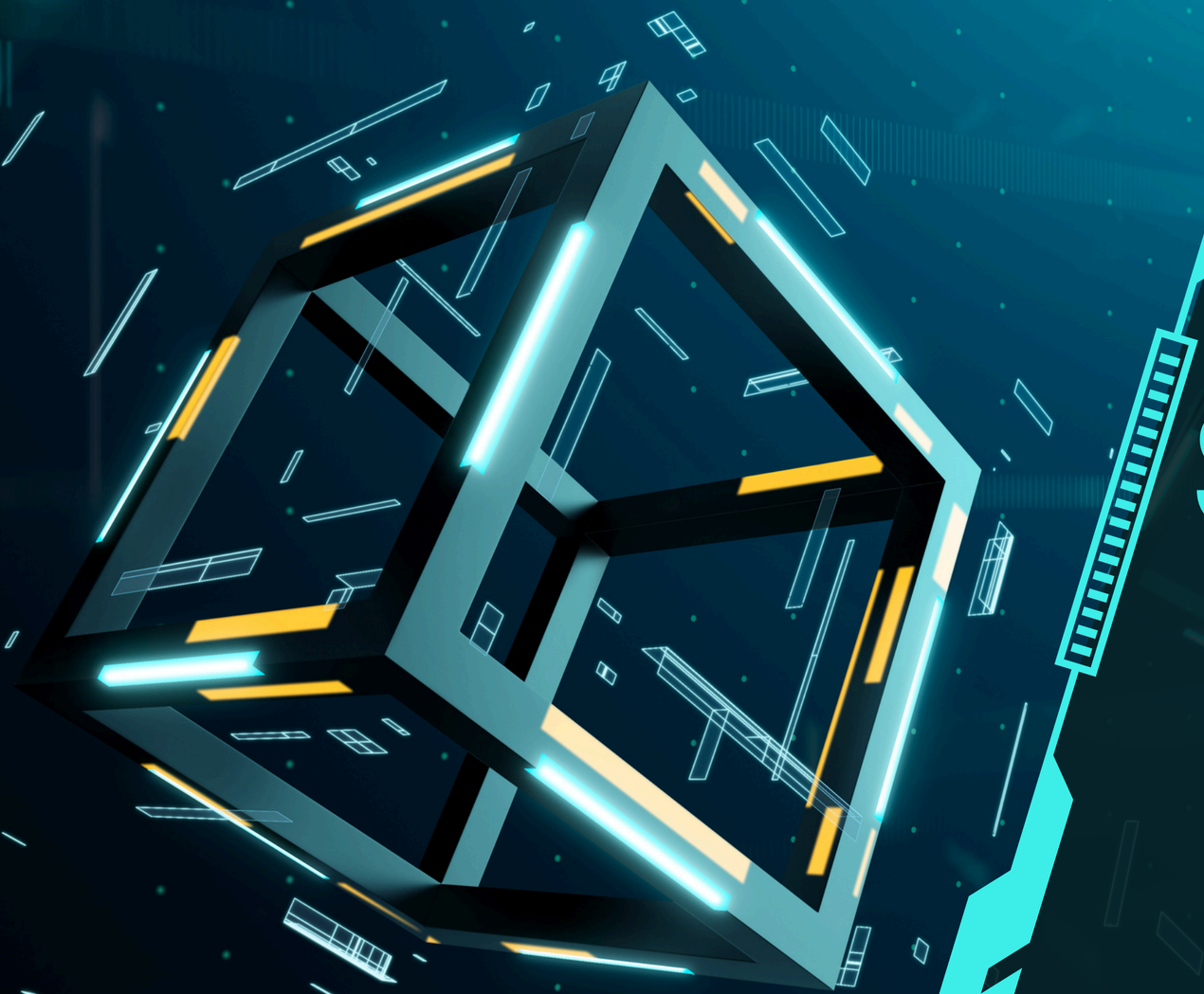
NADHIF RIF'AT RASENDRIYA



ibimbing

STUDENT SCORE PREDICTION

MACHINE LEARNING

Models

Linear Regression

Decision Tree

Random Forest

Overview

In this project, we explore different machine learning models to predict student scores based on available data. The goal is to compare the performance of Linear Regression, Decision Tree, and Random Forest in predicting student performance, identifying which model provides the most accurate and reliable results.

