STUDENT PERFORMANCE PREDICTION USING LINEAR REGRESSION

1. Introduction

Artificial Intelligence plays a major role in educational analytics, helping institutions analyze student patterns and predict academic outcomes. In this mini project, a simple yet powerful machine learning algorithm—**Linear Regression**—is used to predict a student's average academic score based on two key behavioral factors:

- **Daily Study Hours**
- **Attendance Percentage**

The objective of the project is to demonstrate how supervised learning models can identify trends, learn from historical data, and make accurate predictions.

2. Problem Statement

Modern learning environments often struggle to track which students are at risk of low performance.
The goal of this project is to **build a predictive model that estimates a student's average score** using linear regression, based on their study habits and attendance.

3. Objectives

1. To preprocess and analyze the student dataset.
2. To train a Linear Regression model for score prediction.
3. To evaluate the model using RMSE and R^2.
4. To interpret model coefficients and understand feature influence.
5. To visualize relationships between study hours, attendance, and performance.

4. Dataset Description

The dataset contains **40 student records** with the following columns:

- MathScore
- ScienceScore

- ReadingScore
- StudyHours
- Attendance

A new feature **AverageScore** is computed:

$$AverageScore = \frac{Math + Science + Reading}{3}$$

The model uses:

- **Features (X):** StudyHours, Attendance
- **Target (y):** AverageScore

5. Methodology

Step 1 — Data Preprocessing

- Loaded dataset using pandas.
- Calculated AverageScore.
- Selected required features.
- Verified for missing values.
- Split into Training (80%) and Testing (20%).

Step 2 — Model Training

Used **LinearRegression** from scikit-learn.

The training produced the following parameters:

- **Coefficients:** `[3.5416, 0.6283]`
- **Intercept:** `6.3131`

Step 3 — Model Equation

$$\hat{y} = 6.31 + (3.54 \times StudyHours) + (0.63 \times Attendance)$$

Step 4 — Model Evaluation

- **RMSE:** 2.1541
- **R² Score:** 0.9381

This means **93.8% accuracy** in explaining performance variation.

6. Results & Interpretation

Feature Impact

- **StudyHours:** Strongest influence. Increasing study time by 1 hour increases predicted score by ~3.54 points.
- **Attendance:** For every 1% increase in attendance, score increases by ~0.63 points.

