DATA SERIES 17.0

ARTIFICIAL INTELLIGENCE MACHINE LEARNING

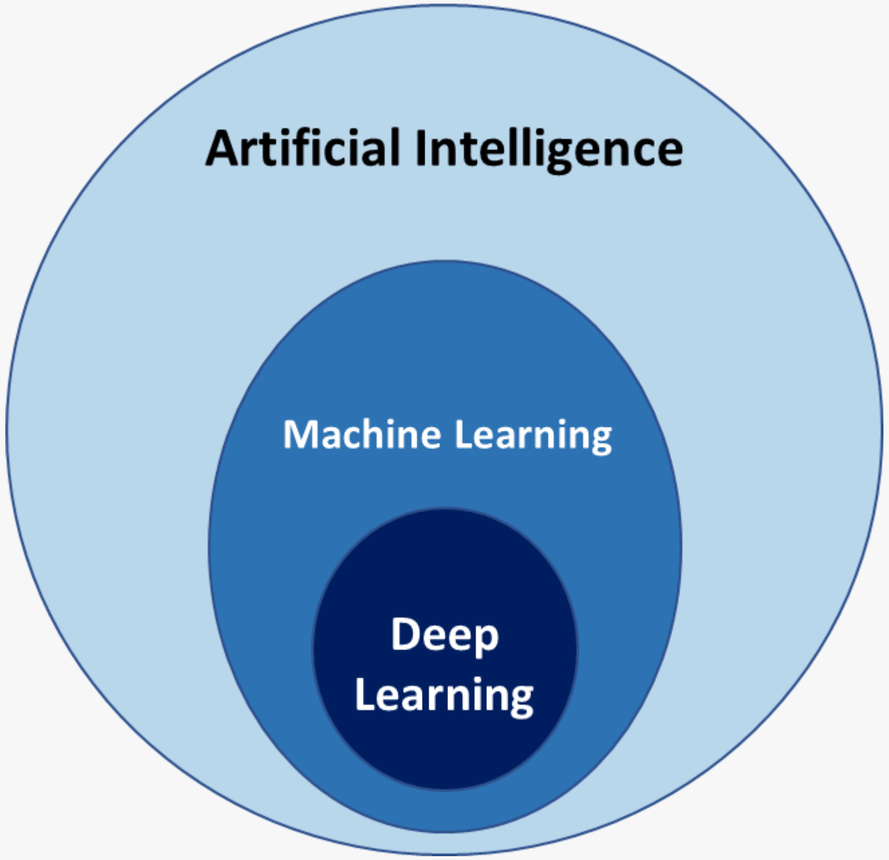
RAIHAN
PRATAMA

#DATASERIES17

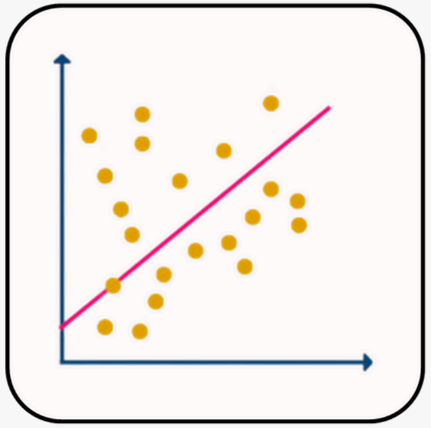
WHAT IS AI/ML?

ARTIFICIAL INTELLIGENCE (AI) is a technology that enables computers and machines to AI is a way for machines to ‘learn’ and ‘think’ like humans who can do understanding, problem solving, decision making and creativity.

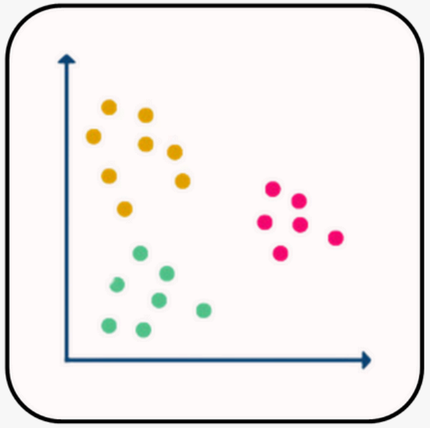
MACHINE LEARNING (ML) is a way for computers to ‘learn’ from experience (data) and make predictions, without having to be constantly instructed by humans. In other words, the more ‘training’ a computer receives, the smarter it becomes at recognizing patterns, making decisions, or predicting things.