Why This Project?

Student performance prediction is crucial in educational settings, helping educators and institutions identify students who may need additional support. By leveraging machine learning, we can analyze patterns in student data and make data-driven predictions that can enhance learning strategies and outcomes.

Full Code & Datasets

GITHUB

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

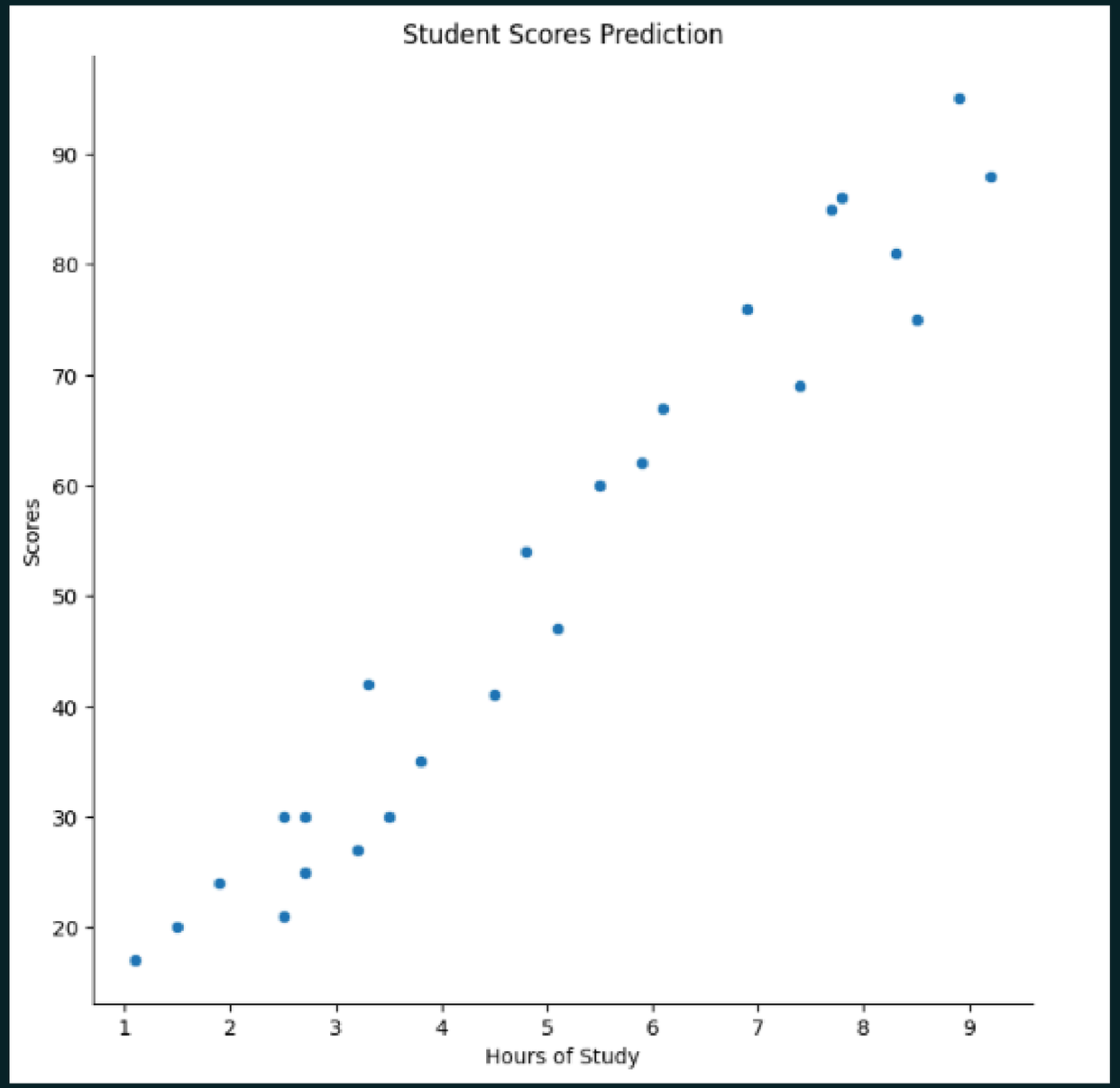
student_scores.csv

1 to 25 of 25 entries Filter

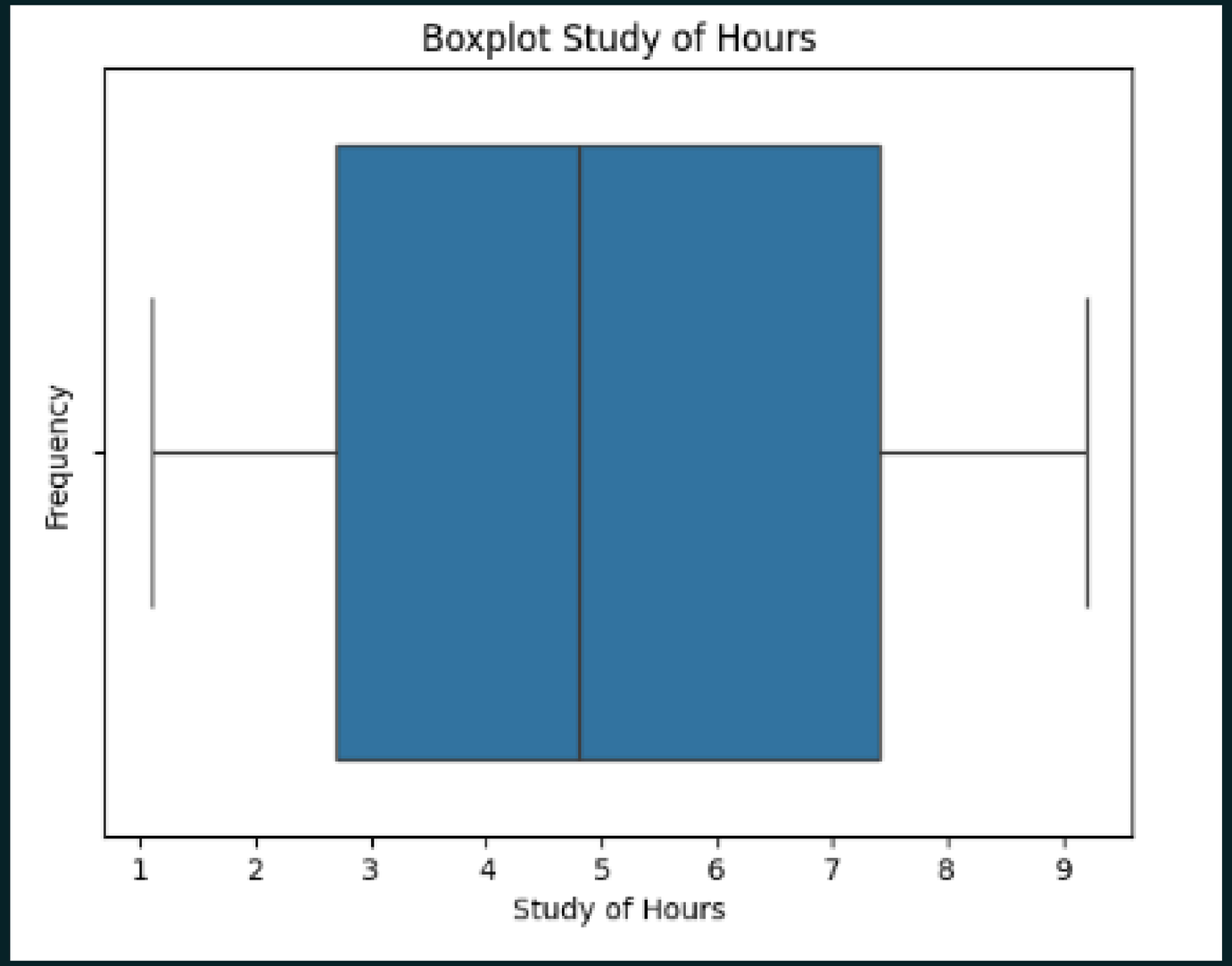
Hours	Scores
2.5	21
5.1	47
3.2	27
8.5	75
3.5	30
1.5	20
9.2	88
5.5	60
8.3	81
2.7	25
7.7	85
5.9	62
4.5	41
3.3	42
1.1	17
8.9	95
2.5	30
1.9	24
6.1	67
7.4	69
2.7	30
4.8	54
3.8	35
6.9	76
7.8	86

