STAT 530 - Applied Regression

Project Report

PREDICTION ANALYSIS OF STUDENT PERFORMANCE

Team Members:

- Sai Sanjith Sivapuram
- Lohit Arun Saravanan
- Pranitha Velusamy Sundararaj

INTRODUCTION

Overview of Student Performance Prediction:

Student performance prediction is vital in enhancing academic outcomes by enabling institutions to provide tailored support and interventions. This project aims to build predictive models that utilize various student-related factors to forecast academic performance. The insights from these models can aid curriculum improvements, better resource allocation, and enhance institutional rankings.

Importance of Performance Prediction:

- Facilitates personalized support for students.
- Helps in designing effective course structures.
- Provides data-driven insights for resource allocation.
- Enhances institutional reputation and rankings.
- Promotes better data recording and utilization.

Objective of the Analysis:

The primary objective is to build robust regression models to predict student performance based on multiple variables. The project evaluates model adequacy, compares predictive capabilities, and identifies the best-fit model for academic use.

DATASET OVERVIEW AND PRE-PROCESSING

Dataset Details:

- **Source:** Kaggle
- **Dataset URL:** [Student Performance (Multiple Linear Regression)]
- **Size:** 10,000 rows and 6 columns
- **Variables:**
    - Hours Studied
    - Previous Scores
    - Extracurricular Activities
    - Sleep Hours
    - Sample Question Papers Practiced
    - Performance Index (Target Variable)

Pre-Processing Steps:

- Checked for null values; the dataset contained no missing entries.
- Applied one-hot encoding to transform categorical variables into numeric format.

- Summarized data for statistical insights, identifying correlations and distributions.
- Split the data into training (80%) and testing (20%) sets for model development.

EXPLORATORY DATA ANALYSIS

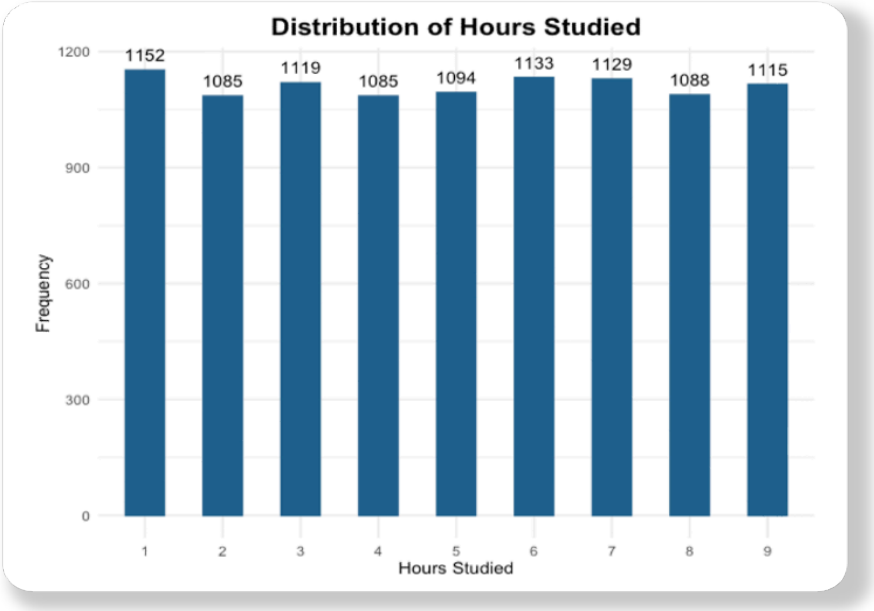
Descriptive Statistics:

- Average study hours: 5.6 hours/day
- Mean sleep hours: 7.2 hours/day
- The strongest correlation was observed between "Previous Scores" and "Performance Index."

Visualization Techniques:

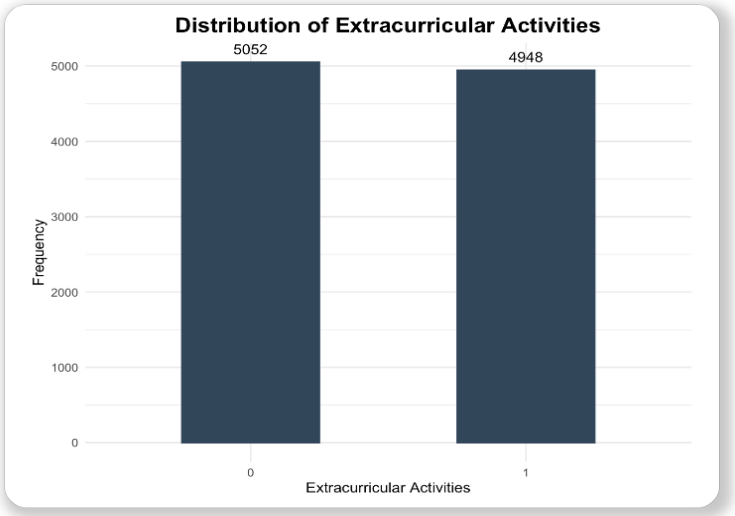
1. Bar Plot: Distribution of Hours Studied

- The distribution of hours studied is nearly uniform, with frequencies ranging between 1085 and 1152 for all categories.
- There are no extreme peaks or troughs, indicating consistent study patterns across students.



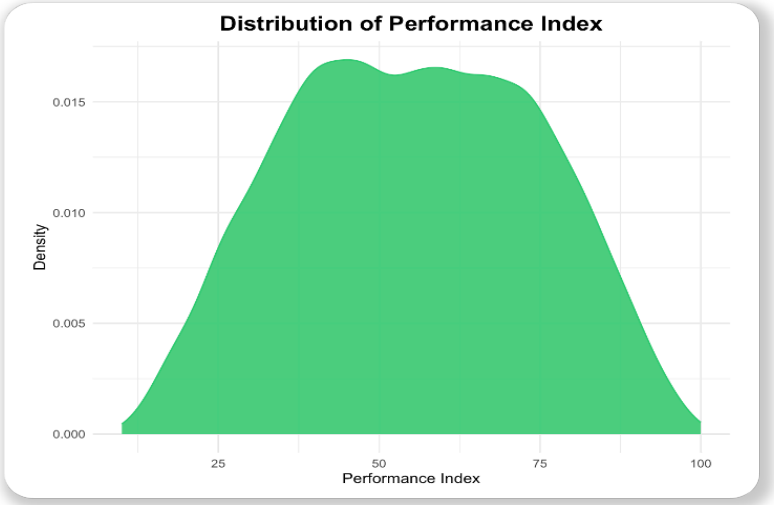
2. Bar Plot: Distribution of Extracurricular Activities

- Students who do not participate in extracurricular activities (0) slightly outnumber those who participate (1), with counts of 5052 and 4948, respectively.
- The difference is minimal, suggesting balanced participation in extracurricular activities across the dataset.



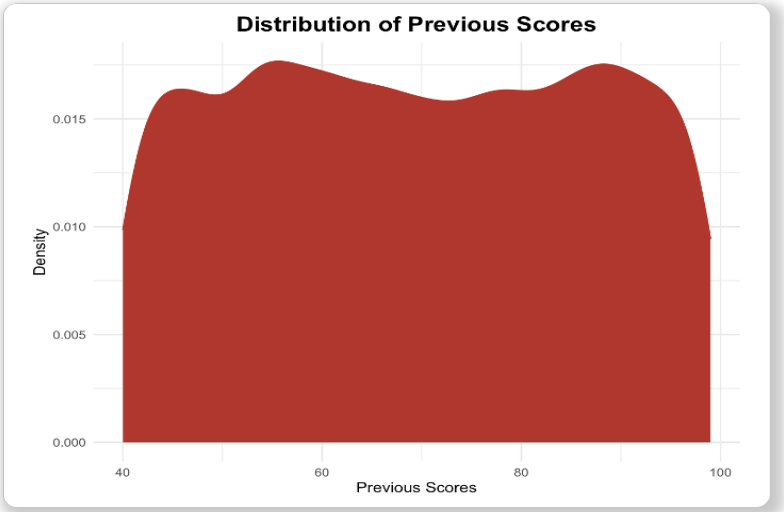
3. Density Plot: Distribution of Performance Index

- The performance index follows a symmetrical bell-shaped distribution, peaking near the center.
- No extreme outliers or heavy tails are observed, indicating the majority of students cluster around the average performance.