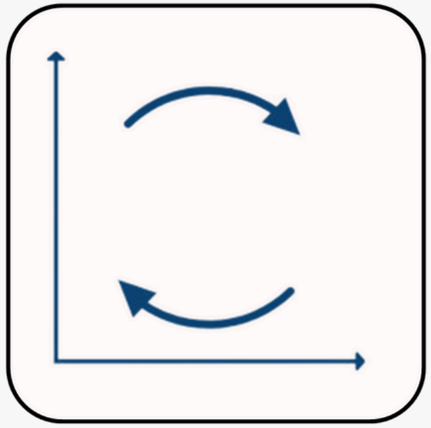
CATEGORY OF MACHINE LEARNING



Supervised Learning

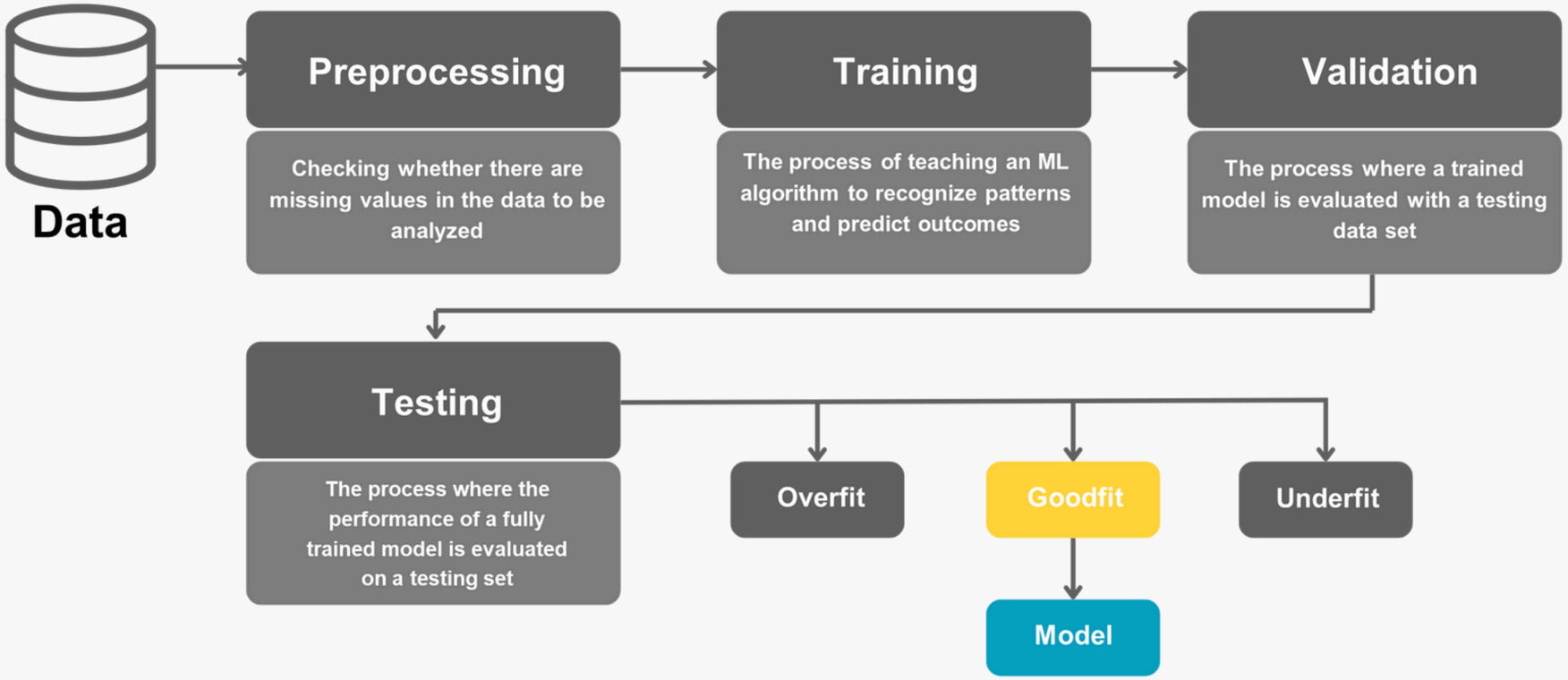


Unsupervised Learning



Reinforcement Learning

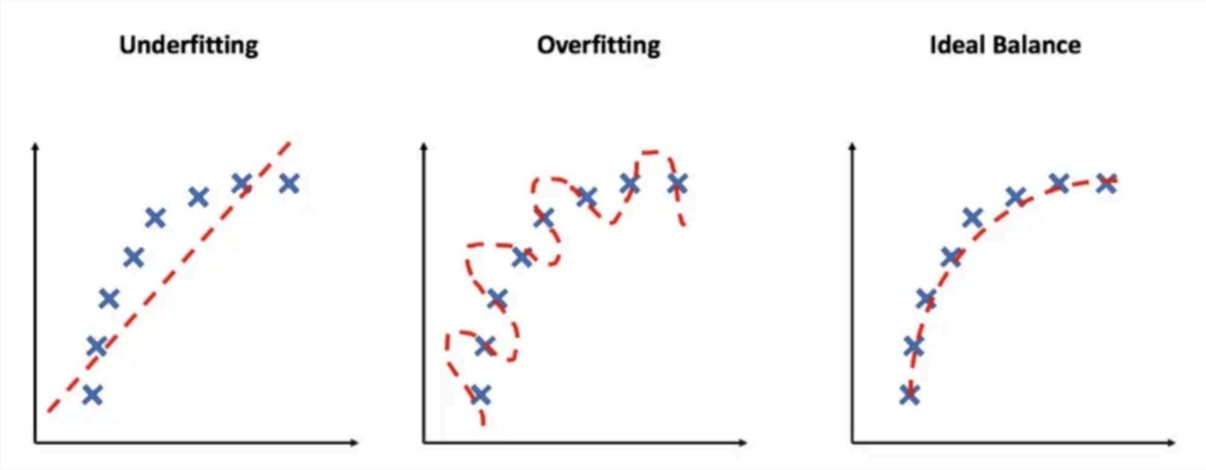
MACHINE LEARNING PROCESS



WHAT IS
OVERFITTING/ UNDERFITTING?

Overfitting is when a model becomes too focused on certain details and noise in its training data. As a result, the model may perform very well on the training set but fail to generalize when faced with new or previously unseen data.

Underfitting is the opposite scenario, where a model does not learn enough from the training data. It lacks the capacity to capture the essential patterns, leading to poor performance on both the training set and any new data.



SUPERVISED LEARNING
SALARY PREDICTION
FINAL PROJECT

#DATASERIES17

SUPERVISED LEARNING SALARY PREDICTION

Imagine a system that can estimate an individual's salary simply by looking at their years of professional experience. That's precisely what this supervised learning project aims to achieve. Leveraging a dataset with 100 records—containing an ID, years of experience, and salary—the study puts three different models to the test: Linear Regression, Decision Tree, and Random Forest. The objective is to pinpoint which model delivers the most accurate salary predictions.

