

Data Series 17.0

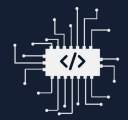


Artificial Intelligence Machine Learning



With Final Project Salary Prediction





What is AI/ML?

Artificial Intelligence (AI) is a technology that enables computers to perform tasks that normally require human intelligence. AI can learn, reason, plan, and solve problems.

Machine learning (ML) is a branch of AI focused on enabling computers and machines to imitate the way humans learn, perform tasks autonomously, and improve their performance and accuracy through experience and exposure to more data.

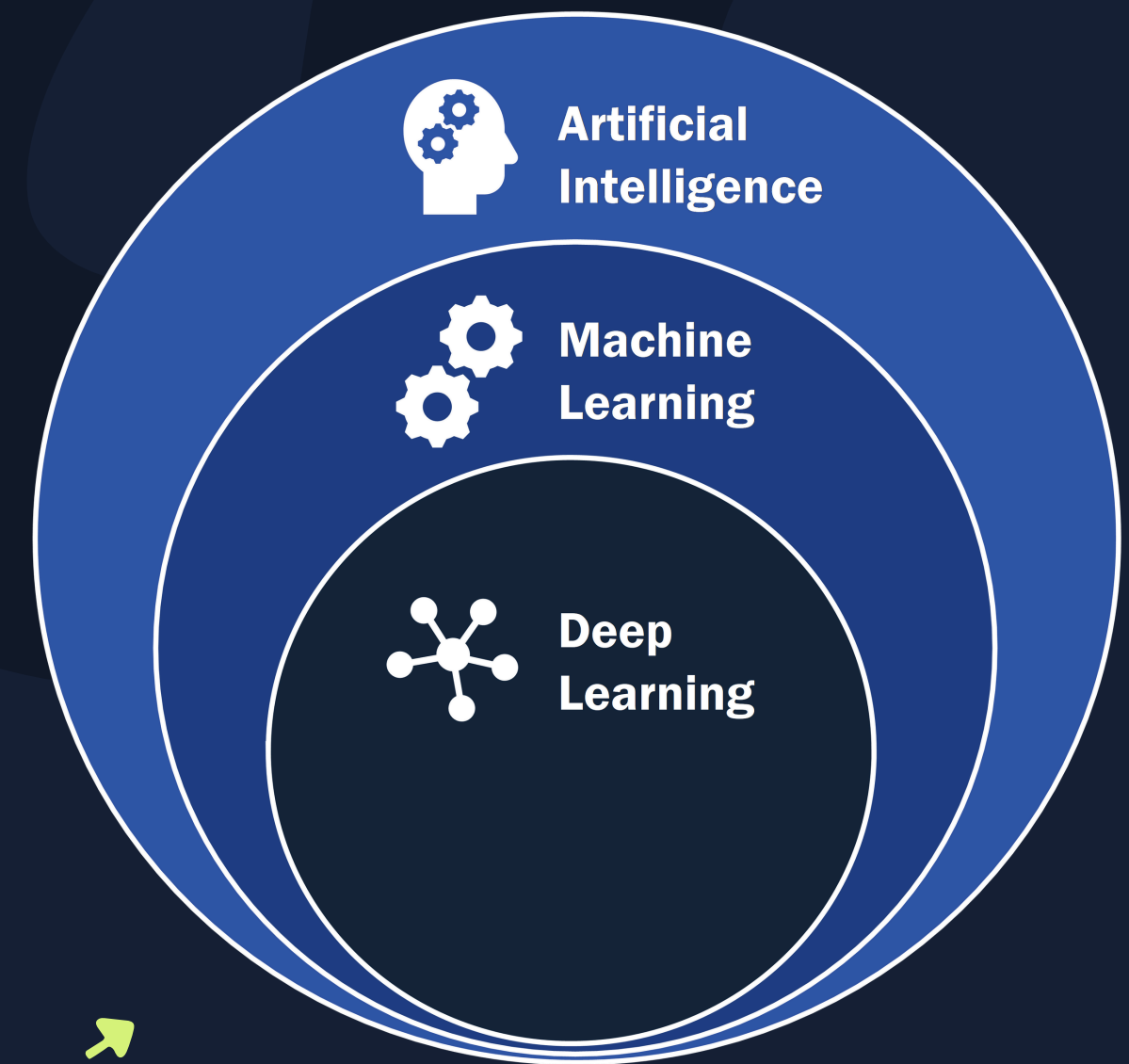
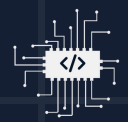
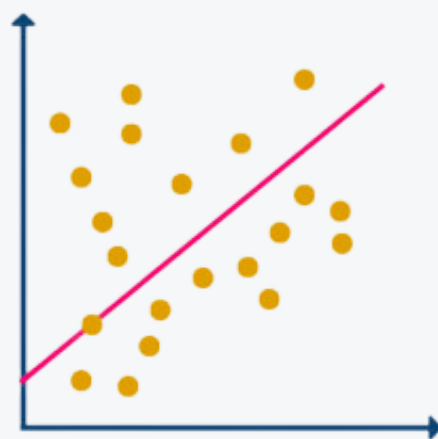


