Farezki Maulana Putra

Student scores Prediction

ibimbing

Project Objective

This mini portfolio project aims to explore and predict student performance based on their study habits. Specifically, the objectives are:

01 To analyze the relationship between the number of study hours and the scores achieved by students in an exam.

02 to perform data analysis through summary statistics and visualization to gain insights from the dataset.

03 To apply machine learning regression models to predict scores based on study hours using real data.

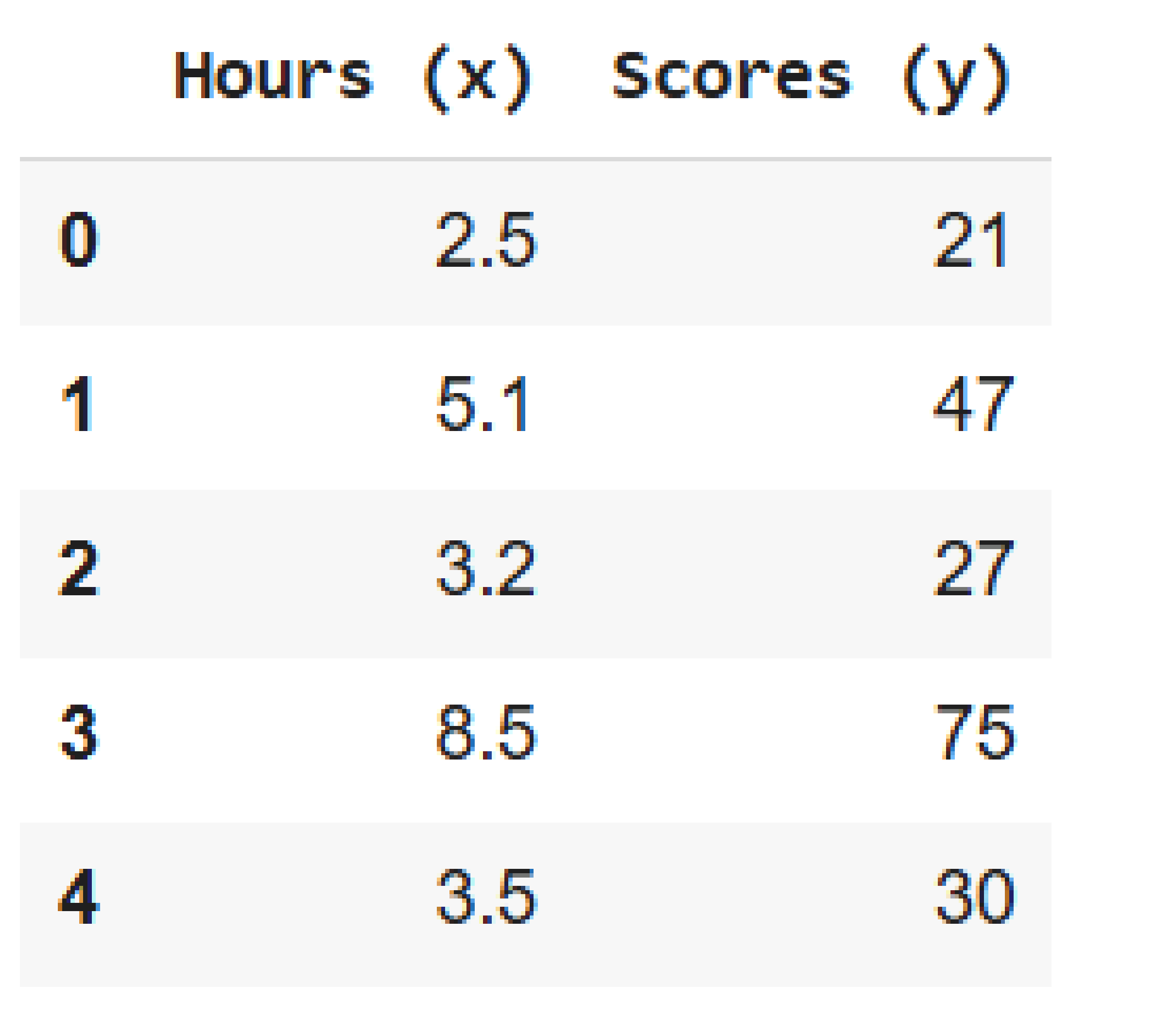
04 To compare different regression models (Linear Regression and Decision Tree Regressor) and evaluate their performance.

