Predicting Student Exam Performance Using Regression Models

Abstract

This research investigates the factors influencing student exam performance using a dataset containing academic, personal, and socio-economic attributes. Multiple regression-based models were applied, including linear regression, polynomial regression, and various feature-engineered datasets to evaluate their predictive performance. The goal was to identify which features and preprocessing approaches yield the highest R-squared value.

1. Introduction

Student performance is influenced by a wide range of factors, from study habits and attendance to family background and access to educational resources. Accurately predicting performance can assist educators and policymakers in identifying at-risk students and improving academic support strategies. This study applies regression techniques to the Student Performance dataset to analyze these factors and compare different modeling approaches.

2. Dataset Description

The dataset contains 6,607 records with 20 features related to student demographics, study habits, and environmental factors. Attributes include hours studied, attendance percentage, parental involvement, access to resources, extracurricular activities, sleep hours, previous scores, motivation level, internet access, tutoring sessions, family income, teacher quality, school type, peer influence, physical activity, learning disabilities, parental education level, distance from home, and gender. The target variable is 'Exam_Score', representing the student's final exam result.

3. Methodology

3.1 Data Preprocessing

Data preprocessing involved handling missing values, encoding categorical variables using One-Hot Encoding, and scaling numerical features with StandardScaler. Categorical variables with ordinal meaning were also considered for alternative encoding approaches.

3.2 Model Development

Multiple modeling strategies were applied:

1. Linear Regression on the full dataset after preprocessing.
2. Feature-engineered dataset introducing combined parental impact variables.
3. Datasets with selected features removed (e.g., without parental factors, without teacher factors).
4. A dataset focusing on student diligence (study hours, attendance, etc.).
5. Polynomial Regression (degree=2, interaction only) on the full dataset.

The performance of each approach was compared using R-squared and Mean Squared Error.

4. Results

The baseline Linear Regression model on the full dataset achieved the highest R-squared value of approximately 0.77, outperforming polynomial regression and all alternative feature subsets. Polynomial Regression yielded a slightly lower R-squared, suggesting limited benefit from introducing interaction terms for this dataset. Removing certain feature groups generally led to reduced accuracy, indicating that most features contribute to predictive performance.

5. Discussion

The results highlight the importance of a wide range of academic and socio-economic features in predicting student performance. While feature engineering can sometimes improve predictive accuracy, in this case, the full dataset performed best, suggesting that all features provide valuable information. Polynomial features did not improve results, possibly due to overfitting or the linear nature of relationships between predictors and the target variable.

6. Conclusion

This study demonstrates that student performance can be predicted with reasonable accuracy using linear regression models applied to a rich set of demographic, academic, and socio-economic features. The analysis found that the inclusion of all available features produced the highest accuracy, and alternative modeling approaches did not outperform the baseline model. Future work may explore advanced non-linear models such as Random Forests or Gradient Boosting to assess potential gains in predictive performance.