Image Source

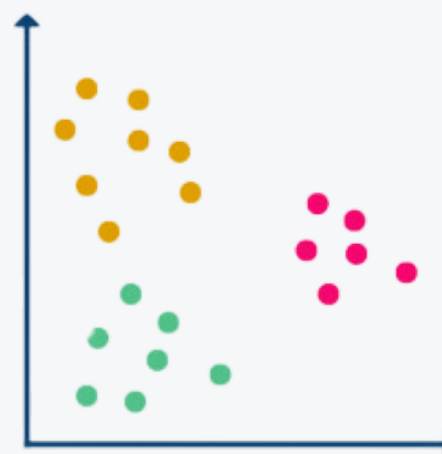


Category of Machine Learning

Supervised

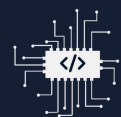


Unsupervised

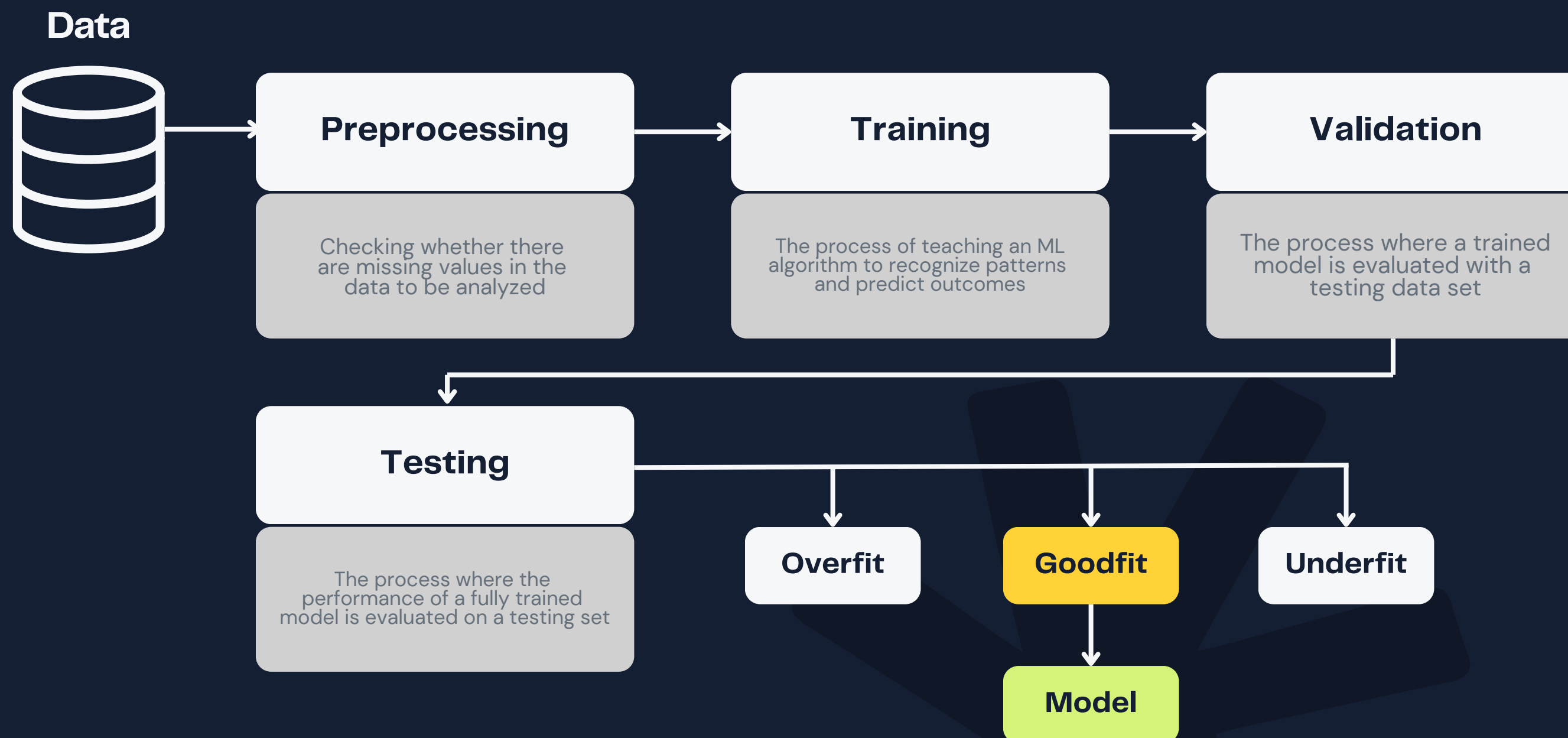


Reinforcement





Machine Learning Process





s.erzv

serzv.my.id

What is Overfitting/Underfitting?

Overfitting is a machine learning issue that occurs when a **model is too closely trained