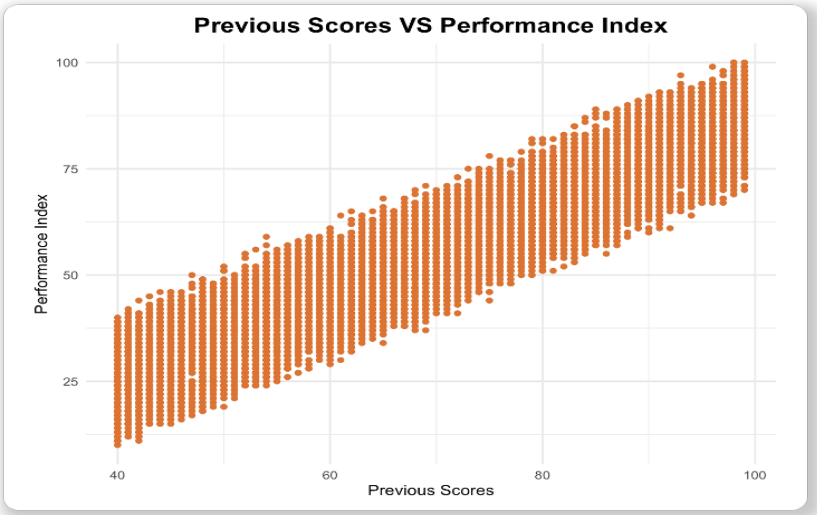
4. Density Plot: Distribution of Previous Scores

- The previous scores show a slightly flatter density curve compared to the performance index, indicating a wider spread of values.
- Students are distributed relatively evenly within the score range, with no noticeable skew or clustering.



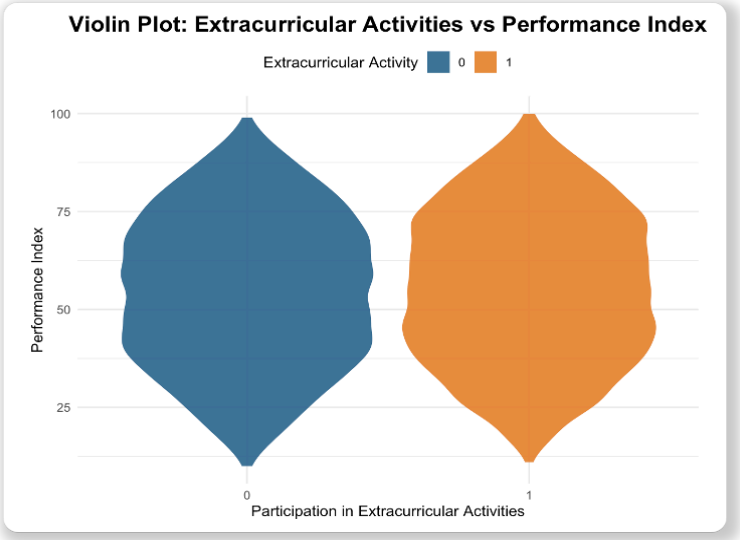
5. Scatter Plot: Previous Scores vs Performance Index

- A clear and strong positive linear trend exists, indicating that higher previous scores are directly associated with higher performance index values.
- The data points are densely clustered along the trend line, with minimal deviation, supporting a strong correlation.



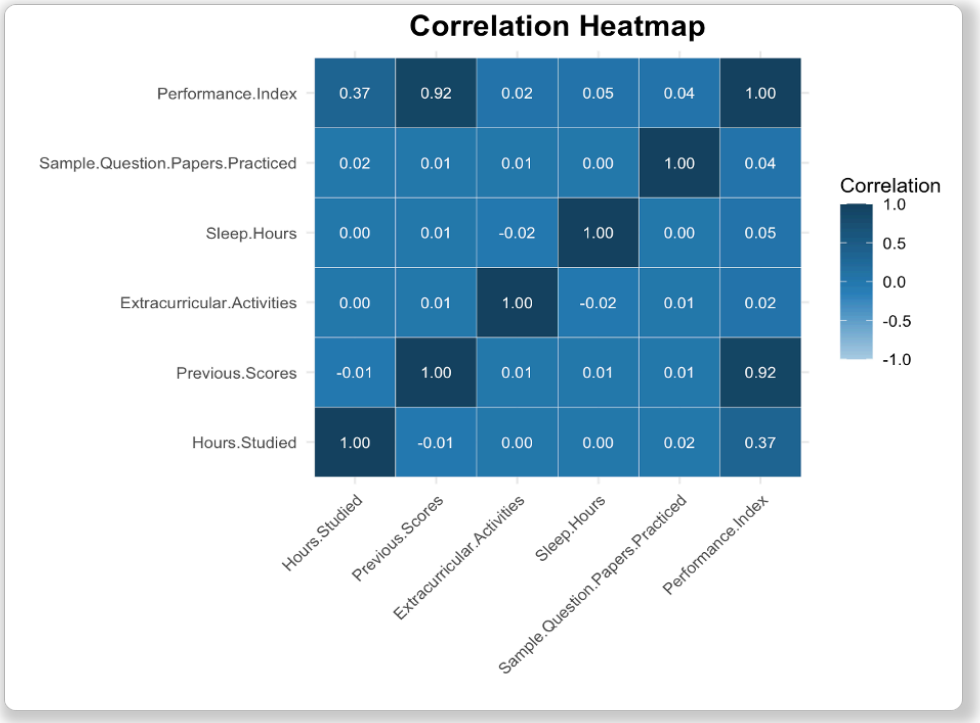
6. Violin Plot: Extracurricular Activities vs Performance Index