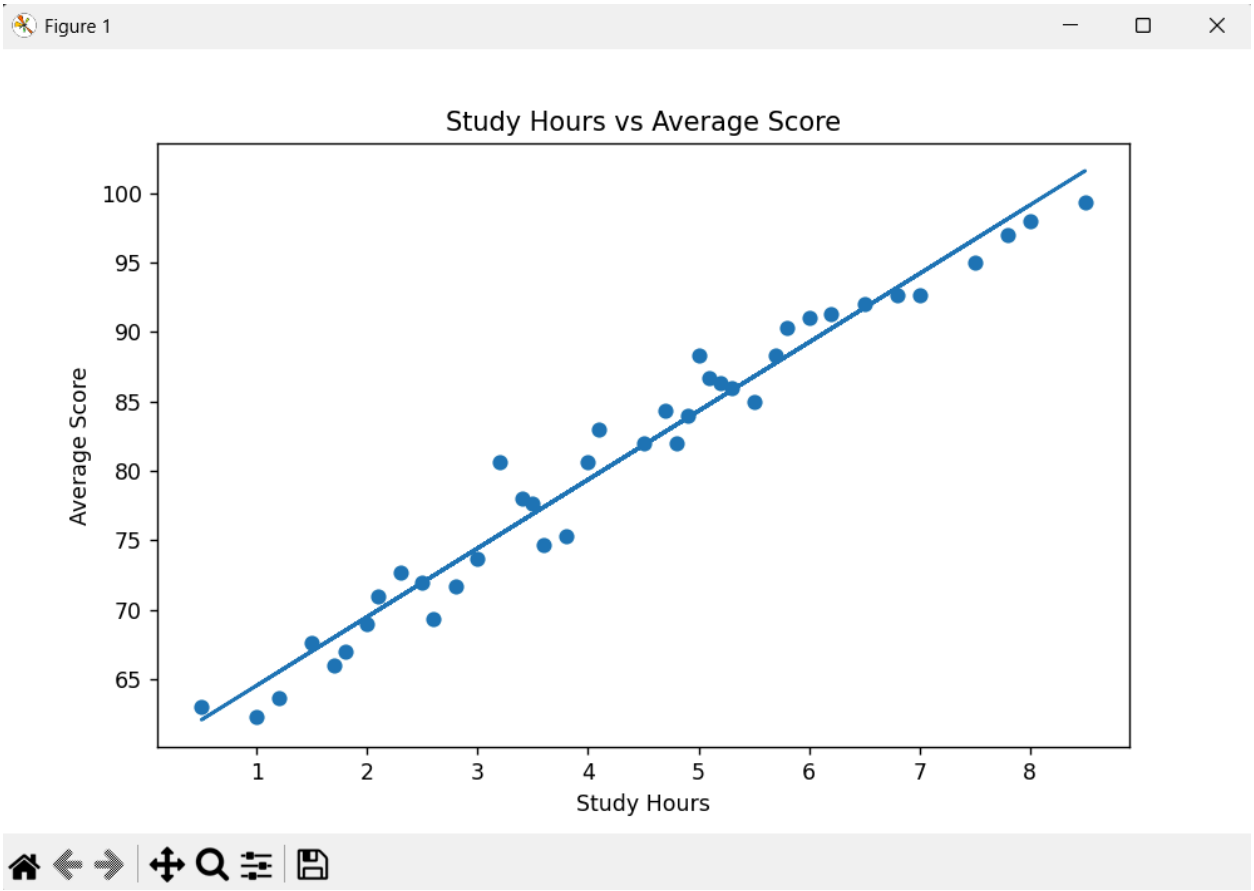
Insights

- Consistency + discipline → better scores.
- High study hours and high attendance result in 90+ scores.
- Model predicts with very low error, suitable for academic environments.

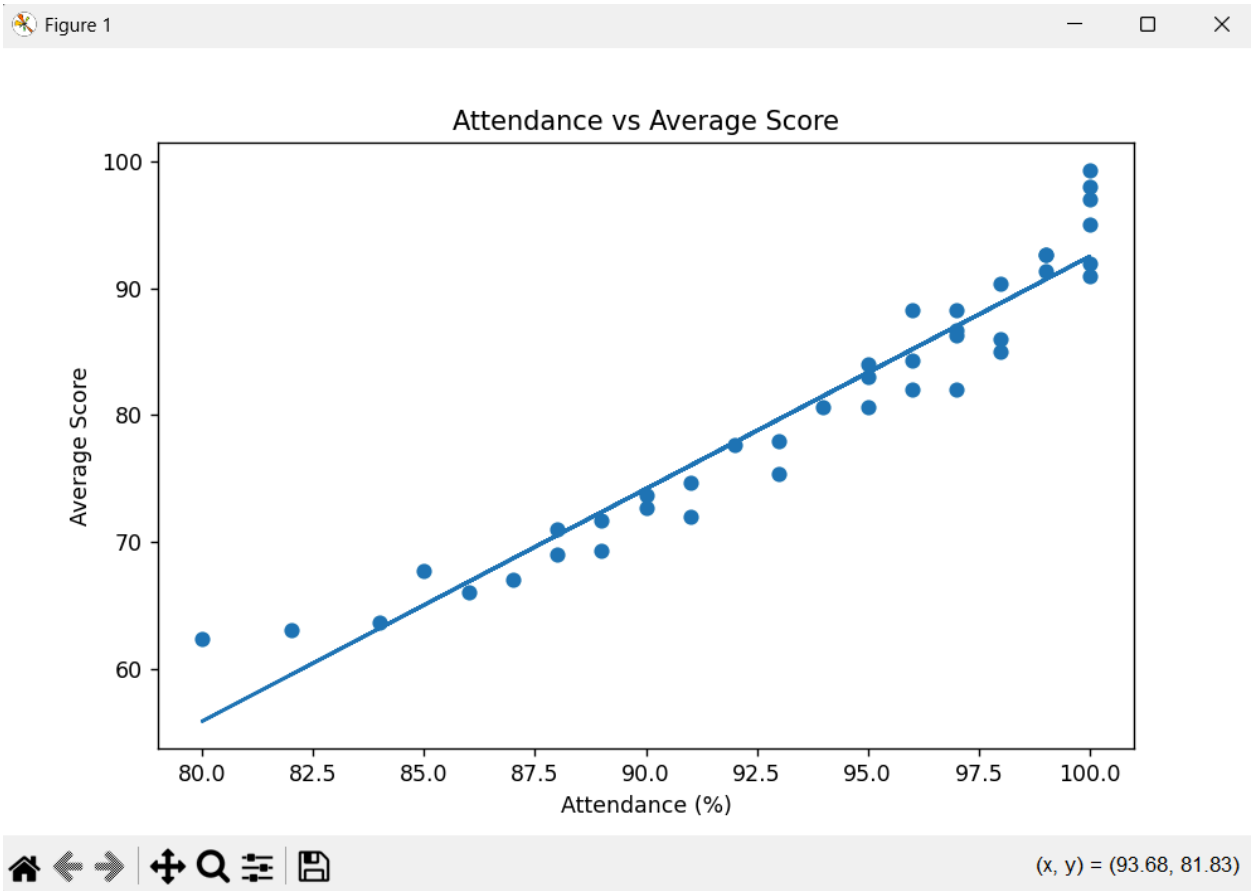
7. Visualizations

Recommended graphs:

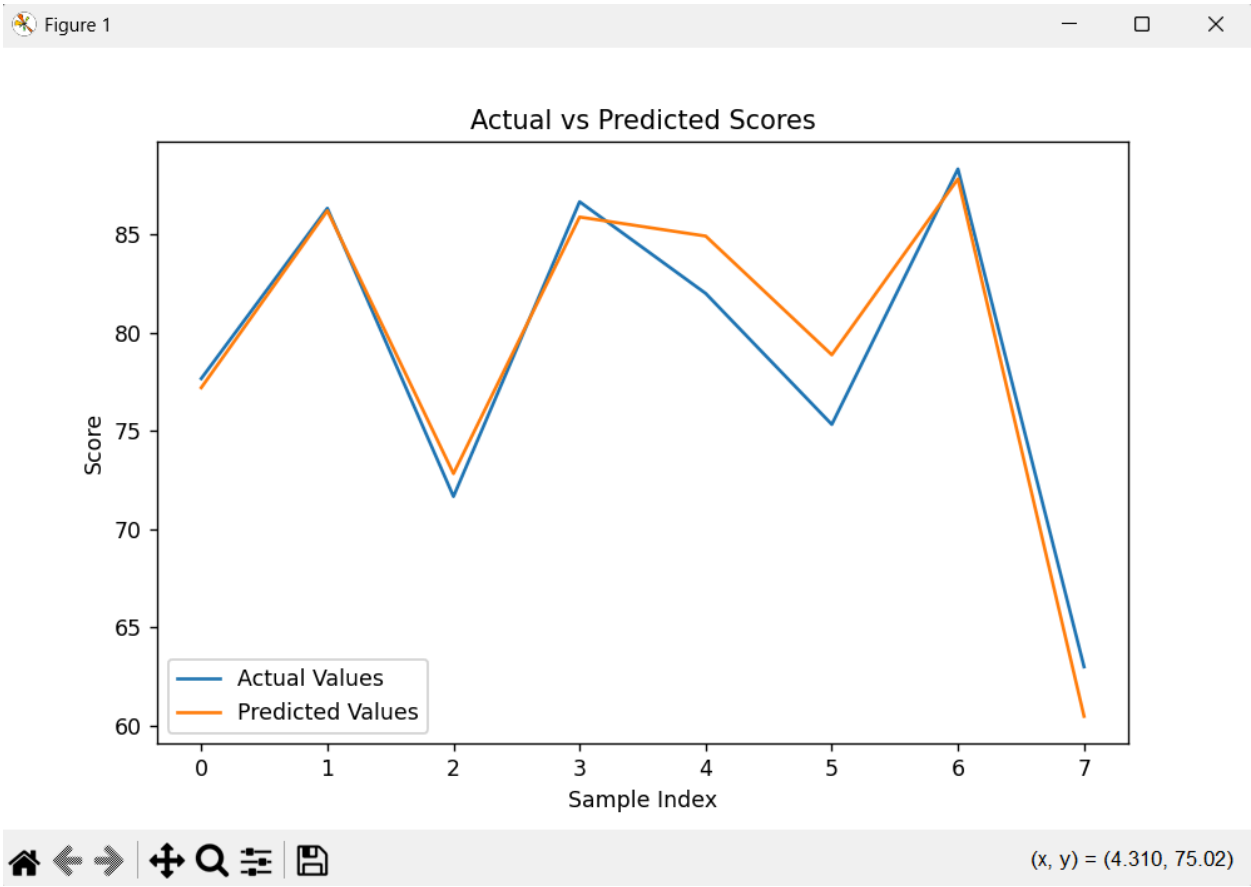
1. Study Hours vs Average Score



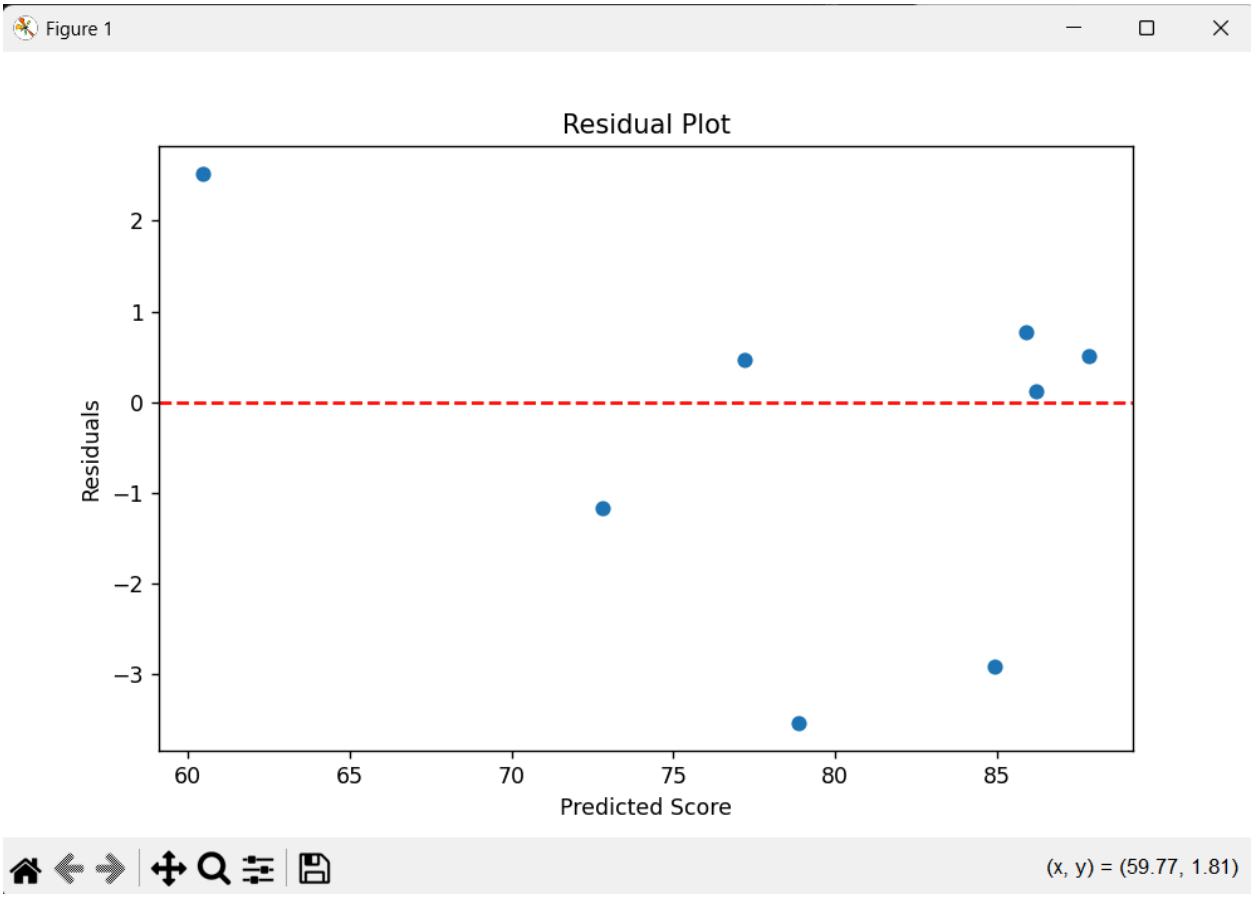
2. Attendance vs Average Score



3. Actual vs Predicted scores



4. Residual Plot



8. Applications

- Identifying at-risk students early.
- Academic planning and counselling.
- Personalized learning dashboards.
- AI-based performance monitoring systems.

9. Conclusion

The project successfully demonstrates the use of **Linear Regression** as an AI technique to predict student performance.
With an R^2 score of **0.93**, the model shows excellent predictive capability.

This mini project highlights how simple AI models can be applied to real-world educational environments and assist in early prediction and decision support.

10. Future Scope

- Add more features like sleep hours, background, and stress levels.
- Experiment with algorithms like Random Forest, SVM, or Gradient Boosting.
- Deploy the model as a simple web application.
- Use classroom real-time data for continuous learning.

11. Technologies Used

- Python
- Pandas
- NumPy
- Matplotlib
- Scikit-learn
- Jupyter / VSCode

12. References

- Scikit-learn Machine Learning Documentation
- Python Data Science Handbook
- Research articles on Educational Data Mining