Show 25 per page

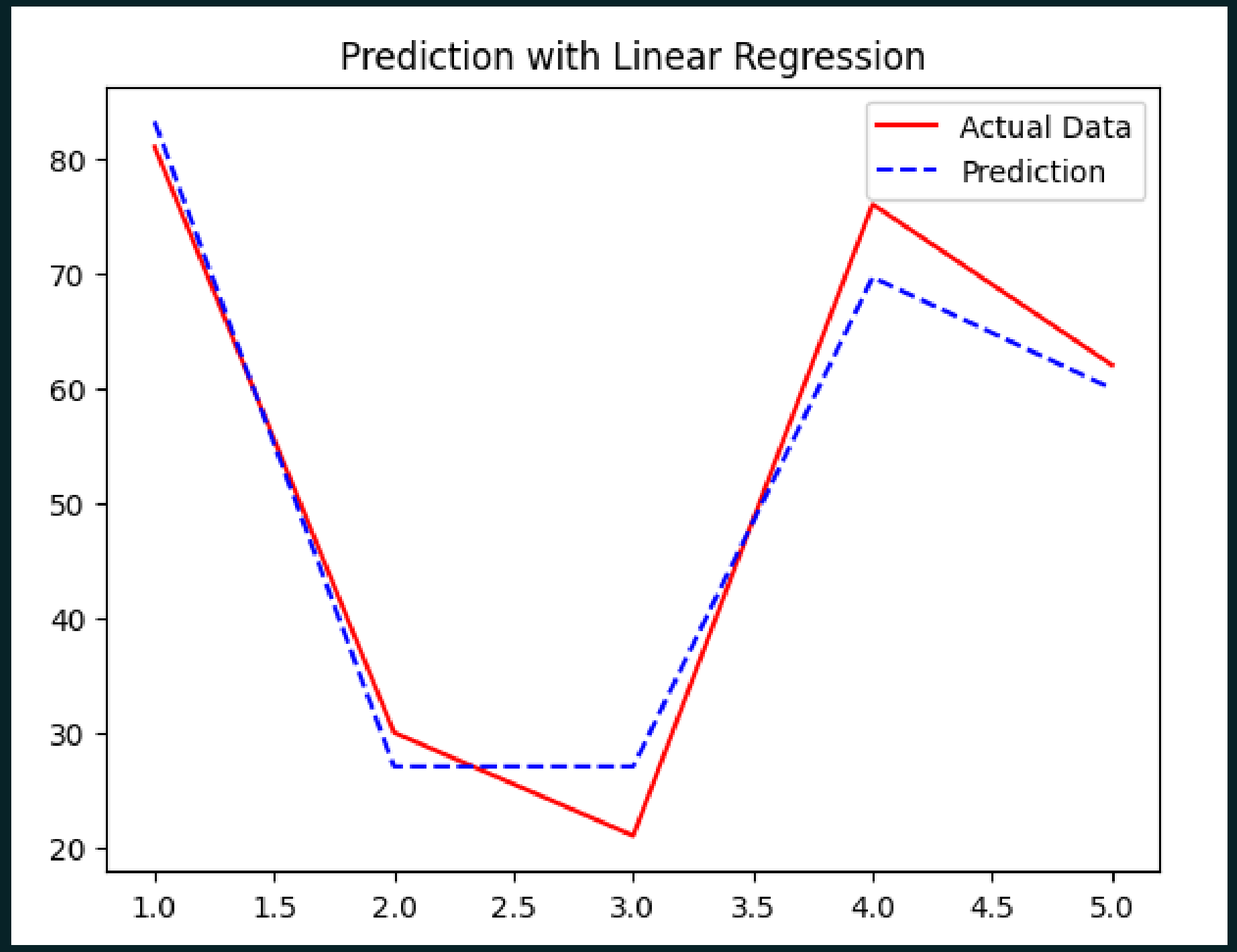
Plotting


Student Scores Prediction
Scores
Hours of Study
90
80
70
60
50
40
30
20
1
2
3
4
5
6
7
8
9