to a training set**, making it unable to accurately predict new data.

Underfitting is a machine learning issue that occurs when a **model is too simple to capture the patterns** in training data. This results in poor performance on both training and test data.

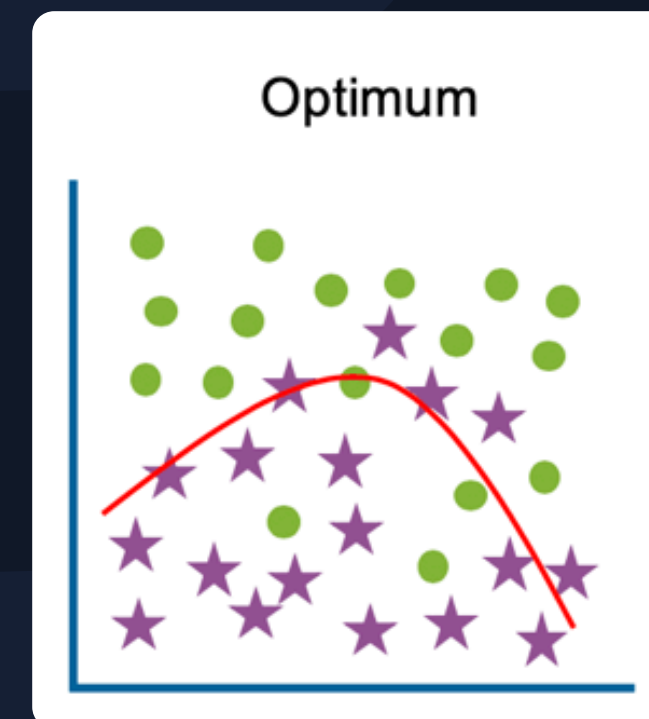
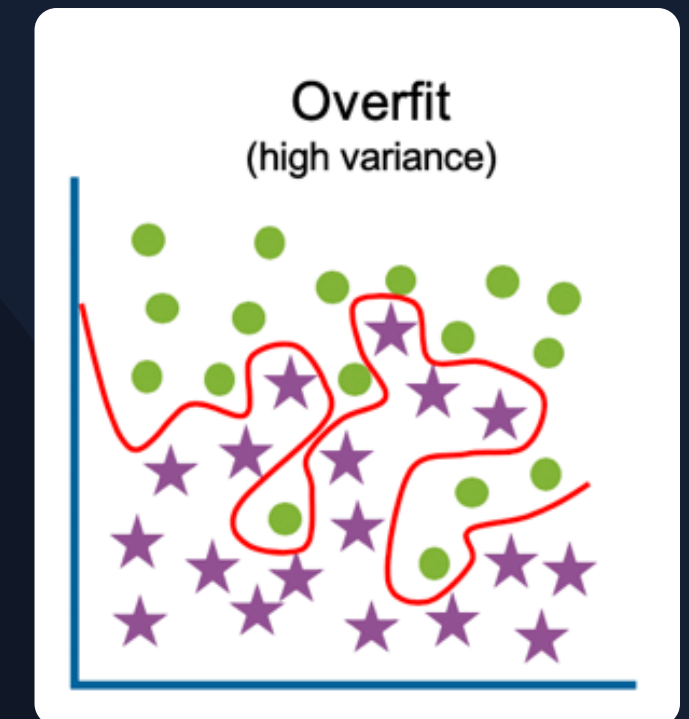
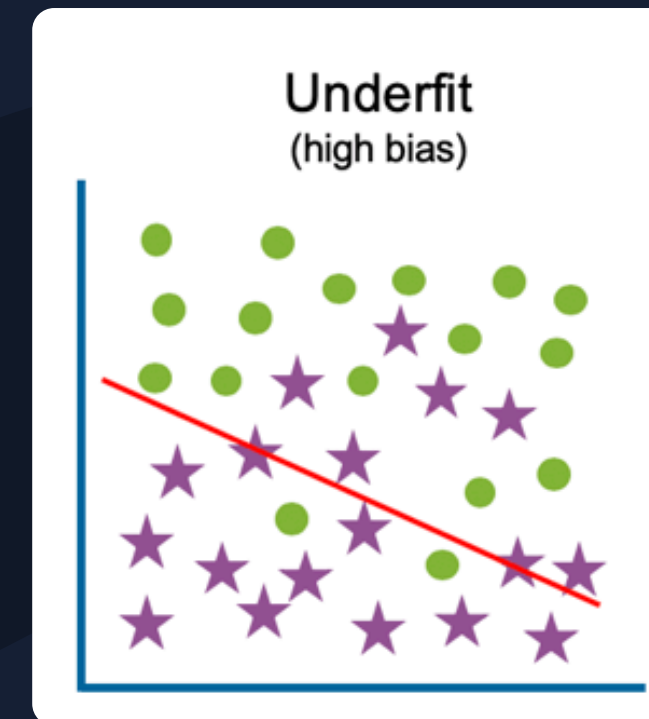


