Student Performance Prediction Report

Shiyi Li, Supeng Yu, Shilong Wang

**Dataset**

[StudentPerformanceFactor](https://www.kaggle.com/datasets/lainguyn123/student-performance-factors)

**Abstract**

In the field of education, understanding the factors that influence students' academic performance is critical for optimizing learning outcomes. We aim to provide insights into how factors such as study habits, socio-economic status, parental involvement, and others impact students' academic success. The findings of this study can serve as a foundation for educators, policymakers, and parents to make informed decisions and interventions to improve student performance. Especially for parents and teachers, understanding these dynamics can offer practical guidance on nurturing a positive and productive learning atmosphere for students. Ultimately, the analysis provides a comprehensive perspective on the factors that drive student success and fosters a collaborative approach to enhancing education outcomes for all.

Specifically, we will employ Multiple Linear Regression and Random Forest models to analyze the data and address our problem statement effectively. This dataset, with its 6,607 records and 20 features, provides a robust foundation for our analysis, allowing us to apply data analytics and machine learning techniques to uncover meaningful insights and predictors of student success.

**Background**

Student performance is a multifaceted issue influenced by a combination of personal, familial, and institutional factors. Numerous studies have identified key factors that affect academic outcomes, such as study habits, attendance, prior academic achievement, socio-economic status, and access to educational resources. For instance, consistent attendance has been linked to better comprehension and retention of course material, leading to higher exam scores. Similarly, students with access to tutoring or additional learning resources often perform better due to the extra support they receive.

Machine learning techniques have become invaluable tools in educational data mining, allowing for the analysis of complex datasets to identify patterns and predictors of academic success. Multiple Linear Regression offers a straightforward approach to modeling the linear relationship between independent variables and exam scores, providing interpretable coefficients for each predictor. On the other hand, Random Forest models can capture non-linear relationships and interactions between variables, offering robustness against overfitting and the ability to handle high-dimensional data.

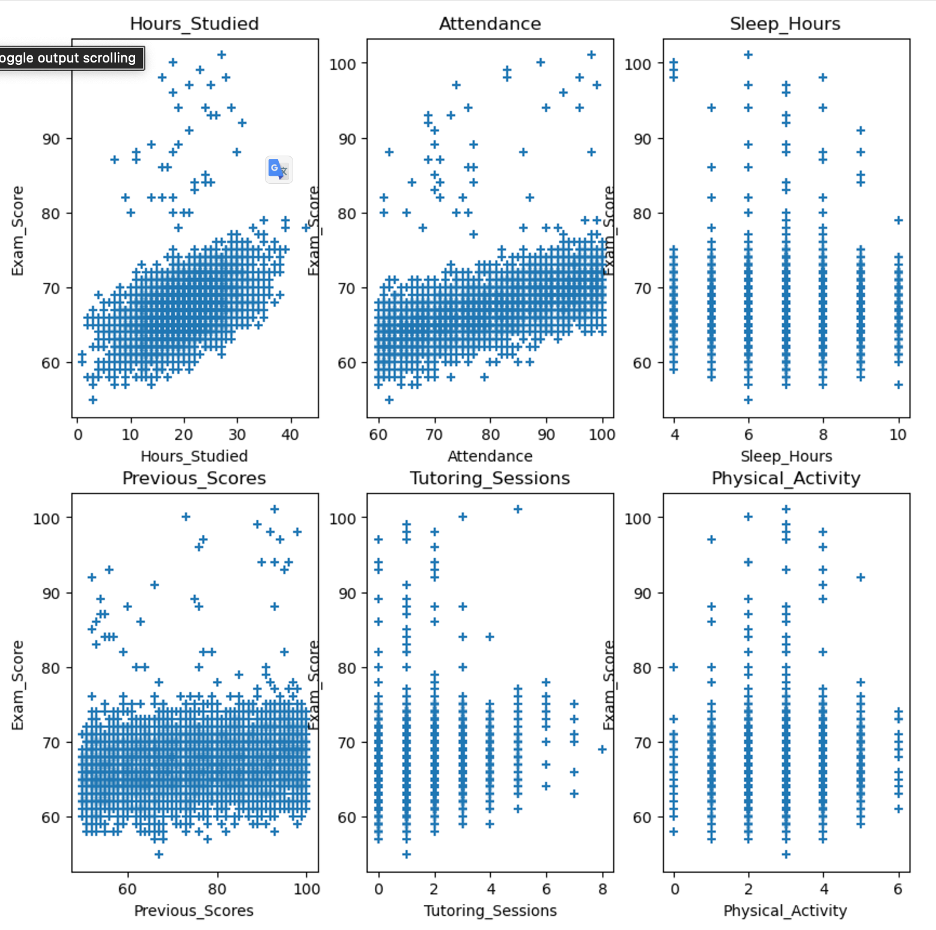
By applying these models to the Student Performance Factors dataset, we aim to determine which factors most significantly impact exam scores and assess the predictive power of each modeling approach.