- The distribution of performance index is similar for students who do (1) and do not (0) participate in extracurricular activities, with overlapping density shapes.
- Slight variability is observed within each group, but the overall effect of extracurricular activities on the performance index appears negligible.



7. Correlation Heatmap

- Previous scores and performance index exhibit the strongest correlation (0.92), highlighting the significant impact of prior academic achievement on current performance.
- Hours studied show a moderate positive correlation (0.37) with performance index, while other variables, such as extracurricular activities and sleep hours, have near-zero correlations, suggesting limited influence.



MODELS IMPLEMENTED

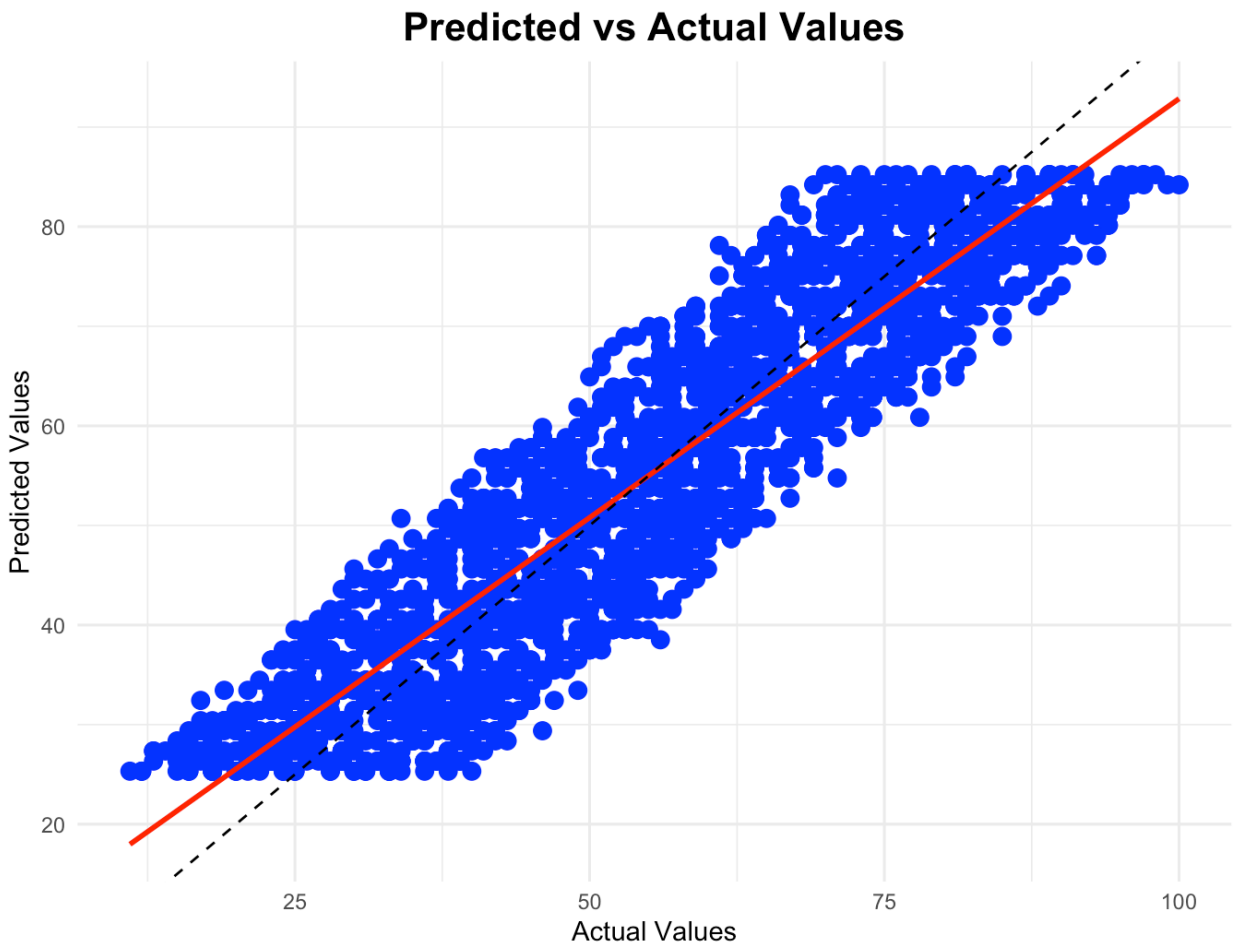
Model 1: Simple Linear Regression

- **Target Variable:** Performance Index
- **Explanatory Variable:** Previous Scores