05 To determine the most accurate model for predicting student scores, and draw conclusions about model effectiveness on small datasets.

Dataset Overview

The dataset used in this project is named student_scores.xlsx. It is a simple dataset containing records of students' study habits and corresponding exam scores.

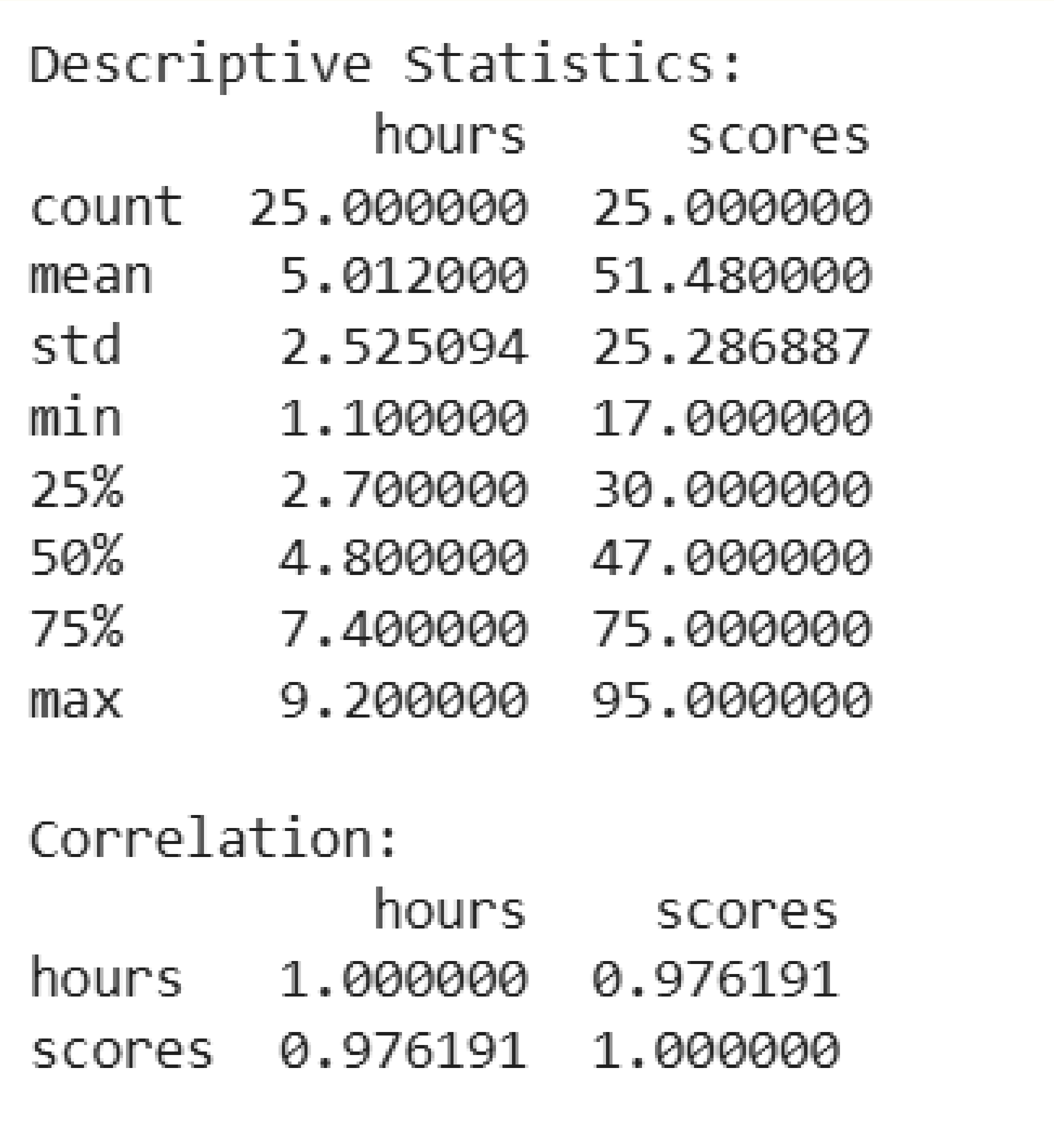
The dataset is structured as follows:



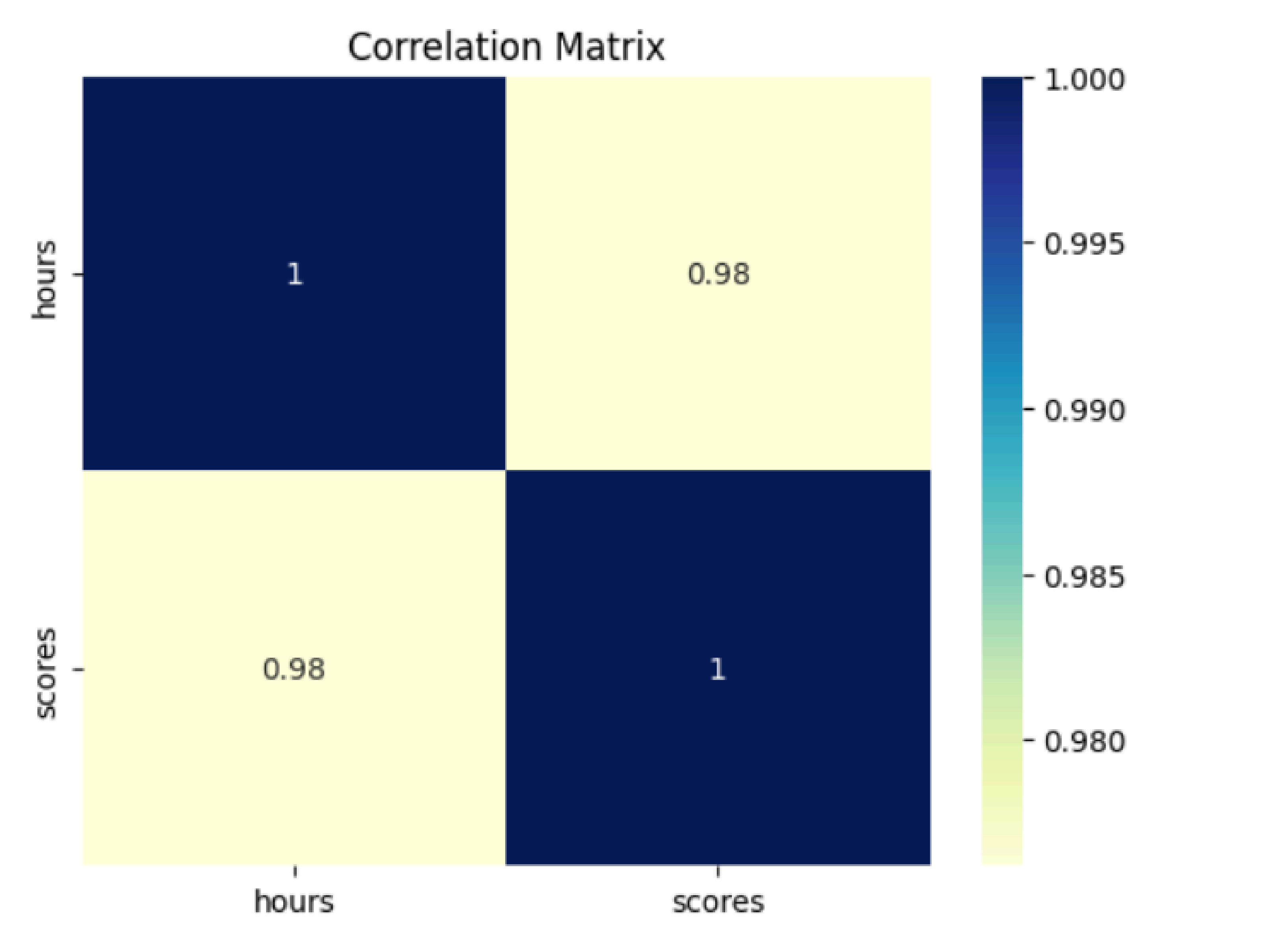
	Hours (x)	Scores (y)
0	2.5	21
1	5.1	47
2	3.2	27
3	8.5	75
4	3.5	30

Exploratory Data Analysis (EDA)

In this stage, we explore the dataset to understand its structure, patterns, and relationships between variables. EDA helps identify trends, anomalies, and prepare for modeling.