BloxPlot



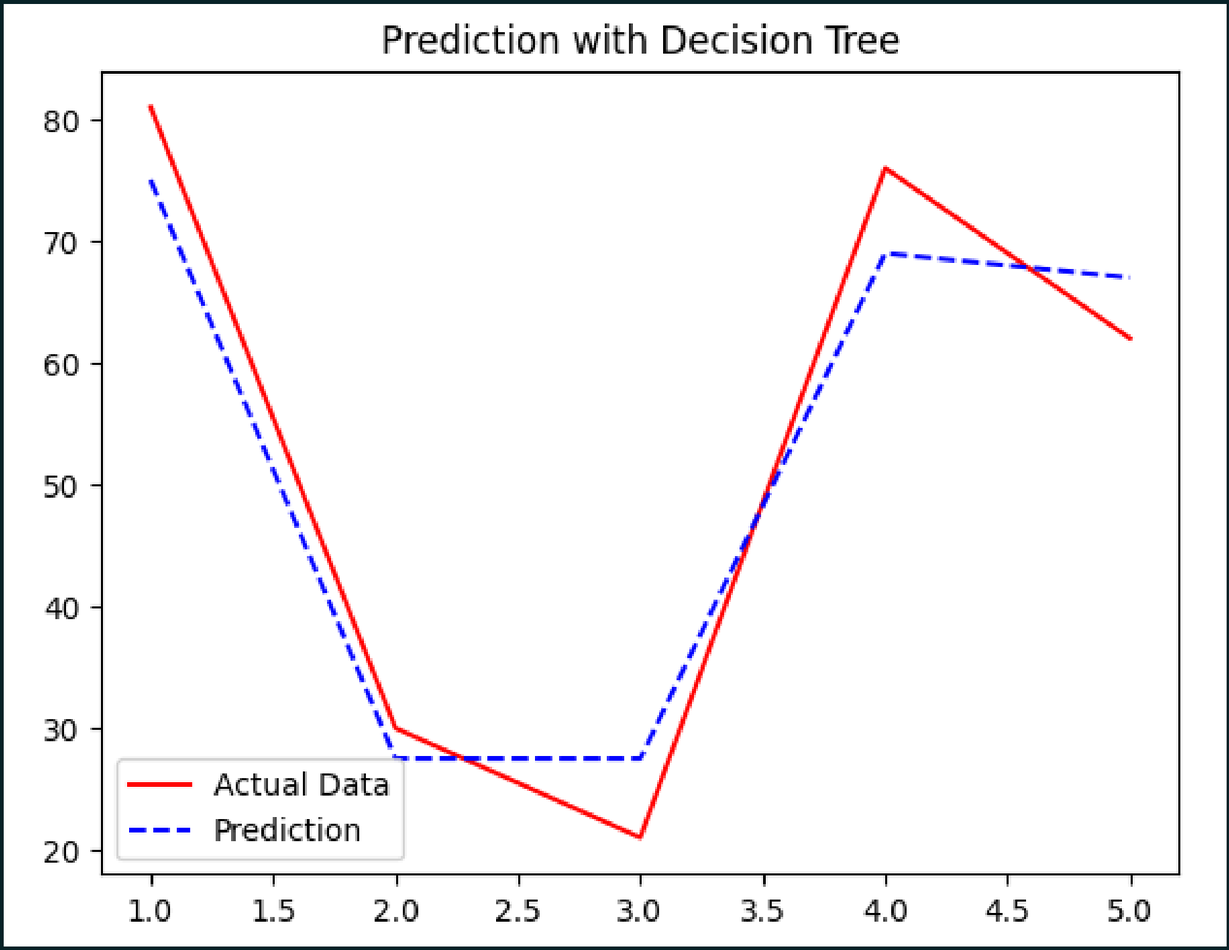
Result



The Linear Regression model effectively captures the relationship between study hours and student scores, as shown by the close alignment between the actual data (red line) and the predicted values (blue dashed line). While the model follows the general trend well, minor deviations occur at certain points, indicating that it may struggle with more complex variations in the data. Overall, the model provides a reasonable prediction but may benefit from enhancements such as polynomial regression or more advanced machine learning techniques to improve accuracy.