Image Source



Final Project

Supervised Learning Salary Prediction




s.id/salaryPredictionML




Supervised Learning Salary Prediction

This project develops a supervised machine learning model to predict salaries based on years of work experience using the salary_data.csv dataset, which contains 100 records and 3 columns (years of experience, salary, and an ID). It compares the performance of linear regression, decision tree, and random forest algorithms to identify the most effective approach for salary prediction.



	A	B	C
1	employee_id	experience_years	salary
2	EM_101	16.8	3166.9
3	EM_102	10.7	3126.9
4	EM_103	14.1	3278.8
5	EM_104	9.1	2828.8
6	EM_105	8.9	2728.7
7	EM_106	7.9	2762.6
8	EM_107	4.4	2142.6
9	EM_108	16.2	3214.5
10	EM_109	2	1518.9
11	EM_110	0	1049.7
12	EM_111	3.6	1867.9
13	EM_112	6.1	2390.7
14	EM_113	14.7	3405.8
15	EM_114	6.7	2449.8
16	EM_115	18.2	3158.5
17	EM_116	2.8	1212.5
18	EM_117	15.4	3257.5
19	EM_118	15.6	3217
20	EM_119	2.4	1522.7
21	EM_120	6.3	267
22	EM_121	11.1	3191



Linear Regression

The Linear Regression model is moderately effective, with R^2 scores of 0.77 (train) and 0.63 (test), but shows signs of underfitting as it struggles to capture data complexity, especially with outliers.

Mean Squared Error:

