Predicting Student Performance Report

1. Introduction

Student academic performance is influenced by various factors, ranging from cognitive abilities to socio-economic and behavioral characteristics. Social Factors Affecting Academic Performance: Further Evidence a piece written by Thomas R. Ford in 1957 highlighted that academic performance should be understood within the broader context of social influences, particularly the impact of parental interest and the societal definition of gender roles.
The findings highlight the need for educators to consider these factors when developing strategies to motivate students and improve academic outcomes. Overall, the study suggests that addressing the nuanced relationships between social status, parental involvement, and student attitudes can lead to more effective educational interventions. The social influences Thomas R. Ford looked at were Demographics, Parental Involvement, Attitudes Toward School, Social Influences, and Social Class

Identifying and understanding these factors can help educators and policymakers implement targeted interventions to enhance learning outcomes. In this project, I aim to explore the use of machine learning regression models to predict student exam scores based on a variety of factors, such as hours studied, attendance, sleep hours, and previous academic performance. My goal is to uncover the relationships between these factors and predict exam scores accurately.

2. Objective

The primary objective of this project is to build and evaluate different regression models to predict student exam scores. Identify the most influential features that contribute to student performance. Compare the performance of various regression models based on evaluation metrics such as r^2(r-squared score), rmse (root mean squared error), mae (mean absolute error), and mse (mean squared error).

3. Dataset Overview

The dataset used in this project includes multiple predictors related to student behavior and academic history:

Hours Studied: Number of hours spent studying per week.
Attendance: Percentage of classes attended.
Parental Involvement: Level of parental involvement in the student's education (Low, Medium, High).
Access to Resources: Availability of educational resources (Low, Medium, High).
Extracurricular Activities: Participation in extracurricular activities (Yes, No).
Sleep Hours: Average number of hours of sleep per night.
Previous Scores: Scores from previous exams.
Motivation Level: Student's level of motivation (Low, Medium, High).
Internet Access: Availability of internet access (Yes, No).
Tutoring Sessions: Number of tutoring sessions attended per month.
Family Income: Family income level (Low, Medium, High).
Teacher Quality: Quality of the teachers (Low, Medium, High).
School Type: Type of school attended (Public, Private).
Peer Influence: Influence of peers on academic performance (Positive, Neutral, Negative).
Physical Activity: Average number of hours of physical activity per week.
Learning Disabilities: Presence of learning disabilities (Yes, No).
Parental Education Level: Highest education level of parents (High School, College, Postgraduate).
Distance from Home: Distance from home to school: (Near, Moderate, Far).
Gender: Gender of the student (Male, Female).
Exam Score: Final exam score.

4. Methodology

4.1 Data Preprocessing

The data underwent several preprocessing steps to prepare it for model building. Missing values within the dataset were appropriately handled by being exempted. The missing values only contributed to a small percentage of the entire dataset.

Categorical variables were encoded and Feature Scaling was implemented. However, we saw that implementing the scaling slightly reduced the performance of the models

4.2 Exploratory Data Analysis (EDA)

Before modeling, exploratory data analysis (EDA) was conducted to understand the distribution of features and their relationships with the target variable. Key visualizations and correlations were examined.

The frequency distribution plot assumes to take a bell shape which implies the population is normally distributed and therefore is the kind of trend you would like to see with regards to the exam score. However, the trend may also show that there are outliers when it comes to students who scored high in the exam from grades 80 to 100
The box plot clearly showed the number of outliers in the data set and how far apart they were from the higher grades starting from the 75 mark to 100

4.3 Feature Selection

To reduce model complexity and improve interpretability, a feature selection process was implemented based on statistical significance. Pearson's correlation coefficients and p-values were computed for each feature against the target variable. Features with a p-value less than 0.05 were considered significant and included in the final model.

4.4 Model Building

Four regression models were trained and evaluated:

Linear Regression: A baseline model to understand the linear relationships between features and the target variable.
Lasso Regression: A linear regression model with L1 regularization to penalize large coefficients and perform feature selection.
Ridge Regression: A linear regression model with L2 regularization to handle multicollinearity and prevent overfitting.
Random Forest Regressor: An ensemble model that builds multiple decision trees and averages their results for better predictive performance.

4.5 Model Evaluation

Each model was evaluated using the following performance metrics:

R^2 (R-squared): Measures the proportion of variance in the target variable explained by the model. A higher R^2 indicates a better fit.
RMSE (Root Mean Squared Error): Indicates the average magnitude of the prediction errors in the same units as the target variable.
MAE (Mean Absolute Error): The average of the absolute differences between predicted and actual values.
MSE (Mean Squared Error): The average of the squared differences between predicted and actual values, which is sensitive to large errors.

5. Results and Discussion

5.1 Model Performance Comparison

The performance of each model was compared based on the evaluation metrics this was after the model had been scaled and feature selection was applied. The results are summarized in the table below:

model name	r-squared	rmse	mae	mse
Linear Regression Model	**0.730480**	**2.046497**	**0.544610**	**4.188151**
Lasso Regression Model	0.724714	2.068276	0.613751	4.277765
Ridge Regression Model	0.730478	2.046505	0.544670	4.188184
Random Forest Model	0.622423	2.422250	1.203174	5.867293

From the results, it can be seen that the Linear Regression Model achieved the highest R^2 score (0.730480), indicating it explained the most variance in the exam scores. It also had the lowest RMSE (), MAE, and MSE, suggesting that it made the most accurate predictions compared to the other models.

5.2 Insights from Feature Selection

The analysis showed that features such as Sleep Hours, Gender, and Peer Influence had a lesser impact on exam scores, as their p-values were higher than 0.05, suggesting that their relationship with the exam score was not statistically significant. These findings can help guide further feature engineering or data collection.

6. Conclusion

This project demonstrates the application of machine learning techniques to predict student exam scores based on behavioral and social factors. The Linear Regression Model outperformed other models, offering the best predictive accuracy. These insights can help educators prioritize interventions for students and make predictions based on these key factors.

Future work could include exploring additional features or using more advanced machine learning techniques to further improve the prediction accuracy.

7. Recommendations

Based on the results of this analysis, I recommend that:

Increased Study Time: Encouraging students to spend more time studying could significantly improve their exam performance.
Promoting Regular Attendance: Ensuring that students attend classes regularly may contribute to better learning outcomes.
Utilizing Previous Scores as a predictor can help identify students who may need additional support.

8. Future Work

1. **Develop Machine Learning Application:** Turning the machine learning model into an interactive web application that allows parents to engage with the model to make predictions about their child's performance based on various social and personal factors.
2. **Incorporate More Features:** Additional data on socio-economic status, parental involvement, and mental health could provide further insights into student performance.
3. **Model Optimization**: Experiment with hyperparameter tuning like GridSearch and other ensemble methods that could enhance the performance of predictive models.
4. **Longitudinal Analysis:** A longitudinal study tracking students over multiple terms could provide deeper insights into how various factors affect performance over time.