- **Formula:** $Y = \beta 0 + \beta 1X1 + \varepsilon$
    - Y: Performance Index
    - X1: Previous Scores
    - ε: Error term
- **R-Squared Metrics:**
    - Train: 0.8382
    - Test: 0.8351

Findings:

1. A strong linear relationship between Previous Scores and the Performance Index resulted in reasonable model performance.
2. Training and testing values are closely aligned, indicating that the model does not overfit and generalizes well.
3. The model's simplicity caused it to overlook the impact of other relevant variables, limiting its ability to explain multi-factor influences.

Results from Model 1:



Model 2: Multiple Linear Regression

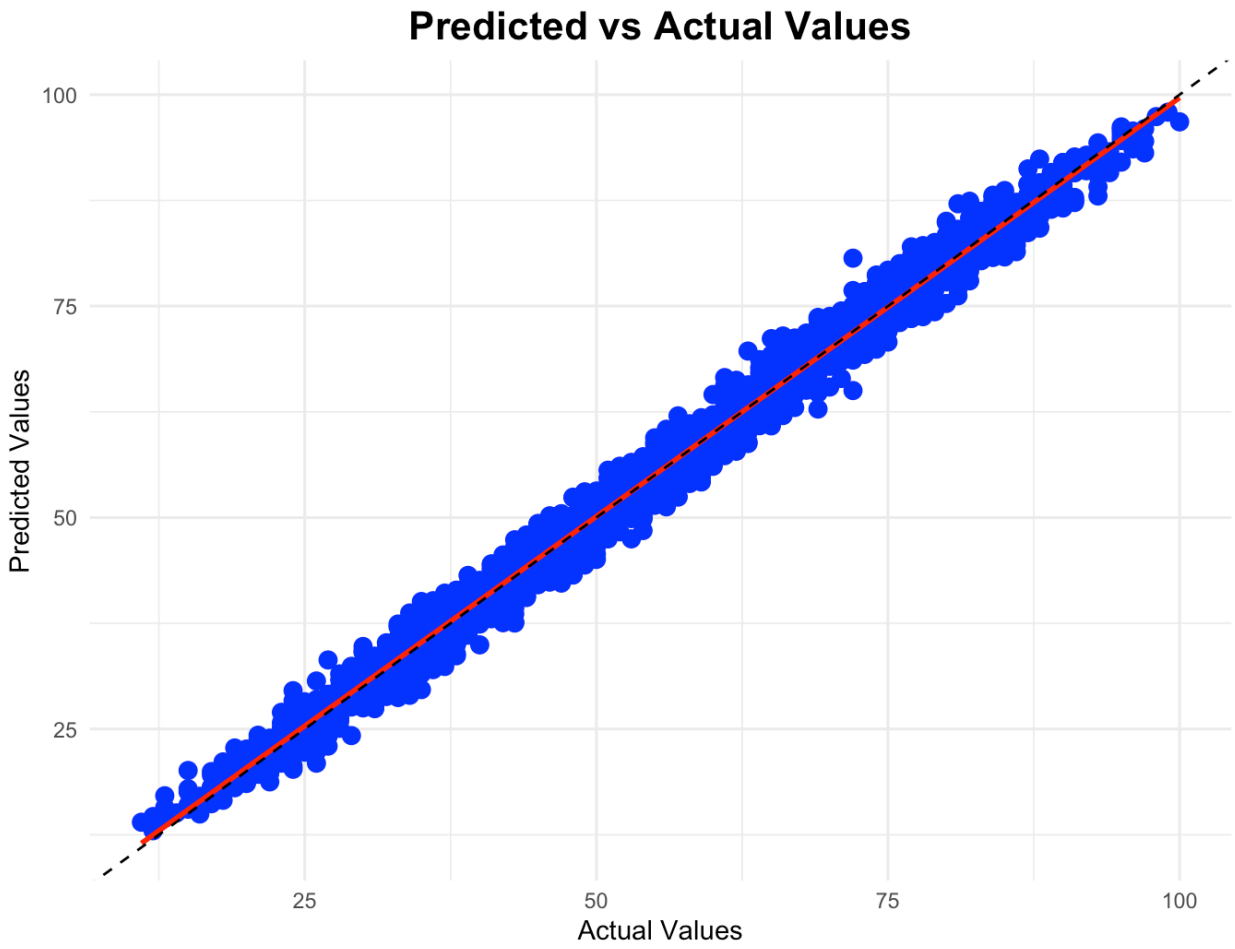
- **Explanatory Variables:** Included all six variables.
- **Formula:** $Y = \beta0 + \beta1X1 + \beta2X2 + ... + \betanXn + \varepsilon$
- **R-Squared Metrics:**
    - Train: 0.9888
    - Test: 0.9884