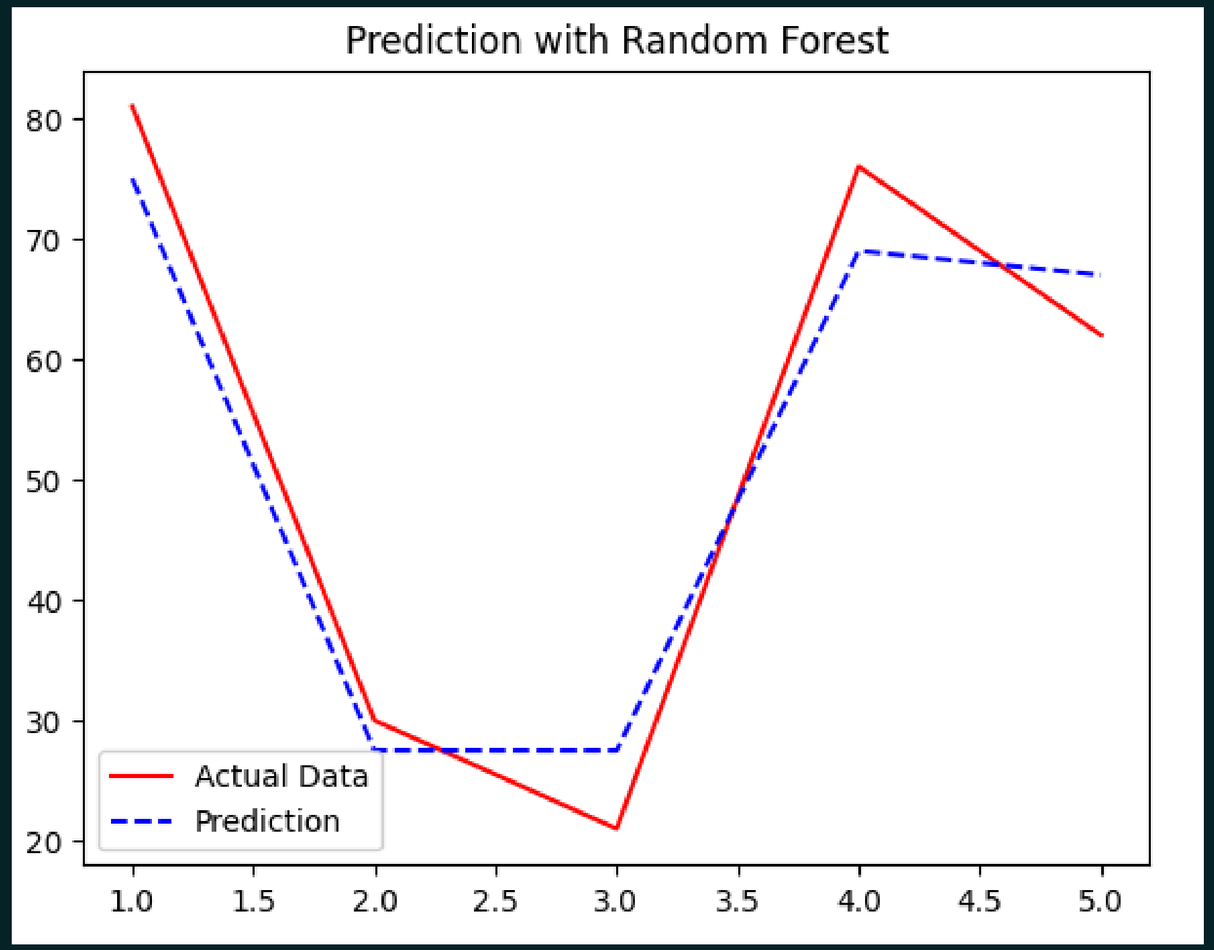
Linear Regression

Result



The Decision Tree model's predictions (blue dashed line) closely follow the actual data (red line), capturing sharp changes more effectively than linear regression. This indicates that the model can handle non-linear relationships well. However, minor discrepancies still exist, suggesting potential overfitting or sensitivity to small variations in the dataset. While Decision Trees provide more flexibility than linear regression, fine-tuning parameters such as depth and pruning can help improve generalization and avoid overfitting.

Result



The Random Forest model's predictions (blue dashed line) closely follow the actual data (red line) and show smoother transitions compared to the Decision Tree model. This suggests that Random Forest is better at reducing overfitting and generalizing to unseen data by averaging multiple decision trees. The prediction accuracy appears improved, as it captures the overall trend while minimizing drastic fluctuations. However, minor deviations still exist, indicating that further parameter tuning, such as adjusting the number of trees, might enhance performance.

Random Forest

Model Comparison

1. Linear Regression
    - Prediction Style: Produces a straight-line relationship between input and output.
    - Strengths: Works well for simple, linear relationships.
    - Weaknesses: Struggles with complex patterns or non-linear relationships, leading to poor predictions when the data is not well-represented by a straight line.
    - Result from The Plot: Likely had significant deviations from actual values due to its inability to capture the non-linearity in student scores.
2. Decision Tree
