Predicting Student Exam Scores Based on Study Hours Using Regression Analysis

Bhavesh Naidu Kulluru

717541.

For this project, I used **Linear Regression**, a supervised machine learning algorithm that models the relationship between a dependent variable (exam score) and an independent variable (hours studied). Linear regression is simple yet effective for problems with a clear linear trend. The dataset, "Student Marks vs. Study Hours," was sourced from **Kaggle**, a public repository of datasets. To implement the analysis, we utilized popular Python libraries such as **pandas** for data manipulation, **scikit-learn** for model building and evaluation, **seaborn** and **matplotlib** for data visualization, and **numpy** for numerical operations.

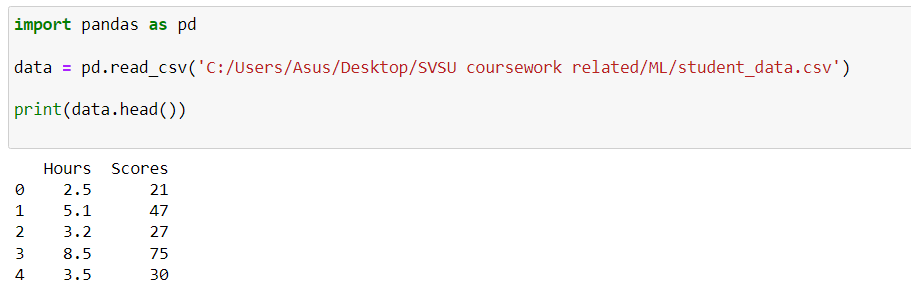
(The reason for this specific dataset and model selection is because I have created the same ML model in my previous internship few years ago in my undergrad, almost 90% of the code is re-used from my github <https://github.com/RiskyTrick>)

**Dataset Overview:**

**Dataset link :** <https://www.kaggle.com/datasets/mykeysid10/tsf-ds-ba-task-1>

The dataset used for this analysis contains two key variables:

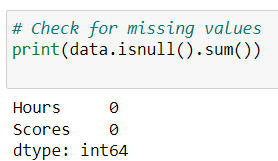
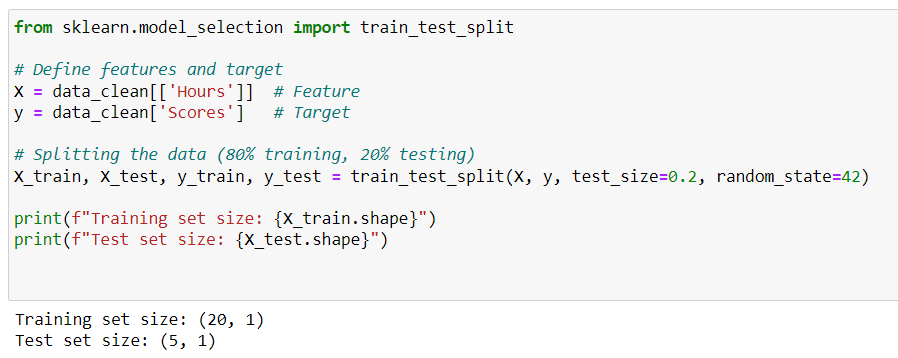
* **Hours Studied:** The number of hours students spend studying.
* **Exam Score:** The percentage score achieved in the exam.



The objective of the study is to predict the student's exam score based on the number of hours they study.

**Data Preprocessing:**

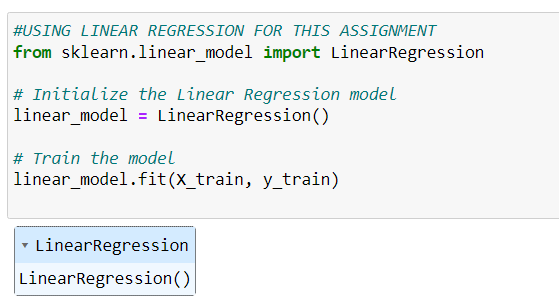
* **Missing Values:** No missing values were found in the dataset, so no imputation or removal of records was necessary.
* **Outliers:** Upon visualizing the dataset, no significant outliers were identified.
* **Data Splitting:** The dataset was split into training and testing sets with an 80/20 ratio. This ensures that the model’s performance is evaluated on unseen data (the test set).

**Model Selection:**

For this task, **Linear Regression** was chosen due to the simple, linear relationship between hours studied and exam score.

* **Linear Regression:** This model assumes a direct linear relationship between the input feature (hours studied) and the target variable (exam score). The regression line is fitted by minimizing the sum of squared differences between the actual and predicted values.