**Preprocess the data set**

1. Identify the **shape** of this dataset: there are 20 columns and 6607 rows in this data set. Except the response variable Exam\_Score, there are 6 categorical variables and 13 numerical variables as the predictors.
2. Get rid of the rows with any **missing values** (remaining 6378 observations)
3. Convert categorical variables into numerical format using **one-hot encoding** to make them suitable for machine-learning models. (after one-hot encoding we got 40 predictor variables including dummy variables).
4. **Standardize** numerical to ensure all features contribute equally to model performance and to minimize issues caused by varying scales of data
5. Split the data set into a train set with 80% observations and a test set with 20% observations (**train set: 5102 observations, test set: 1276 observations).**

**Linear regression model analysis**

Multiple Linear Regression (MLR) is supervised learning, and it can handle multiple predictors efficiently, making it ideal for understanding and interpreting complex systems and making predictions. While it assumes linearity and is sensitive to outliers, it serves as a strong baseline model before considering more complex approaches.



Scatter Plots for Continuous Variables vs. Exam\_Score

First, to observe any relationship between the 6 continuous variables and the response variable, Exam\_Score, we plotted 6 scatter plots for them. The scatter plots suggest some strong positive linear relationships for Hours\_Studied, Attendance, and Previous\_Scores with Exam\_Score and some weak linear or unclear relationships for Sleep\_Hours, Tutoring\_Sessions, and Physical\_Activity with Exam\_Score, which may be due to the limited number of unique values for Sleep\_Hours, Tutoring\_Sessions, and Physical\_Activity. In addition, the plots also indicate some outliers in the dataset that deviate from the general trend. The outliers can be described from the plots by a few students who scored relatively low in the exam, but studied about 30 to 40 hours per week, had about 100 percent attendance, had very high previous scores of about 100, or had more than 6 tutoring sessions. Therefore, we need to carefully consider the effect of the outliers on our model results, and we can adjust the outliers if we get poor predictive accuracy to make the model better at predicting the "Exam\_Score" in the future.