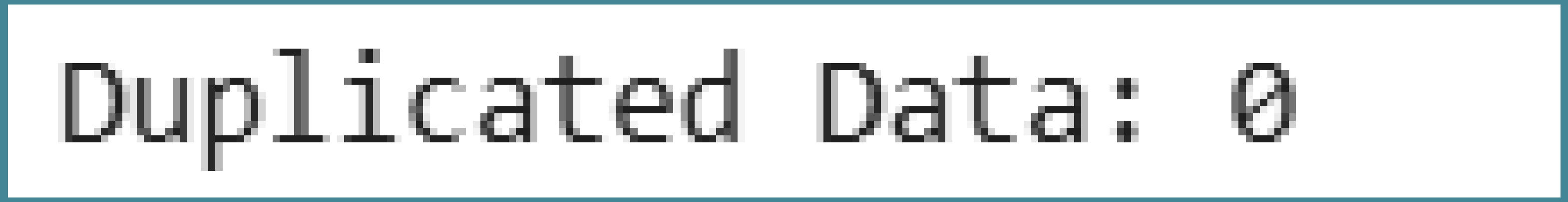
```
Descriptive Statistics:
           hours     scores
count  25.000000  25.000000
mean    5.012000  51.480000
std     2.525094  25.286887
min     1.100000  17.000000
25%     2.700000  30.000000
50%     4.800000  47.000000
75%     7.400000  75.000000
max     9.200000  95.000000

Correlation:
           hours    scores
hours   1.000000  0.976191
scores  0.976191  1.000000
```



Feature Engineering and Data Preparation

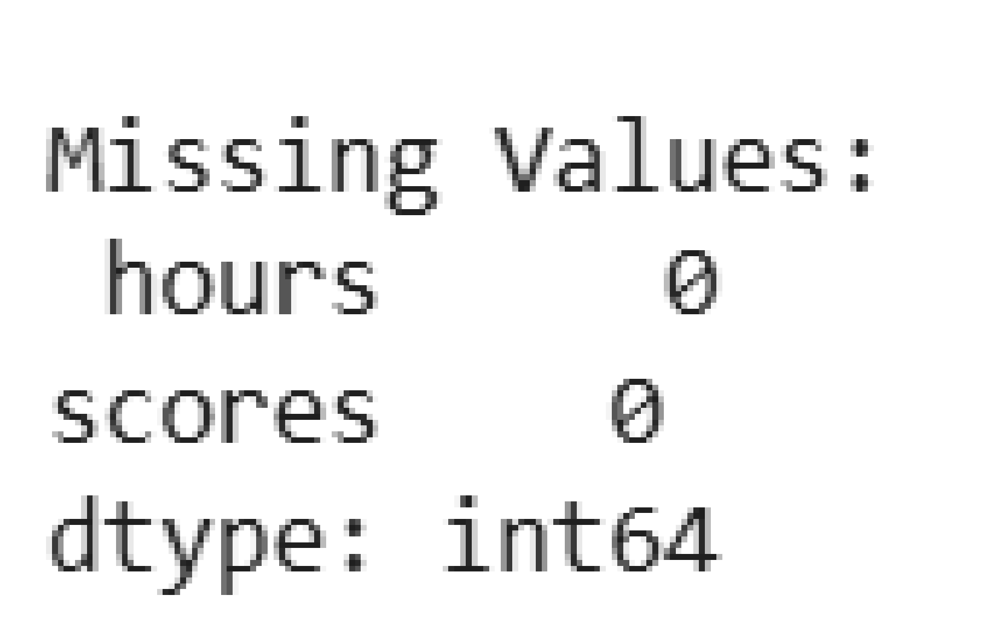
Before building machine learning models, we perform data quality checks and refine the dataset to ensure accuracy and reliability of the results.