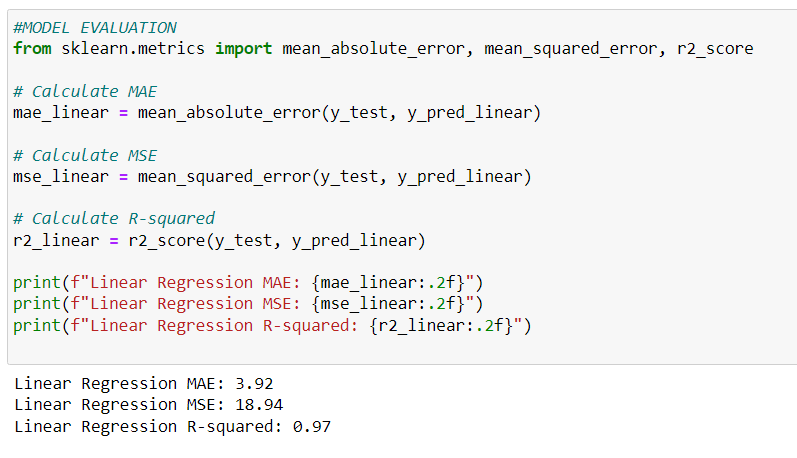


**Model Evaluation:**

After training the model, its performance was evaluated using key regression metrics:

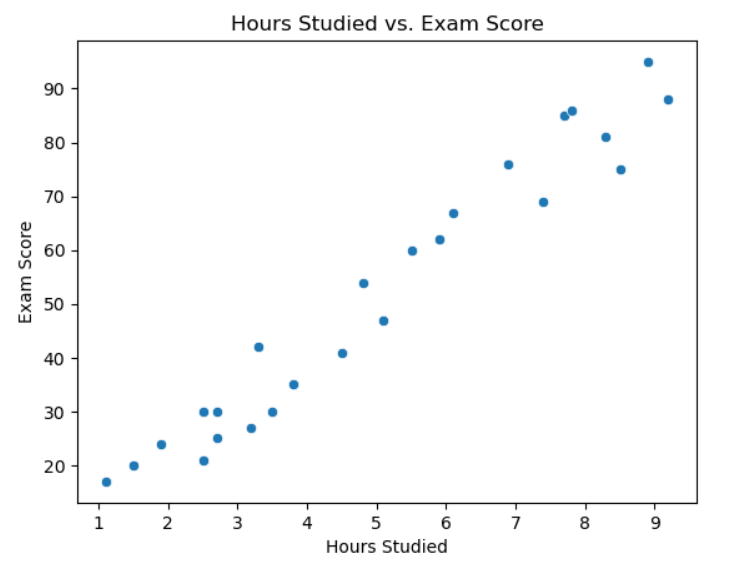
* **Mean Absolute Error (MAE):** 3.92
* **Mean Squared Error (MSE):** 18.94
* **R-squared (R²):** 0.97

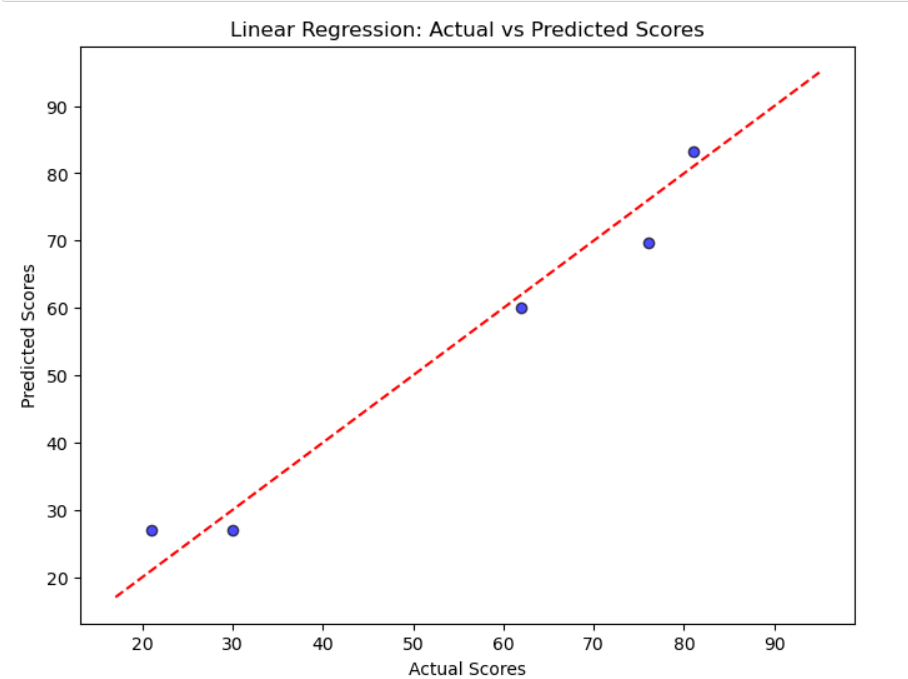
The high R-squared value of 0.97 indicates that 97% of the variance in exam scores is explained by the number of hours studied. The model performed very well, predicting student scores with minimal error.



**Visualizations:**

* **Hours Studied vs. Exam Scores Scatter Plot:** A scatter plot of hours studied versus exam scores shows a clear positive correlation, where students who studied for more hours tended to have higher exam scores.





Bhavesh Naidu Kulluru

717541.