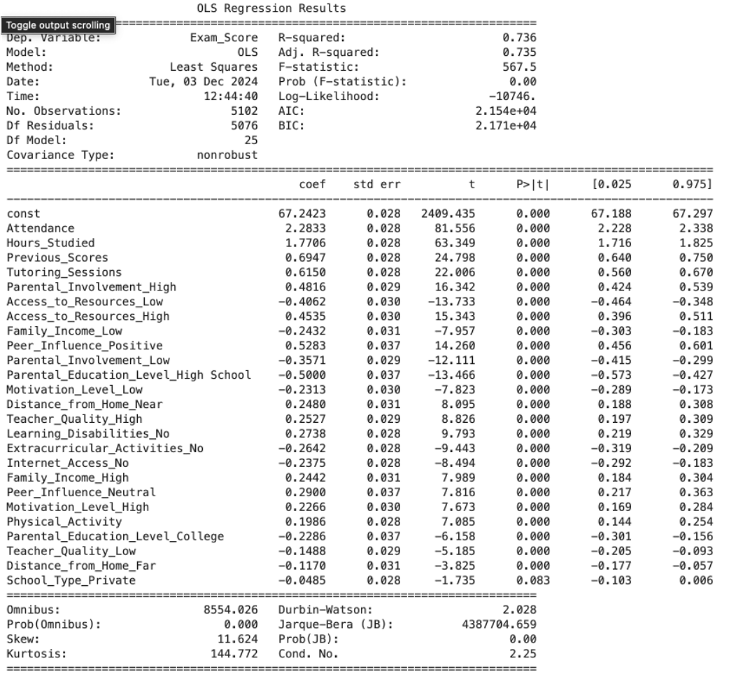
**Full MRL model Analysis**



Full Model Summary Output

Then, we used the standard scaling method to standardize the dataset and built a full multiple linear regression model with all the variables that we have in our dataset, including 6 numeric and 34 dummy variables after using the one-hot encoding method to deal with the 13 categorical variables. We obtained an r-squared value of 0.5740340101583764, suggesting that 57.40% of the variance can be explained by this full model, and a mean square error value of 7.832270134640281 on the test set, which means that the predictive accuracy of this full model is not bad. However, the summary output of this full model shows that only five predictors, Hour\_Studied, Attendance, Previous\_Scores, Tutoring\_Sessions, and Physical\_Activity, have small p-values, which are equal to 0.00, and the p-values for the other predictors are all very high. This suggests that these five predictors have significant effects on the Exam\_Score in this full model. Therefore, the overall performance of this model is not as good as expected, and some improvements are still expected.

**Reduced MLR Model analysis**



Stepwise Regression Result

To find a better multiple linear regression model, we used the stepwise regression method and ordinary least squares modeling to select the "best" subset of predictors for a better multiple linear regression model. As shown in the output, the selected "best" subset of predictors includes 5 continuous variables, Hour\_Studied, Attendance, Previous\_Scores, Tutoring\_Sessions, and Physical\_Activity, and 20 dummy variables. By building a reduced multiple linear regression model using the "best" subset of predictors, we obtained an R-squared value of 0.7364870905041896, indicating that 73.65% of the variance can be explained by this reduced model, which is higher than the full model, and a lower mean square error value of 5.526956754131526 on the test set compared to the full model, meaning that the predictive accuracy of this reduced model is better than the full model. In this reduced model, we have only one predictor, School\_Type\_Private, with a large p-value, which is equal to 0.083, and the p-values of the other predictors are all very small, suggesting that all the predictors in this reduced model have significant effects on the Exam\_Score, except School\_Type\_Private. Therefore, the performance of this reduced model is, overall, better than the full model that we discussed before.

**Conclusion**

Our final best linear regression model is the reduced model, which includes only 5 continuous variables and the other 20 dummy variables, with an r-squared value of 0.7364870905041896 and a mean square error value of 5.526956754131526 on the test set. In this reduced model, the coefficients of the predictors indicate that the percentage of attendance and the number of hours of study per week are the two factors that most influence the exam score, with the highest coefficients greater than 1 and all having a positive relationship with the exam score. Therefore, a higher percentage of attendance and more hours of study per week are the two most important factors for a student to obtain a high exam score and achieve student success.

**Random Forest model analysis**

Building upon the insights gained from MLR model we employed a Random Forest Regression model to capture more complex, non-linear relationships within the data. Random Forest operates by constructing numerous decision trees—each trained on a randomly selected subset of data and features—and then aggregating their predictions. This ensemble method often enhances prediction accuracy and reduces overfitting, making it well-suited for datasets that combine both numerical and categorical variables.

**Reduced MLR Model analysis**

Our dataset contained 20 features, including both continuous and categorical variables. Random Forest naturally accommodates high-dimensional data without the need for extensive feature selection and provides an intrinsic measure of feature importance. By averaging predictions from multiple trees, the model generally improves robustness and is less prone to fitting noise from a single training subset. This property made Random Forest a strong candidate for this analysis, complementing the linear methods and potentially offering improved predictive capabilities.

**Hyperparameter Tuning**

In this study, we extensively tuned the Random Forest hyperparameters to identify the combination that minimized the Mean Squared Error (MSE). We iteratively explored various configurations of the number of trees, maximum depth, feature consideration strategies, minimum leaf size, and bootstrap usage. We use the following configuration produced the best performance:

'n\_estimators': [100, 200, 300, 400, 500, 650, 800],​

'max\_depth': [10, 15, 20, None],​

'max\_features': ['auto', 'sqrt', 'log2', 19],​

'min\_samples\_leaf': [1, 2, 4],​

'bootstrap': [True, False]

**Analysis and Conclusion**

After evaluating multiple rounds of experimentation, the following parameter produced the best performance:

'n\_estimators': 650

'max\_depth': 15

'max\_features': 19

'min\_samples\_leaf': 1

'bootstrap': True

Evaluating the optimized Random Forest model on the test set yielded a Test MSE of 6.3978934824776585, indicating the average squared difference between predicted and actual exam scores. Furthermore, the R² on the training set was 0.6141101210793264, reflecting that approximately 61.41% of the variance in exam scores could be explained by the chosen features within this modeling framework.