Check for Duplicates



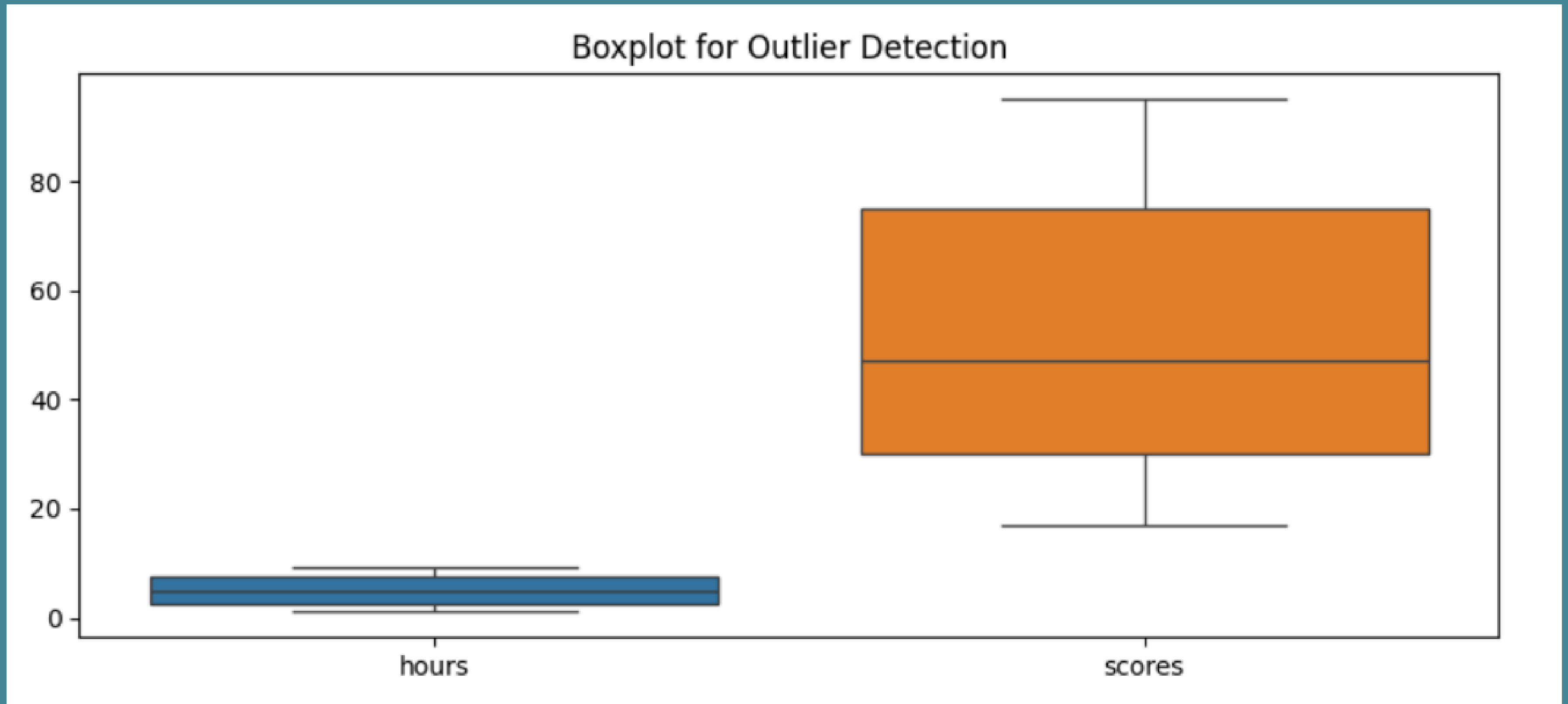
```
Duplicated Data: 0
```

Check for Missing Values



