NADHIF RIF'AT RASENDRIYA

STUDENT SCORES PREDICTION

USING MACHINE LEARNING MODELS

ibimbing

Overview

Predicting Student Scores Based on Study Hours Using Regression Models.

This project aims to predict students' exam scores based on the number of hours they studied using various regression models. The dataset contains two features: study hours and corresponding scores. The process involves data cleaning, exploratory data analysis, training three machine learning models (Linear Regression, Decision Tree Regressor, and Random Forest Regressor) and evaluating them using performance metrics like Mean Squared Error (MSE) and R-squared (R^2). The goal is to identify the most accurate model for predicting student performance.

Approach

- Data Cleaning & Feature Validation
- Exploratory Data Analysis (EDA)
- Model Training: Linear Regression, Decision Tree, Random Forest
- Evaluation using MSE and R^2 Score
- Visual Comparison of Predictions

Full Code & Datasets

GITHUB

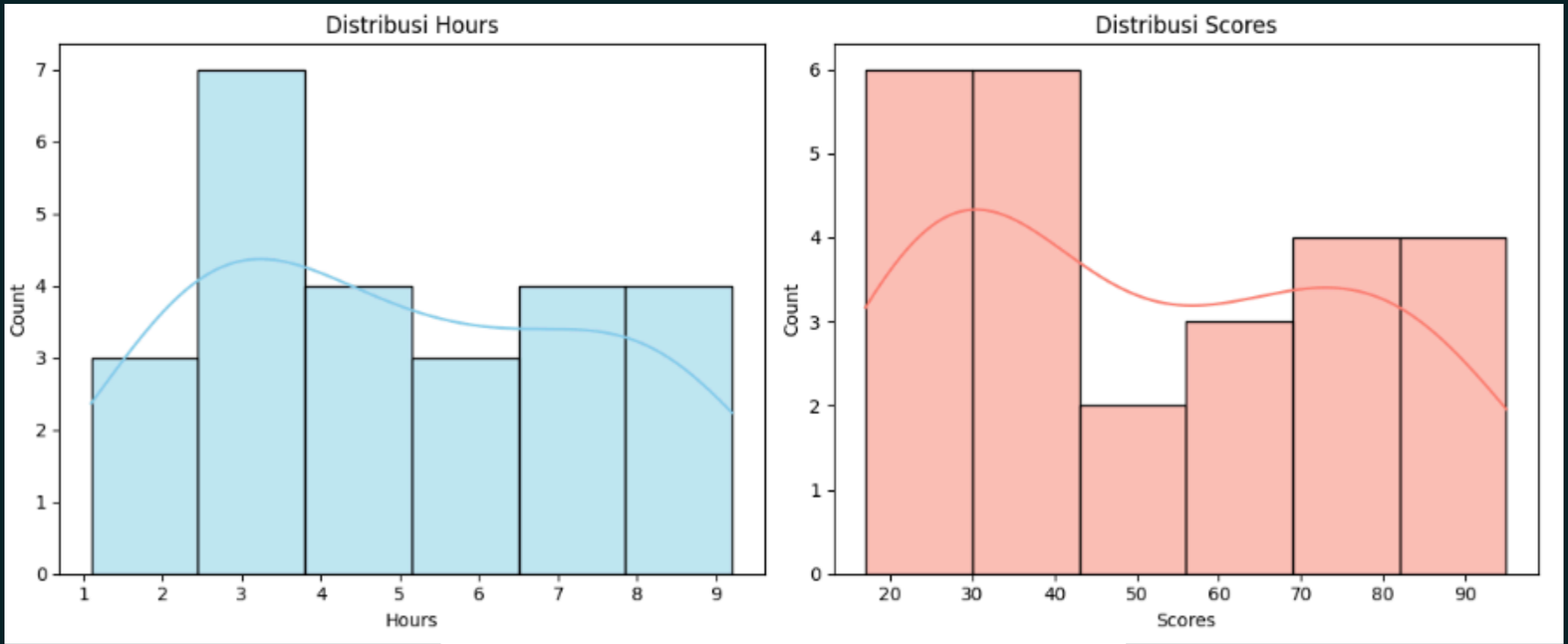
https://github.com/nadhif-royal/DSF38_MachineLearning