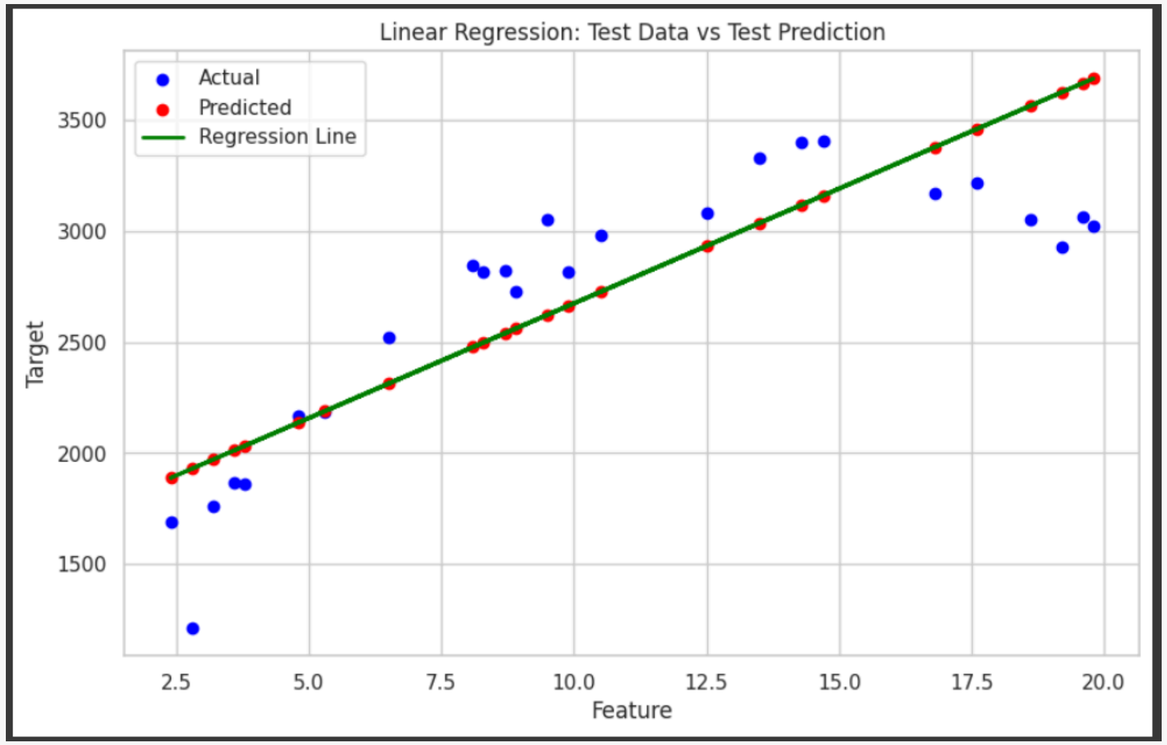
	A	B	C
1	employee_id	experience_years	salary
2	EM_101	16.8	3166.9
3	EM_102	10.7	3126.9
4	EM_103	14.1	3278.8
5	EM_104	9.1	2828.8
6	EM_105	8.9	2728.7
7	EM_106	7.9	2762.6
8	EM_107	4.4	2142.6
9	EM_108	16.2	3214.5
10	EM_109	2	1518.9
11	EM_110	0	1049.7
12	EM_111	3.6	1867.9
13	EM_112	6.1	2390.7
14	EM_113	14.7	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	1692.7
21	EM_120	6.3	2671.8
22	EM_121	11.1	3191.9

LINEAR REGRESSION

This model applies Linear Regression to predict salaries from years of experience. It achieves an R^2 of 0.77 on the training set and 0.63 on the test set, indicating a decent fit for the training data but somewhat limited predictive power for unseen observations.

Performance Metrics

- Mean Squared Error (MSE)
  - Training Data: 107,699.85
  - Test Data: 128,111.12
  - Difference: 20,411.27
- R^2 Score
  - Train: 0.77
  - Test: 0.63

