**STUDENT EXAM SCORE PREDICTION USING LINEAR REGRESSION**

**A Data Science Project on Predicting Exam Performance from Academic and Socio-Economic Factors**

**By**

**Ahmed Talal**

**September 2025**

GitHub Repository: github.com/ahmaadtalal/StudentScorePredictor

**1. Introduction**

The goal of this project was to analyze factors influencing student exam performance and build a predictive model using **Linear Regression**. The dataset was sourced from Kaggle and contains a mix of academic, demographic, and socio-economic features.

The analysis involved:

1. Data preprocessing and encoding categorical features.
2. Handling missing values.
3. Building baseline linear regression models.
4. Performing cross-validation to evaluate model stability.
5. Analyzing feature correlations.
6. Experimenting with different feature subsets to test predictive power.

**2. Dataset Description**

The dataset contained student-related features, including academic, family, and lifestyle attributes, with the target variable:

* **Target Variable:** Exam\_Score

**Key Features**

* **Academic**: Attendance, Hours\_Studied, Previous\_Scores, Tutoring\_Sessions.
* **Socio-Economic**: Family\_Income, Parental\_Education\_Level, Access\_to\_Resources, Teacher\_Quality.
* **Behavioral/Personal**: Motivation\_Level, Peer\_Influence, Sleep\_Hours, Physical\_Activity, Distance\_from\_Home.
* **Binary Features**: Extracurricular\_Activities, Internet\_Access, School\_Type, Learning\_Disabilities, Gender.

A screenshot of a computer

AI-generated content may be incorrect.

**3. Data Preprocessing**

**3.1 Ordinal Encoding**

Certain categorical features had an **inherent order** and were encoded as ordinal values:

* Parental\_Involvement / Access\_to\_Resources / Motivation\_Level / Family\_Income: 0 = Low, 1 = Medium, 2 = High
* Teacher\_Quality: 0 = Poor, 1 = Average, 2 = Excellent (78 missing values filled with mode)
* Peer\_Influence: 0 = Negative, 1 = Neutral, 2 = Positive
* Parental\_Education\_Level: 0 = High School, 1 = College, 2 = Post Graduate (90 missing values filled with mode)
* Distance\_from\_Home: 0 = Near, 1 = Moderate, 2 = Far (67 missing values filled with mode)

**3.2 Binary Encoding**

Binary features were encoded as **0 and 1**:

* Extracurricular\_Activities / Internet\_Access / Learning\_Disabilities: 0 = No, 1 = Yes
* School\_Type: 0 = Public, 1 = Private
* Gender: 0 = Male, 1 = Female

A screenshot of a computer

AI-generated content may be incorrect.

**4. Model Building**

**4.1 Baseline Linear Regression**

A baseline linear regression model was trained using all available features to predict student exam scores.

* Mean Squared Error (MSE): 3.2379
* R² Score: 0.7709

**Interpretation:**

* The MSE indicates that, on average, the squared difference between the predicted and actual exam scores is low, meaning predictions are fairly close to actual scores.
* The R² score of 0.7709 means the model can explain approximately 77% of the variance in exam scores, suggesting that the combination of academic, socio-economic, and personal factors provides strong predictive power.
* This baseline serves as a reference for evaluating future experiments with feature subsets or alternative models.

**4.2 Cross-Validation**

To evaluate the model’s robustness and generalization, we applied 5-Fold Cross-Validation, which splits the dataset into five subsets, trains the model on four folds, and tests it on the remaining fold, repeating this process for all folds.

* R² Scores (per fold): [0.6257, 0.7236, 0.7094, 0.7336, 0.8404]
* Mean R²: 0.7265
* Standard Deviation (R²): 0.0685
* Mean MSE: 4.1619

**Interpretation:**

* The mean R² of 0.7265 indicates that the model consistently explains about 73% of the variance in exam scores across different subsets.
* Slight variation in R² across folds reflects natural differences in the data splits but does not indicate instability.
* The mean MSE shows that the model’s average prediction error remains low, confirming that it is reliable and generalizes well beyond a single train-test split.

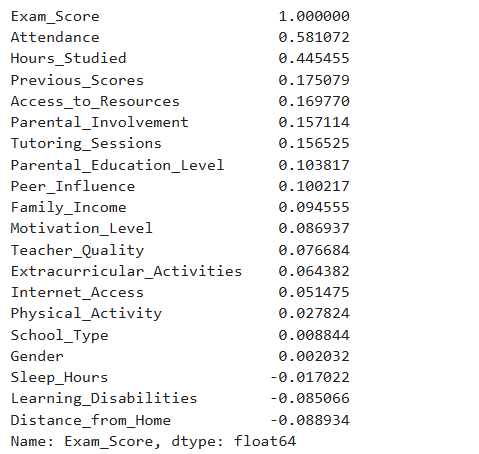
**5. Correlation Analysis**

We computed correlations between each feature and Exam\_Score.

* Top predictors: Attendance (0.581), Hours\_Studied (0.445)
* Moderate predictors: Previous\_Scores, Access\_to\_Resources, Parental\_Involvement, Tutoring\_Sessions
* Weak/negligible: Gender, School\_Type, Sleep\_Hours, Distance\_from\_Home

**Insight:**

* Academic engagement (attendance and study hours) strongly influences exam performance, while some demographic or lifestyle factors have little impact.