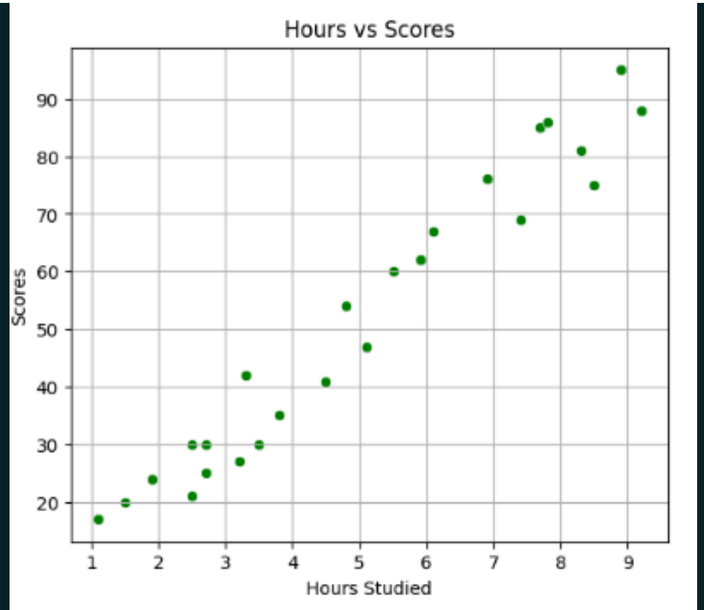
Scan here

Dataset:

- Source: student_scores.csv
- Features:
  - Hours: Hours studied
  - Scores: Exam scores

Exploratory Data Analysis (EDA)





In this phase, several visual and statistical techniques were used to better understand the dataset.

Descriptive statistics revealed the central tendency and spread of the "Hours" and "Scores" variables. Histograms with KDE (Kernel Density Estimation) provided insights into the distribution of both features, where "Hours" showed a right-skewed pattern and "Scores" appeared slightly bimodal.
A scatter plot was used to visualize the relationship between hours studied and student scores, indicating a strong positive correlation. To support this, a correlation heatmap confirmed the linear relationship between the two variables. Lastly, boxplots were included to detect the presence of outliers, and both features appeared to be relatively clean with no extreme outliers.