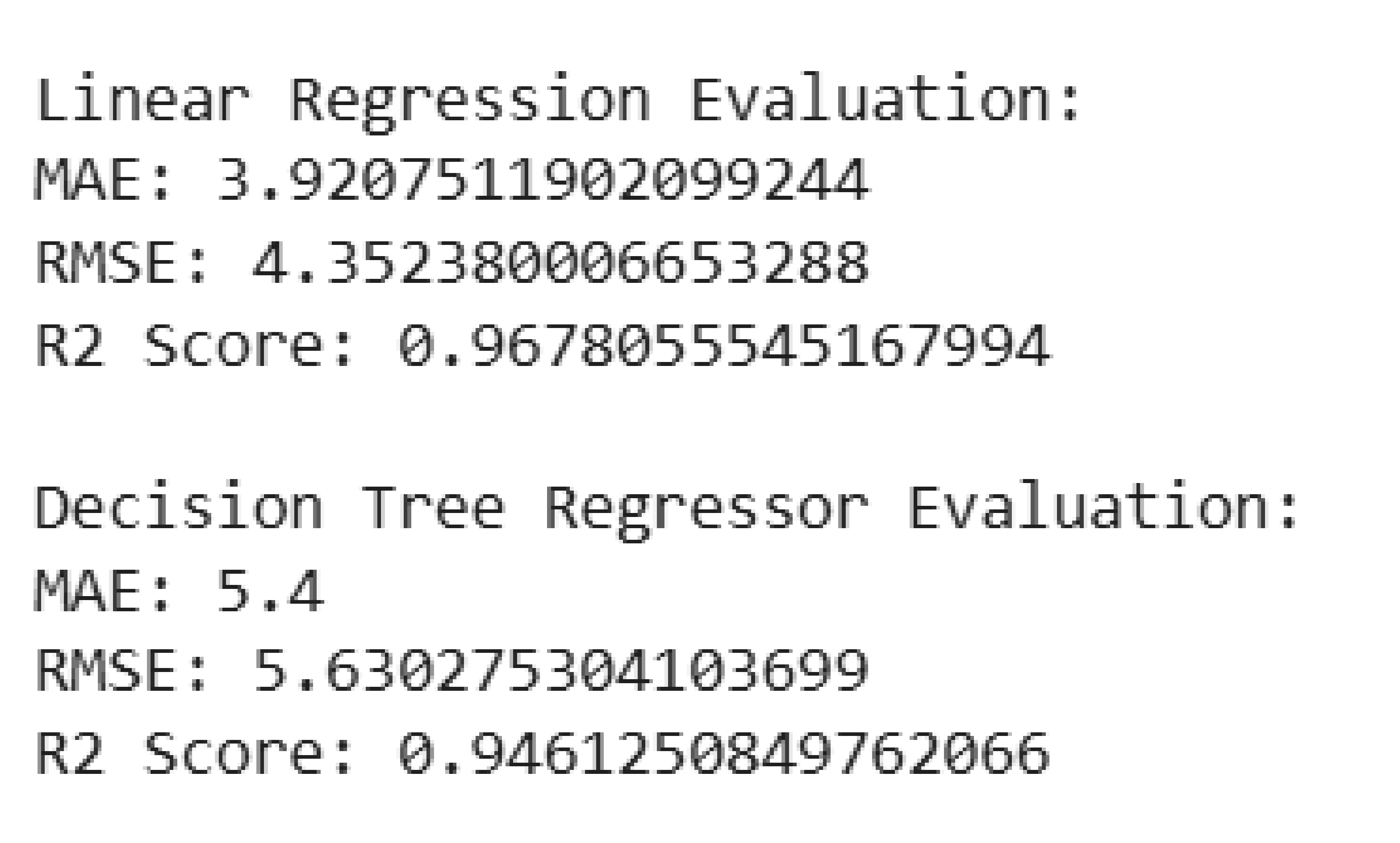
```
Missing Values:
 hours     0
scores    0
dtype: int64
```

