Application of the SEMMA Methodology for Predicting Student Performance: A Case Study Using the Student Performance Factors Dataset

Yash Kumar

October 24, 2024

Abstract

This study applies the SEMMA methodology to develop a predictive model for student exam performance using the Student Performance Factors dataset from Kaggle. The process encompasses data sampling, exploratory analysis, data transformation, model building, and assessment. Leveraging various regression models, this research identifies critical factors influencing exam outcomes and enhances predictive accuracy through thorough model tuning and validation. The results offer valuable insights into effective educational analytics.

1 Introduction

Predicting student performance is a crucial aspect of education analytics, providing insights for tailored interventions. This study uses the SEMMA (Sample, Explore, Modify, Model, Assess) methodology to systematically process the Student Performance Factors dataset from Kaggle, aiming to develop a robust predictive model for exam outcomes.

2 Methodology

The SEMMA methodology, employed here, involves five phases:

2.1 Sample Phase

In this phase, a representative subset of the dataset was selected to ensure computational efficiency and data representativeness. Missing values were handled, and stratified sampling was used to maintain the distribution of the target variable.

2.2 Explore Phase

The exploration phase included descriptive statistics, correlation analysis, and visualization of distributions. Insights into feature relationships were obtained through correlation heatmaps, scatter plots, and distribution plots.