Model Evaluation

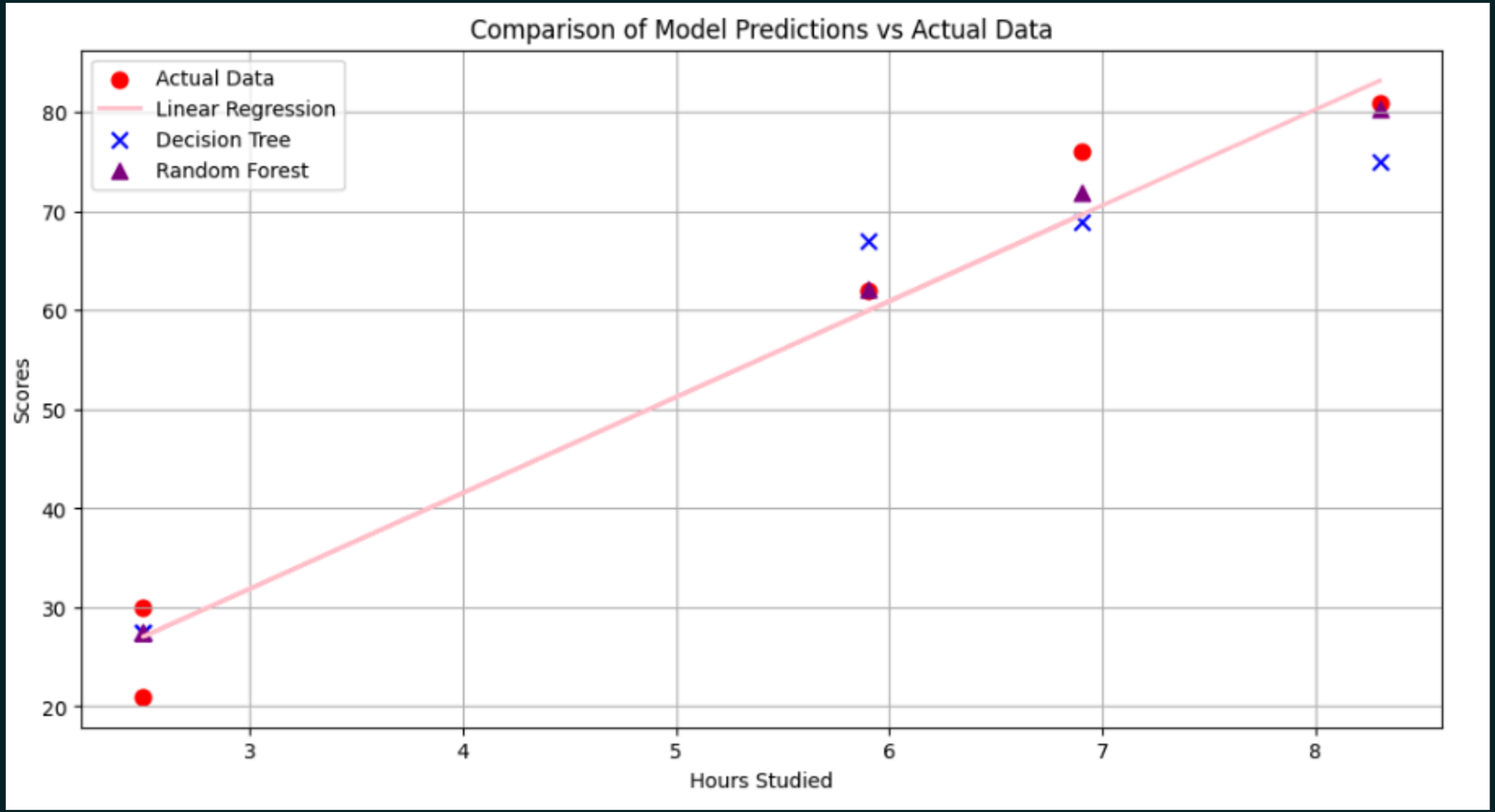
Model	MSE	R^2 Score
Linear Regression	18.94	0.968
Decision Tree Regressor	31.70	0.946
Random Forest Regressor	**13.05**	**0.978**

Linear Regression
- Mean Squared Error (MSE): 18.943211722315272
- R-squared (R2): 0.9678055545167994

Decision Tree Regressor
- Mean Squared Error (MSE): 31.7
- R-squared (R2): 0.946125084976206 6

Random Forest Regressor
- Mean Squared Error (MSE): 13.045153611111104
- R-squared (R2): 0.9778294466160586

Model Comparison and Evaluation



The visualization above compares the actual student scores (in red dots) with predictions made by three different machine learning models: Linear Regression (line), Decision Tree (blue X), and Random Forest (purple triangle).

Each model was evaluated based on how closely their predicted values matched the real data points. From the plot, we can observe that the Linear Regression model provided a smooth and consistent approximation, while the Decision Tree and Random Forest models captured more localized variations.

This comparison helps in understanding the performance of each model in predicting student scores based on hours studied.

Conclusion

The **Random Forest Regressor** is the best-performing model for this dataset because:

- **Lowest MSE (13.05):** This indicates that its predictions are the closest to the actual values compared to the other models.
- **Highest R^2 Score (0.978):** This means the model is able to explain approximately 97.78% of the variance in the target data (Scores), outperforming both Linear Regression ($R^2 = 0.968$) and Decision Tree Regressor ($R^2 = 0.946$).

THANK YOU

https://www.linkedin.com/in/royalnadhif50/