Outlier Detection

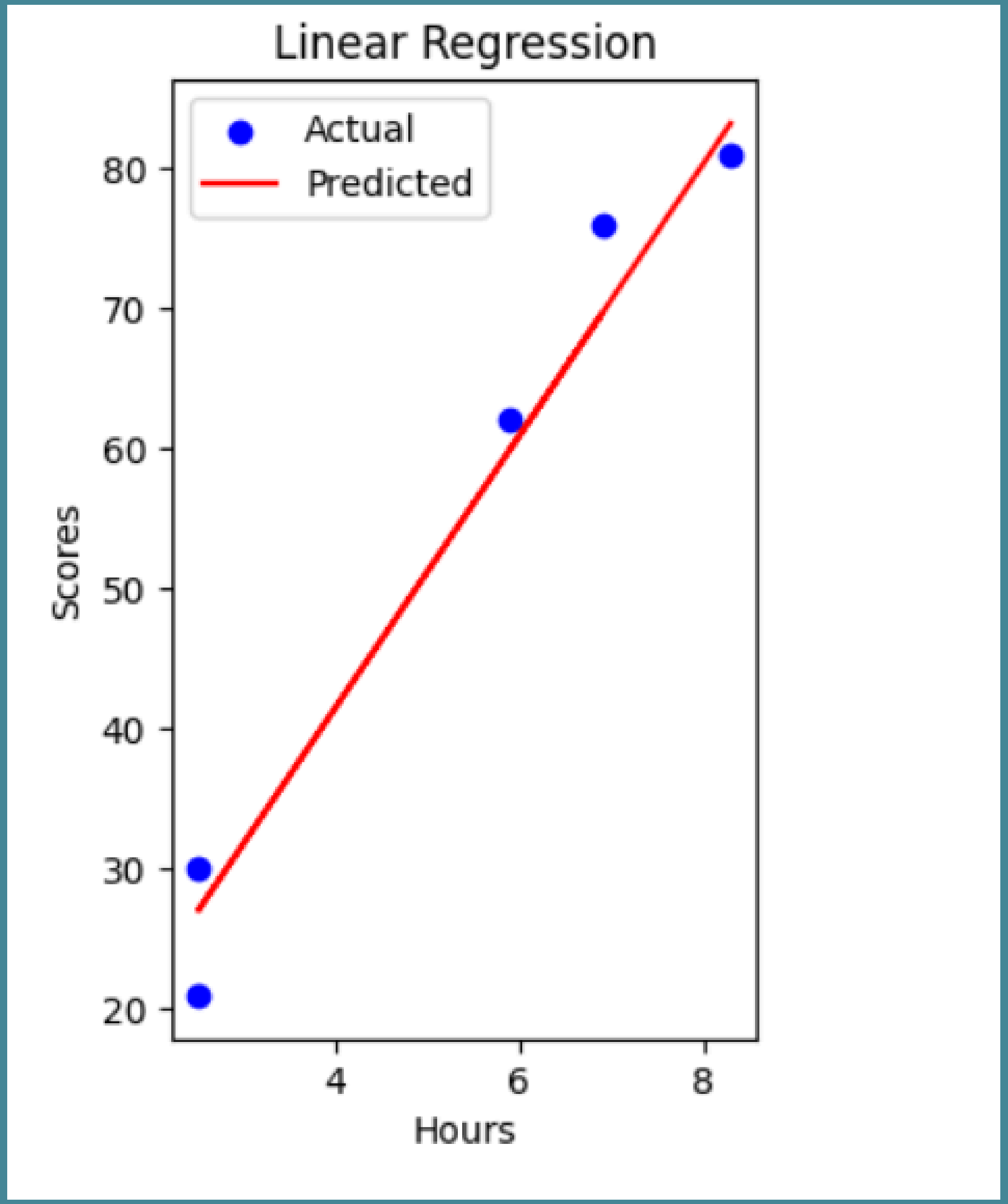


Model evaluation

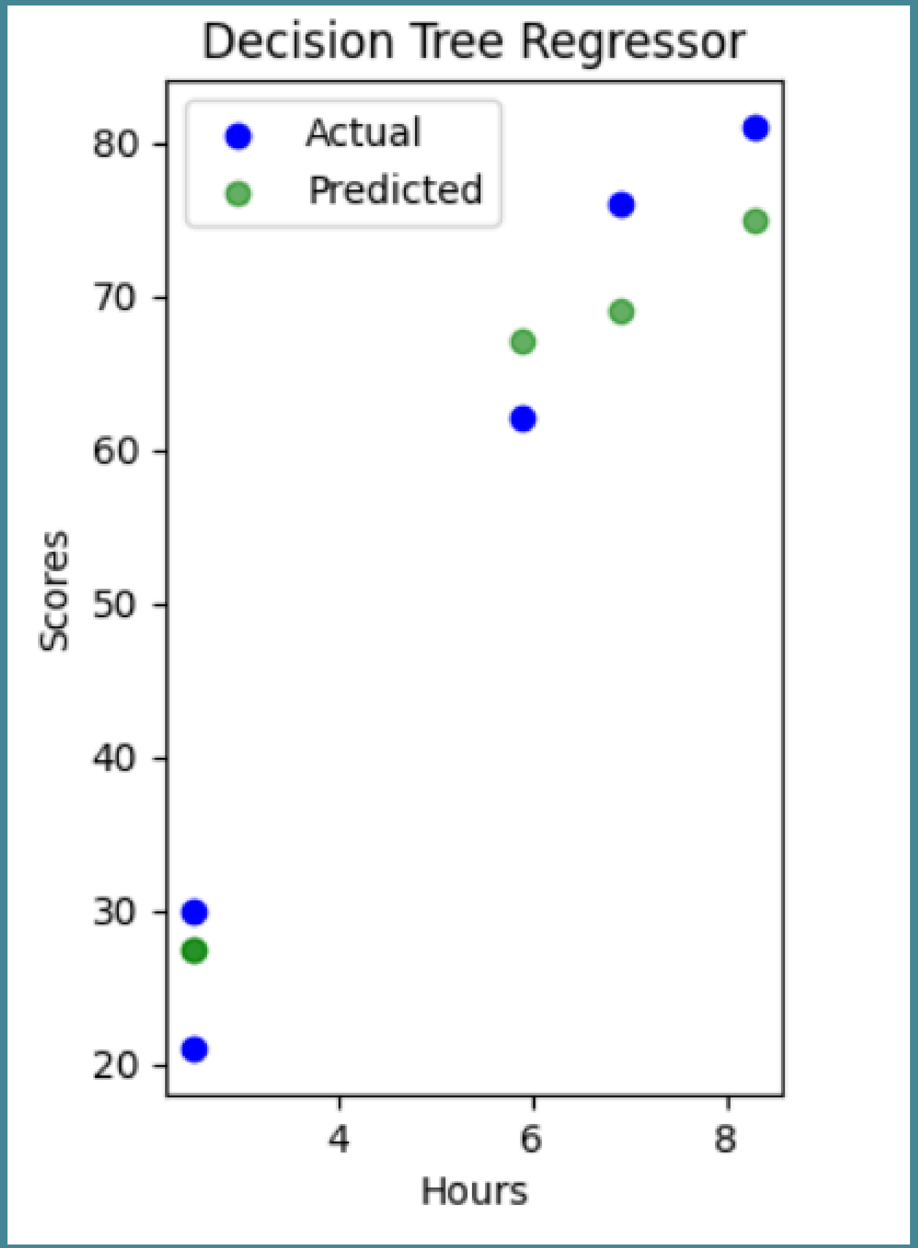
In this phase, we apply supervised machine learning techniques to train models that can predict student scores based on study hours. Since the output variable is numeric, regression models are used.



```
Linear Regression Evaluation:
MAE: 3.9207511902099244
RMSE: 4.352380006653288
R2 Score: 0.9678055545167994

Decision Tree Regressor Evaluation:
MAE: 5.4
RMSE: 5.630275304103699
R2 Score: 0.9461250849762066
```



Model comparasion



Conclusion

After performing data analysis and building regression models, we arrive at the following conclusions:

1. There is a strong linear relationship between the number of study hours and exam scores (correlation ≈ 0.98).
2. The dataset is clean — no missing values or significant outliers.
3. Two models were used: Linear Regression and Decision Tree Regressor.
4. Linear Regression performed better across all evaluation metrics (MAE, RMSE, R^2).
5. Linear models are suitable for simple and strongly linear datasets like this one.

Final note:
Linear Regression is the recommended model for score prediction in this case. It provides accurate results, is simple to implement, and aligns well with the data characteristics.

Thankyou