    - Prediction Style: Creates step-like decision boundaries, making abrupt jumps in predictions.
    - Strengths: Can model non-linear relationships and handle categorical features well.
    - Weaknesses: Prone to overfitting, meaning it memorizes training data but performs poorly on new data.
    - Result from The Plot: Captured the trend better than Linear Regression but showed sharp transitions, indicating overfitting.
3. Random Forest
    - Prediction Style: Averages multiple Decision Trees to smooth out predictions.
    - Strengths: Reduces overfitting by combining multiple trees, leading to more stable and accurate predictions.
    - Weaknesses: Can be computationally expensive and harder to interpret compared to a single Decision Tree.
    - Result from The Plot: Provided smoother and more accurate predictions, following the actual data closely while avoiding extreme fluctuations

Final Verdict

- Linear Regression is too simplistic for this dataset.
- Decision Tree learns well but tends to overfit.
- Random Forest offers the best balance between accuracy and generalization, making it the best choice among the three.

Here are the results of r square (R^2):

```
Linear Regression = 0.967
Decision Tree = 0.946
Random Forest = 0.981
```

Conclusion

The analysis results indicate that the Random Forest model achieved the highest performance ($R^2 = 0.981$), outperforming Linear Regression ($R^2 = 0.967$) and Decision Tree ($R^2 = 0.946$). This suggests that Random Forest is the most accurate model, as it better captures data variability compared to the others. Therefore, Random Forest is recommended as the optimal model for predicting student scores based on study hours.

THANK YOU

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