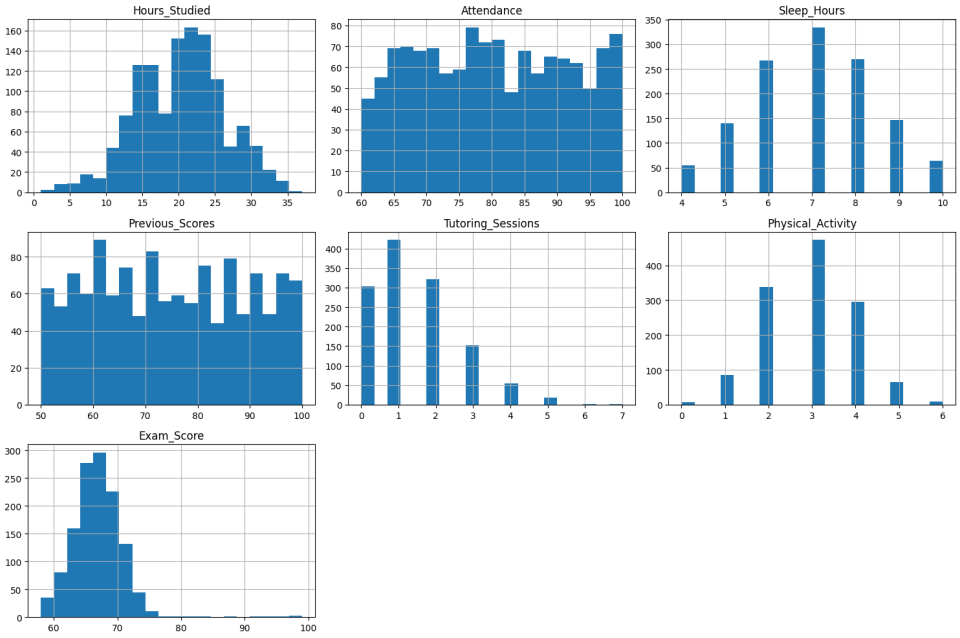

Figure 1: Histogram of the numerical columns

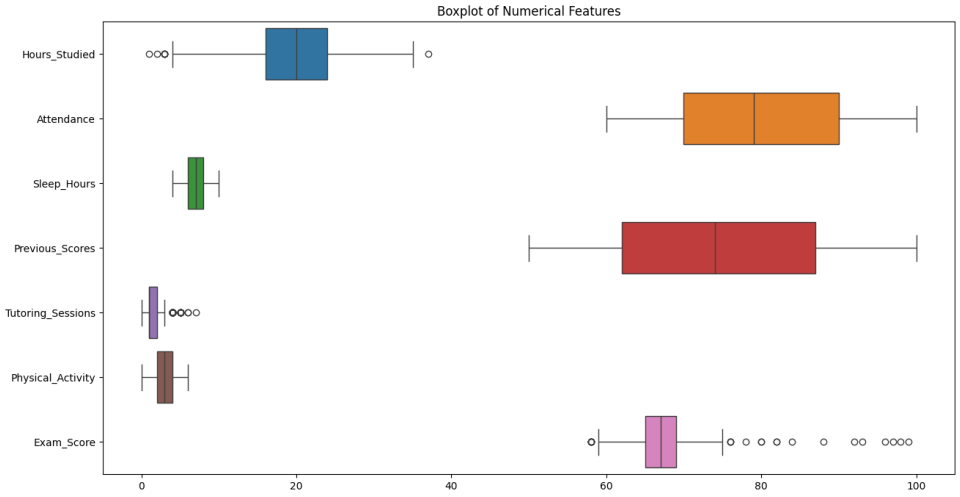

Figure 2: Boxplot of the categorical variables

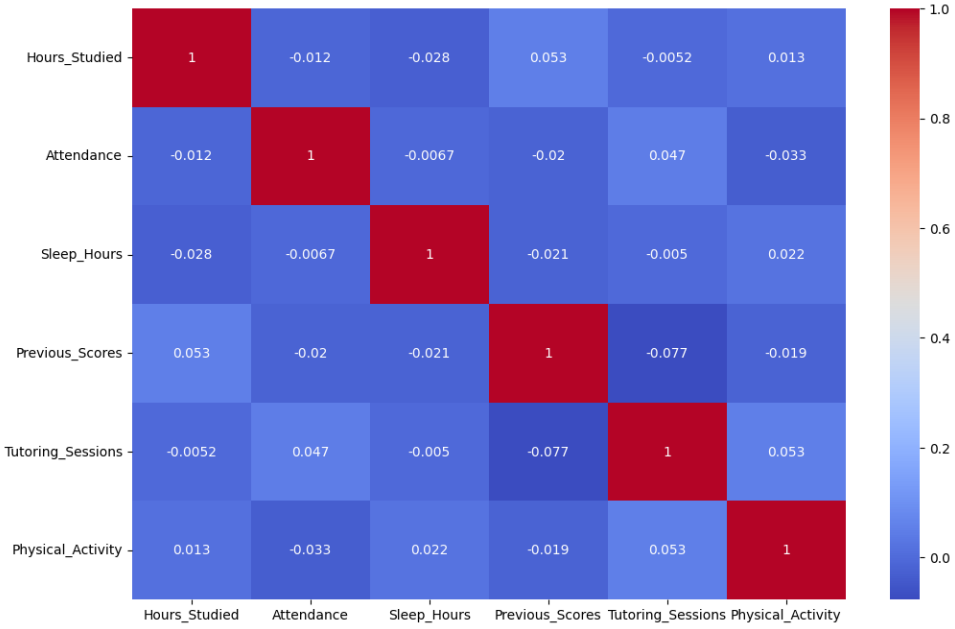

Figure 3: Heatmap of the numerical columns

2.3 Modify Phase

The modification phase involved data cleaning, feature engineering, encoding categorical variables, scaling numerical features, and feature selection. Techniques like one-hot encoding and standardization were applied to transform the dataset for modeling.

3 Model Building

Multiple regression models, including Linear Regression, Decision Trees, and Random Forests, were implemented. Hyperparameter tuning via GridSearchCV improved the model's performance, achieving optimal accuracy for predicting student exam scores.

4 Assessment

The assessment phase focused on evaluating model accuracy through metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). Residual analysis and cross-validation ensured the model's reliability and generalization capabilities.

5 Results and Discussion

The results indicate that the Random Forest model, with optimized hyperparameters, performed best in predicting exam scores, as evidenced by the highest R-squared and lowest error metrics. This study's findings highlight the importance of each SEMMA phase in constructing a reliable predictive model.

6 Conclusion

This case study demonstrates the effectiveness of the SEMMA methodology in developing a predictive model for student performance. The systematic approach ensures that each phase adds value to the data mining process, ultimately leading to accurate and reliable results.

7 Future Work

Future research could explore additional models, such as Gradient Boosting or Support Vector Machines, and incorporate more complex feature engineering techniques. The current approach could also be adapted to predict other educational outcomes, such as graduation rates or course completions.

References

- SEMMA Methodology Overview: `https://www.sas.com/en_us/insights/analytics/what-is-semma.html`
- Kaggle Student Performance Dataset: `https://www.kaggle.com/datasets/lainguyn123/student-performance-factors`