****

**6. Feature Subset Experiments**

**6.1 Top 5 Correlated Features**

To investigate the predictive power of the strongest individual predictors, we trained a linear regression model using only the top 5 features most correlated with Exam\_Score:

* Features used: Attendance, Hours\_Studied, Previous\_Scores, Access\_to\_Resources, Parental\_Involvement
* Cross-Validation Results:
  + Mean R²: 0.6352
  + Mean MSE: 5.5404

**Interpretation:**

* Using only the top 5 correlated features reduced performance compared to the full feature set (baseline R² ≈ 0.77).
* This shows that although these features are individually important, additional factors (like motivation, parental education, and socio-economic indicators) provide significant predictive value.
* The drop in R² highlights the multidimensional nature of exam performance, where combining academic, social, and personal factors gives the best results.

**6.2 Academic-Only Features**

We tested a model using only academic-related features to see how well they alone could predict exam scores:

* Features used: Attendance, Hours\_Studied, Previous\_Scores, Tutoring\_Sessions, Parental\_Education\_Level
* Cross-Validation Results:
  + Mean R²: 0.5973
  + Mean MSE: 6.1111

**Interpretation:**

* Performance dropped compared to both the baseline and the top 5 correlated features model.
* Purely academic predictors were not sufficient to fully explain exam scores.
* This suggests that non-academic factors such as motivation, parental involvement, and access to resources play a significant role in student performance.
* The model confirms that exam success is influenced by a combination of academic and social/environmental factors rather than academics alone.

**6.3 Pruned Feature Set**

To improve model simplicity while maintaining predictive power, we removed weakly correlated features that contributed little to predicting exam scores:

* Removed features: Gender, School\_Type, Sleep\_Hours, Physical\_Activity, Learning\_Disabilities, Distance\_from\_Home
* Cross-Validation Results:
  + R² Scores (per fold): [0.6169, 0.7141, 0.7003, 0.7169, 0.8199]
  + Mean R²: 0.7136
  + Mean MSE: 4.3557

**Interpretation:**

* The pruned model achieves performance almost as strong as the full-feature model (baseline R² ≈ 0.77) despite using fewer features.
* Removing weak predictors reduces noise and makes the model simpler, more interpretable, and computationally efficient.
* This demonstrates that a carefully selected subset of meaningful features can retain predictive accuracy while streamlining the model.

**7. Comparative Results**

We compared the performance of different feature sets to evaluate their predictive power:

| **Feature Set** | **Mean R²** | **Mean MSE** |
| --- | --- | --- |
| Full Model | 0.727 | 4.16 |
| Pruned Model | 0.714 | 4.36 |
| Top 5 Correlated | 0.635 | 5.54 |
| Academic Only | 0.597 | 6.11 |

**Interpretation:**

* The **Full Model** achieved the highest R² and lowest MSE, showing that using all features captures the most variance.
* The **Pruned Model** performs nearly as well but with fewer features, making it more interpretable and efficient.
* Models using only **Top 5 Correlated** or **Academic-Only** features showed reduced performance, confirming that **non-academic and social factors** are important for predicting exam scores.
* Overall, this comparison highlights the **trade-off between model complexity and performance**, with the pruned model providing an optimal balance for practical use.

**8. Model Export**

After evaluating all feature sets, we saved and exported the linear regression model trained on all features for future use. This allows predictions on new student data without retraining the model. The exported model file is:

* File: StudentScorePredictor\_Model.pkl

This ensures that the full-feature model can be reused directly, maintaining the same performance and feature encoding as during training.

**9. Conclusions**

1. **Best Performance:**
   * The full-feature model achieved the highest predictive performance (R² ≈ 0.727), demonstrating the value of incorporating all available academic, socio-economic, and behavioral factors.
   * The pruned model achieved nearly the same performance (R² ≈ 0.714) with fewer features, offering a simpler and more interpretable solution.
2. **Key Drivers:**
   * Attendance and Hours Studied were the most influential predictors of exam scores, highlighting the critical role of consistent academic engagement.
3. **Non-Academic Impact:**
   * Socio-economic and behavioral factors, such as parental involvement, peer influence, access to resources, and teacher quality, also contributed meaningfully, confirming that exam performance is shaped by multiple dimensions beyond academics alone.
4. **Simplification and Interpretability:**
   * Pruning weakly correlated features (e.g., Gender, School\_Type, Sleep\_Hours, Physical\_Activity, Learning\_Disabilities, Distance\_from\_Home) reduced complexity while maintaining strong predictive power.
   * This demonstrates that feature selection can improve model interpretability and efficiency without sacrificing accuracy.

**Overall Insight:**

* A balanced approach using a pruned, meaningful feature set provides an effective and interpretable model for predicting student exam performance, suitable for both analysis and portfolio demonstration.

**Development & Deployment Process**

We created the **Student Score Predictor** as a full-stack web application. The frontend was built with React, giving students an easy-to-use interface with sliders and dropdowns to input their study habits and lifestyle factors. The backend was developed with Flask, where we connected a trained machine learning model that processes the inputs and predicts exam scores.

For deployment, we hosted the frontend on **Netlify**, making the app accessible with a clean web link, and deployed the backend API on **Heroku** so the prediction system could run online. Together, this setup allowed us to build, connect, and launch a fully functional prediction platform that works seamlessly on the web.