incorporating qualitative measures of study quality or detailed attendance patterns.
    - Address redundant or uninformative features identified in this study to improve model efficiency.
- Evaluating the Impact of Skewness:
    - Given the slight right skew in the distribution of exam scores, a log transformation could be applied to the target variable and compared against its performance with the original data. This transformation can help reduce the influence of outliers, and ensure that models, such as linear regression, better adhere to assumptions of normality.
- Addressing Overfitting:
    - Fine-tune hyperparameters and test regularization techniques in models like Random Forest to mitigate overfitting and improve generalizability.

By building on these findings and addressing these areas, future research can provide deeper insights into academic performance, supporting data-driven decision-making for educators, parents, and policymakers.