Findings:

1. The inclusion of all six variables allowed the model to explain nearly 99% of the variance in both training and testing datasets.
2. Variance Inflation Factor (VIF) analysis revealed no multicollinearity among predictors, validating the stability of coefficients.

3. The model achieved excellent generalization, with minimal variance between training and testing, making it ideal for this dataset.

Results from Model 2:



Model 3: Random Forest Regression

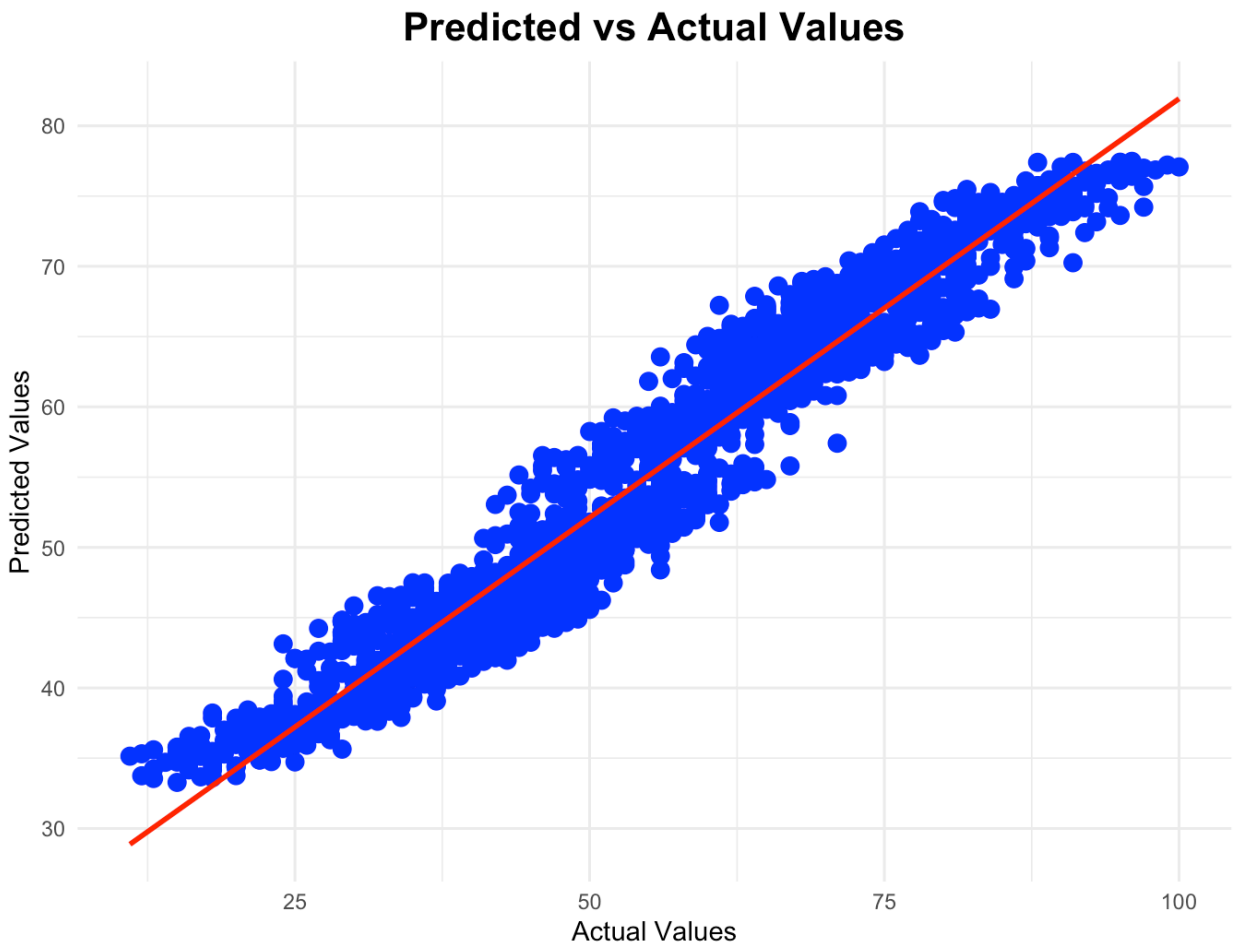
- **Approach:** Ensemble method with decision tree aggregation.
- **R-Squared Metrics:**
  - Train: 0.8230
  - Test: 0.8165