In this study, the top five most important features determined by the Random Forest model were: ​Attention, Hours\_Studied, Previous\_Score, Tutoring\_Session and Sleep\_Hours. These findings closely align with those from the MLR analysis, reinforcing the conclusion that consistent class attendance, dedicated study time, strong prior academic performance, additional tutoring, and sufficient sleep collectively enhance student exam outcomes.

A graph with blue and white text

Description automatically generated

Random Forest Feature Importance Results

While this performance is respectable, it did not surpass the best results obtained by the reduced MLR model. The relatively modest improvement in MSE and R² suggests that, given the size and variability of our dataset, the complexity and non-linear modeling capability of Random Forest did not deliver a clear predictive advantage over the simpler linear approach.

**SHAP Framework**

To increase the model explainability, derived from cooperative game theory, SHAP provides a structured approach to understand the importance of individual features in a predictive model. It assigns each feature a "contribution value", called “SHAP value”, indicating its impact on the prediction.

**SHAP summary plot**

The SHAP summary plot for a Random Forest model is generated using the Tree SHAP algorithm, which efficiently calculates Shapley values to explain feature contributions. For each prediction, Tree SHAP evaluates how features influence the model's output by analyzing their impact on tree splits and aggregating their contributions across all trees in the Random Forest. The plot ranks feature by their global importance (y-axis) and shows the distribution of their SHAP values (x-axis) across all predictions. Each dot represents an individual SHAP value, with the color gradient indicating the actual feature values (e.g., red for high and blue for low). This plot provides insights into the overall importance and variability of features, along with their interactions, in driving model predictions.

**Result**

A screen shot of a graph

Description automatically generated

From the graph, we can see that Features are ranked by importance and with **Attendance**, **Hours Studied**, and **Previous Scores** being the top contributors to the predicted exam scores. High attendance (red points for Attendance) consistently increases predictions because of the positive SHAP values, while low attendance (blue points) decreases them because of the negative SHAP values. Similarly, high study hours (red for Hours Studied) have a strong positive effect, whereas low study hours reduce scores. Less important features, such as **Internet\_Access\_Yes** and **Learning\_Disabilities\_Yes**, show minimal SHAP values, meaning they have little effect on predictions.

**Interaction effect**

There is one thing we noticed that students with higher scores (where Hours Studied has a strong positive SHAP value) tend to cluster more tightly, meaning the impact of Hours Studied is more consistent. Students with **lower scores** show more variability, as reflected in a wider spread of SHAP values for low hours Studied.

To better understand how these features interact with each other and influence individual predictions, we will now focus on their **dependence plots**. This plot will reveal the specific relationships between **Attendance and** **Hours Studied.**

A graph showing the number of hours

Description automatically generated

Results

As Hours Studied increases (moving right on the x-axis), the SHAP value for Hours Studied consistently increases, indicating a strong positive contribution to exam scores. At lower study hours (left side), students with high Attendance (red points) experience a less negative impact, while students with low Attendance (blue points) face a larger negative effect. This interaction suggests that high attendance mitigates the negative effects of low study hours, while low attendance amplifies them. Conversely, as study hours increase, the positive contribution becomes more consistent across all attendance levels, with most points clustering at high SHAP values.

**Overall Conclusion**

The findings indicate that student exam performance is significantly influenced by several key factors. Both the Multiple Linear Regression (MLR) model and the Random Forest model consistently identified attendance, hours studied, and prior academic performance as the most influential predictors. Analyzing feature contributions through methods like SHAP confirmed that changes in these key variables substantially influence individual predictions.

While the Random Forest provided insight into non-linear relationships, the reduced MLR model produced a high R² value and a low mean squared error and making it the most effective model for predicting academic performance.

These modeling results suggest that improving attendance policies, encouraging effective study habits, and offering targeted tutoring or supplemental activities can have a meaningful impact on student performance. In practice, educators and parents could prioritize these areas to boost exam scores.

**Reflection**

In the future, to further improve the predictive accuracy and efficiency of modeling analysis, we can verify the key assumptions such as linearity, normality, and constant variance, and apply transformations of predictors if necessary, address multicollinearity concerns using tools like Variance Inflation Factor (VIF) to manage highly correlated predictors, identify and handle outliers or influential data points through residual analysis and influence metrics, and also incorporate more relevant features and simplifying the dataset using dimensionality reduction techniques like Principal Component Analysis to improve predictive accuracy and efficiency.