Train: 107699.85

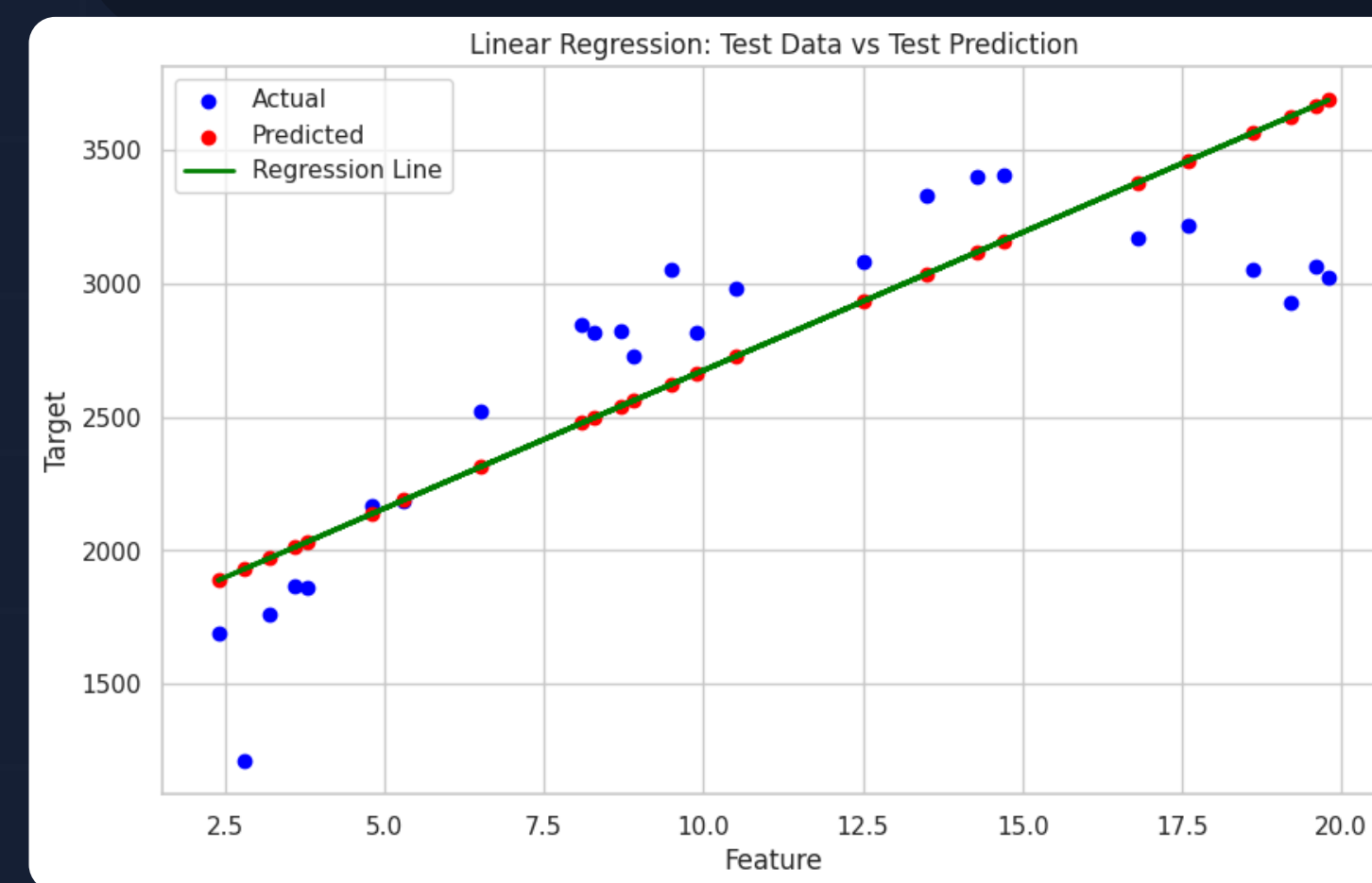
Test : 128111.12

Gap : 20411.27

R^2 Score:

Train: 0.77

Test : 0.63



Decission Tree

The Decision Tree model shows excellent performance on the training data with an R^2 score of 1.00, and a strong performance on the test data with an R^2 score of 0.93. However, the significant gap in Mean Squared Error (MSE) between training (88.12) and test (23,627.99) suggests that the model may be overfitting the training data.

Mean Squared Error:

Train: 88.12

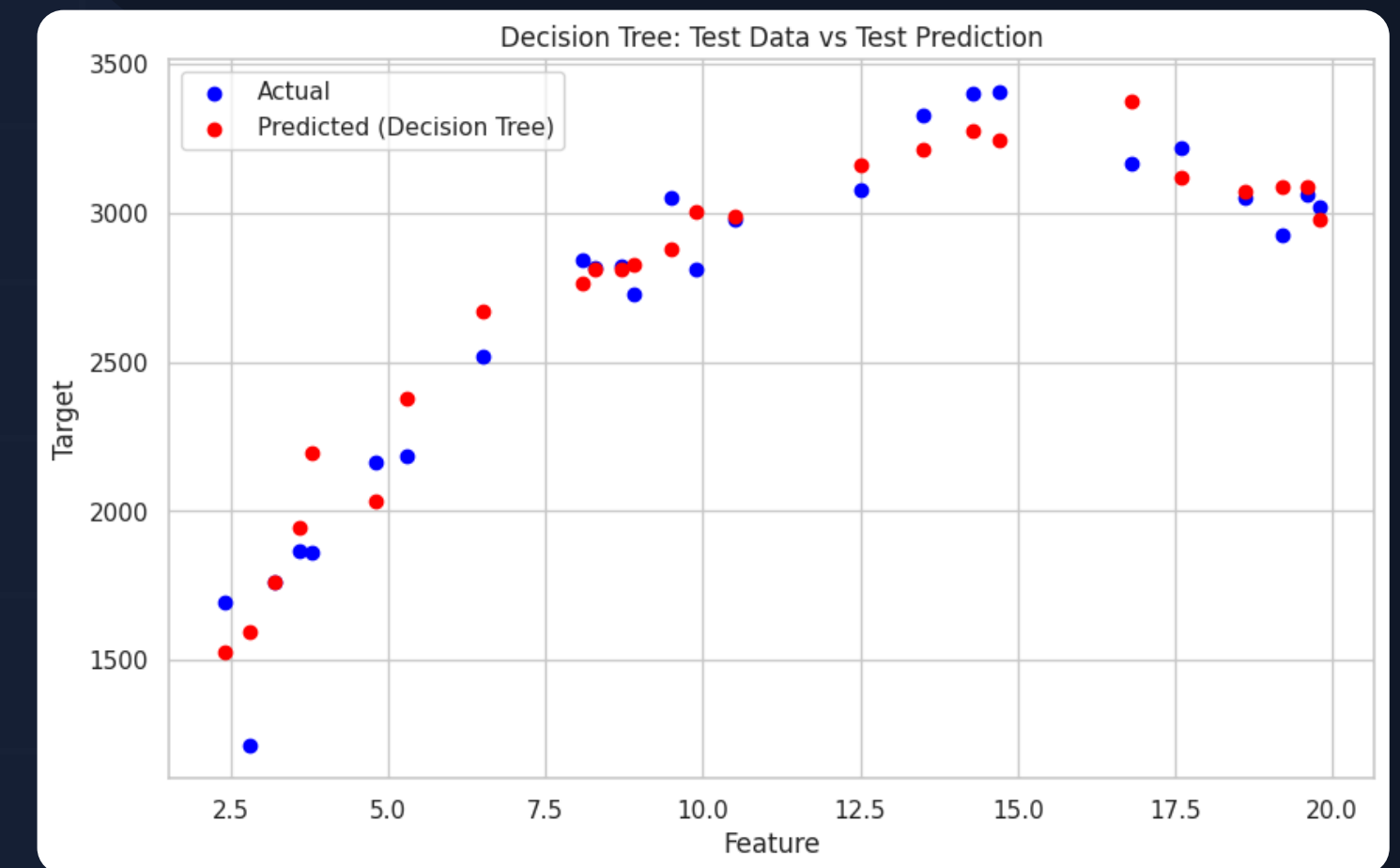
Test : 23627.99

Gap : 23539.87

R^2 Score:

Train: 1.00

Test : 0.93



Random Forest

The Random Forest model demonstrates strong performance with an R^2 score of 0.99 on the training data and 0.94 on the test data, indicating excellent predictive capability and minimal overfitting. While the Mean Squared Error (MSE) is higher on the test data (21,744.73) compared to the training data (3,737.44), the gap (18,007.29) suggests the model generalizes well to unseen data. This result indicates the Random Forest model effectively captures patterns in the data while maintaining robustness.

Mean Squared Error:

Train: 3737.44

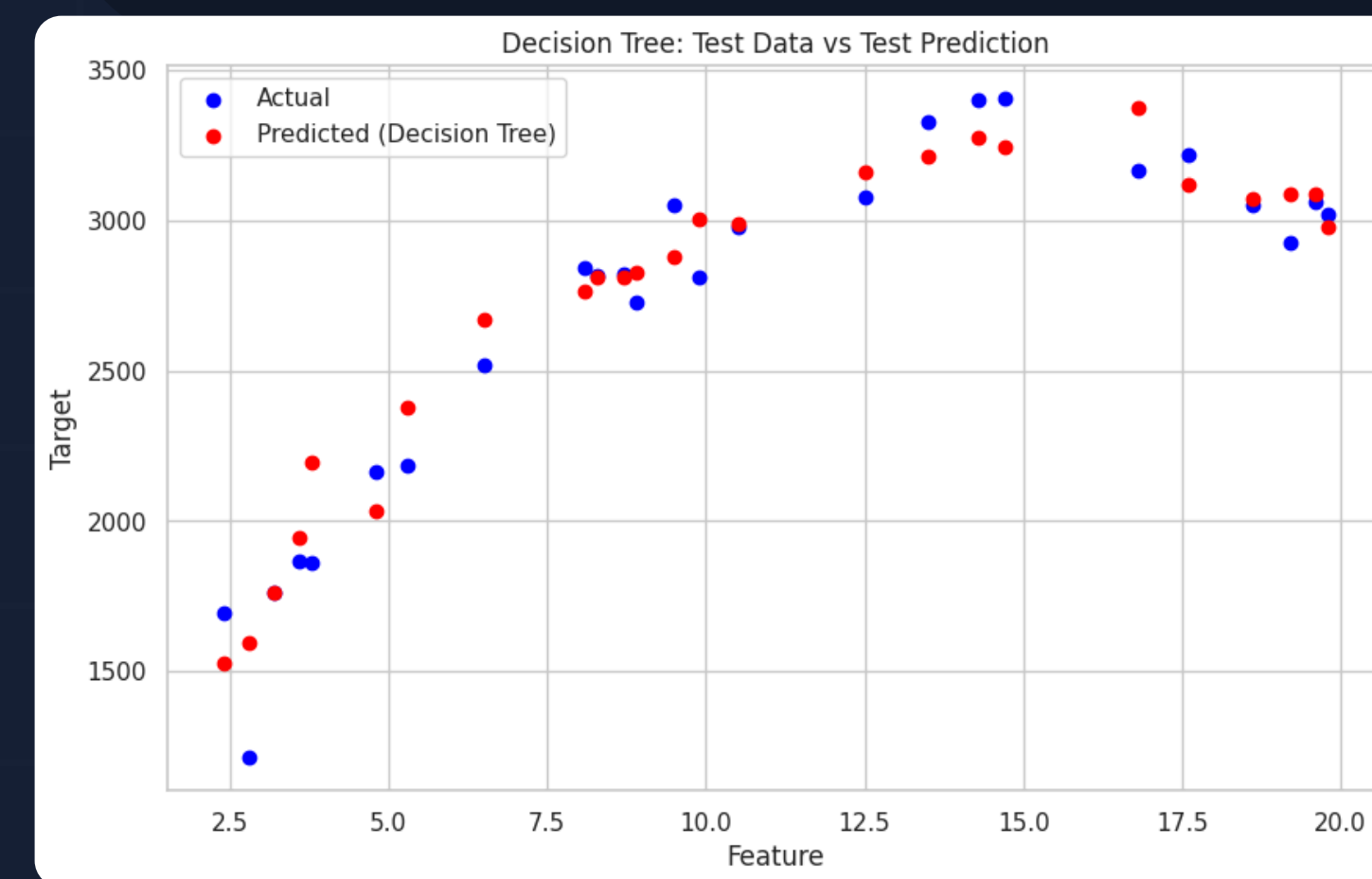
Test : 21744.73

Gap : 18007.29

R^2 Score:

Train: 0.99

Test : 0.94





Comparison

Linear Regression

Mean Squared Error:

Train: 107699.85

Test : 128111.12

Gap : 20411.27

R² Score:

Train: 0.77

Test : 0.63

Decision Tree

Mean Squared Error:

Train: 88.12

Test : 23627.99

Gap : 23539.87

R² Score:

Train: 1.00

Test : 0.93

Random Forest

Mean Squared Error:

Train: 3737.44

Test : 21744.73

Gap : 18007.29

R² Score:

Train: 0.99

Test : 0.94



Conclusion

The Random Forest model demonstrates strong performance with an R^2 score of 0.99 on the training data and 0.94 on the test data, indicating excellent predictive capability and minimal overfitting. While the Mean Squared Error (MSE) is higher on the test data (21,744.73) compared to the training data (3,737.44), the gap (18,007.29) suggests the model generalizes well to unseen data. This result indicates the Random Forest model effectively captures patterns in the data while maintaining robustness.



Thank You So Much