Findings:

1. The model captured non-linear relationships and interactions between predictors, which were missed by linear models.
2. Slightly lower R-squared values in training and testing indicate underperformance, likely due to overfitting on complex relationships.

3. While the model is effective for capturing complex patterns, its reduced performance suggests it is not as robust for this dataset compared to Multiple Linear Regression.

Results from Model 3:



RESULTS AND ANALYSIS

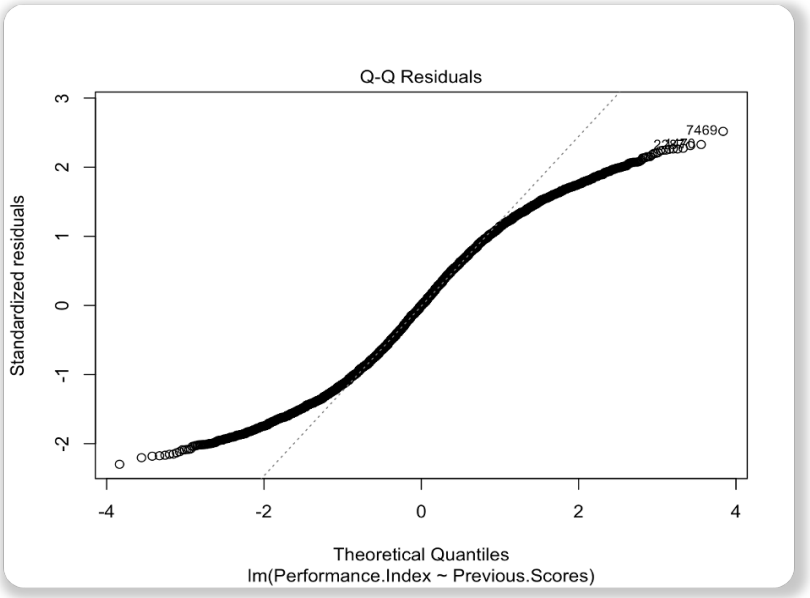
Model Performance Comparison:

Model	**Train R-squared**	**Test R-squared**
Simple Linear Regression	0.8382	0.8351
Multiple Linear Regression	0.9888	0.9884

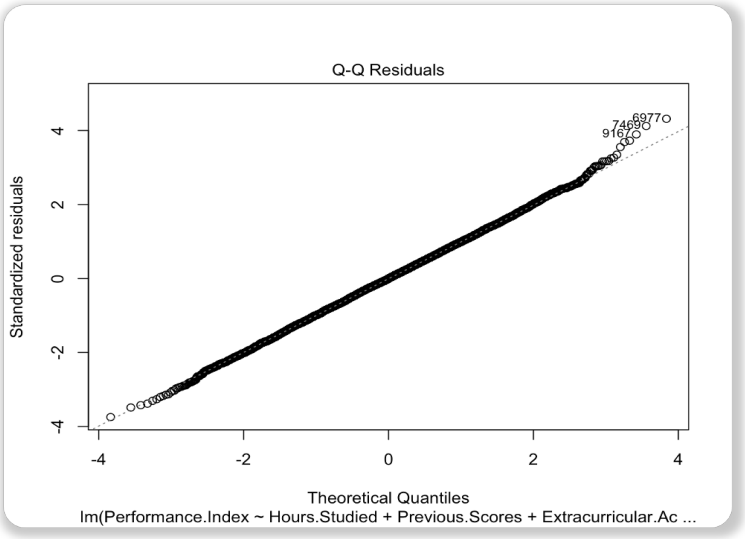
Random Forest Regression	0.8230	0.8165

Residual Analysis:

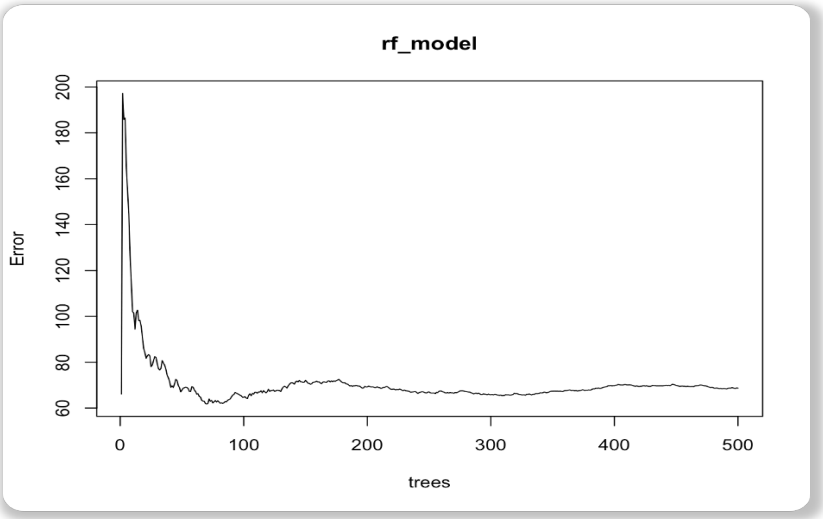
- For the **simple linear regression model**, the residual plot likely reveals a fairly random distribution of residuals, supporting the assumption of linearity and homoscedasticity (constant variance). The R-squared values for training (0.8382) and testing (0.8351) indicate a well-generalized model. However, minor patterns could suggest limitations in using only the "Previous Scores" variable.



- For the **Multiple Linear Regression model**, the residual plot shows ideal behavior, suggesting no significant patterns, ensuring assumptions like linearity, homoscedasticity, and no multicollinearity (as verified by Variance Inflation Factors). The high R-squared values for training (0.9888) and testing (0.9884) highlight strong predictive accuracy and generalizability.



- For the **Random Forest model**, the error decreases rapidly as the number of trees increases, indicating that adding more trees significantly improves the model's performance. However, after approximately **100 trees**, the error stabilizes and exhibits minor fluctuations, suggesting that the model has reached convergence. While the model is not perfect, it demonstrates fair predictive power and efficiency, making it a decent model for the given task



CONCLUSION

In conclusion, this project on student performance prediction effectively demonstrated the application of statistical and machine learning techniques to model and evaluate academic outcomes. Through robust preprocessing, including one-hot encoding for categorical variables and an 80-20 data split, we ensured a reliable dataset for training and testing. Among the models

implemented, Multiple Linear Regression achieved superior performance with R-squared metrics of 0.9888 on the training set and 0.9884 on the test set, indicating excellent generalization and minimal variance inflation. In contrast, Random Forest and Simple Linear Regression exhibited lower predictive accuracy. The analysis underscores the efficacy of incorporating all explanatory variables in a multivariate context for prediction tasks. These results highlight the potential for predictive modeling to guide targeted interventions and optimize educational strategies, paving the way for scalable implementations in academic performance monitoring systems.

Key Findings:

1. **Multiple Linear Regression** emerged as the best-performing model, with minimal variance between training and testing data.

2. **Random Forest Regression** provided reasonable accuracy but was slightly prone to overfitting.

3. **Simple Linear Regression** offered a good baseline but lacked the robustness of multiple regression models.

Applications:

- The models can aid in identifying students at risk and targeting interventions.

- Insights can inform academic policy, resource allocation, and student support systems.

Future Directions:

- Incorporate additional variables like socioeconomic factors and mental health indicators.

- Explore advanced machine learning models such as neural networks for improved accuracy.

- Enhance interpretability with SHAP (SHapley Additive exPlanations) values.