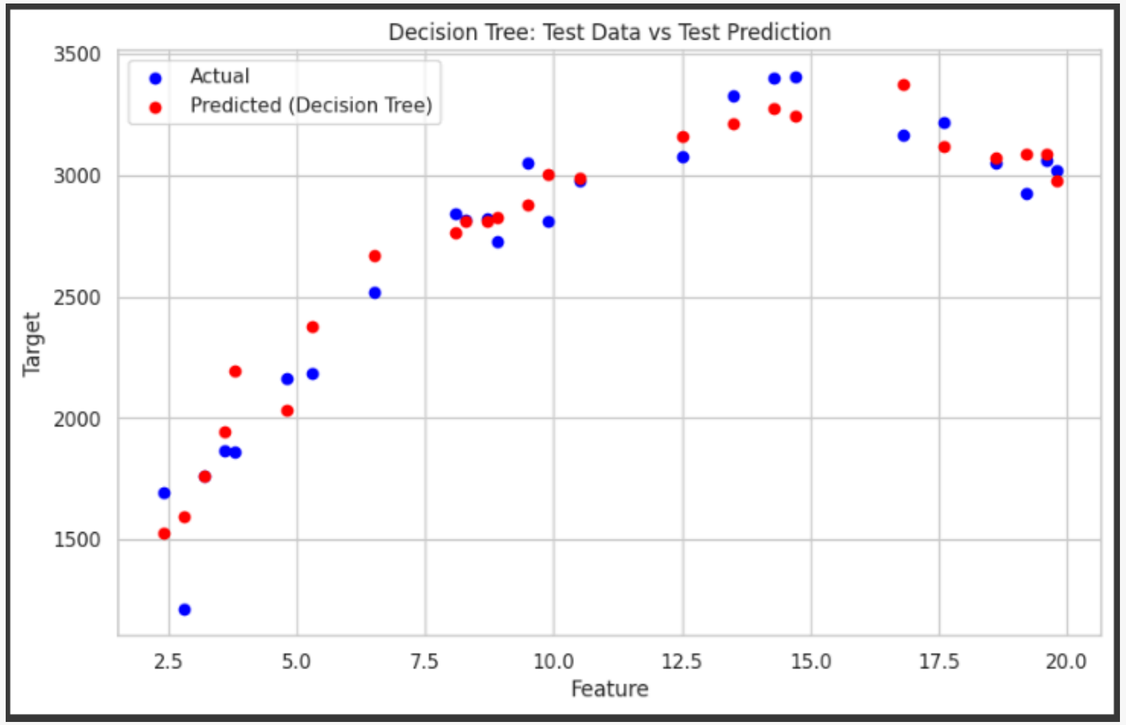


DECISION TREE

The Decision Tree model demonstrates excellent performance on the training data, achieving an R^2 score of 1.00. On the test data, it remains strong with an R^2 of 0.93. However, the large gap in Mean Squared Error (MSE) between the training and test sets suggests potential overfitting—the model may be overly tailored to the training data, which can reduce its predictive power when encountering new or unseen patterns.

Performance Metrics

- Mean Squared Error (MSE)
  - Training Data: 88.12
  - Test Data: 23,627.99
  - Gap: 23,539.87
- R^2 Score
  - Train: 1.00
  - Test: 0.93



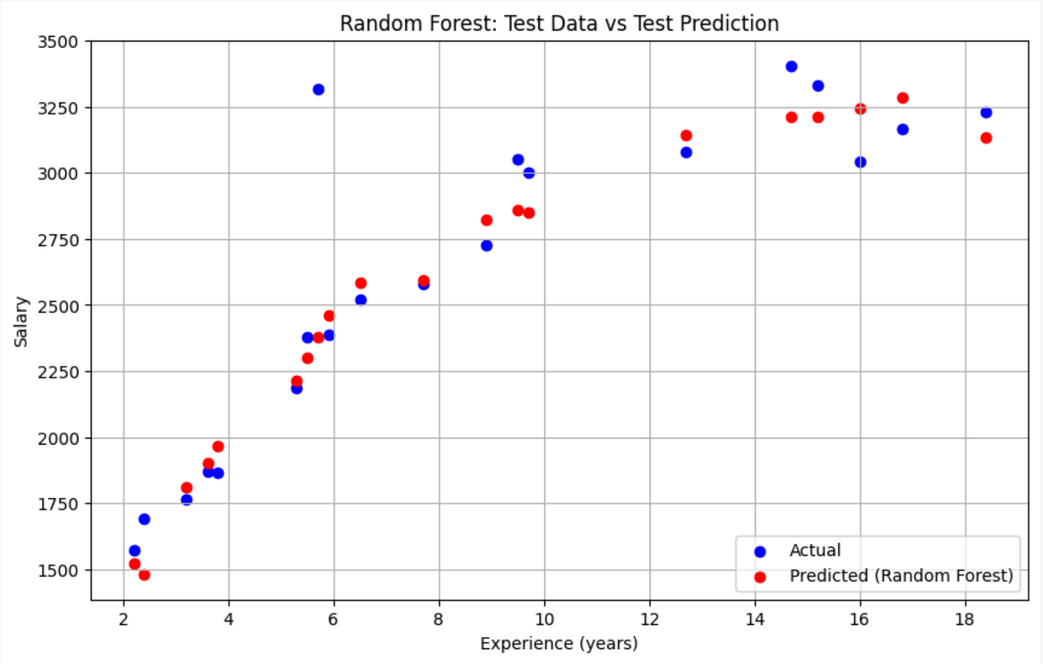
RANDOM FOREST

Random Forest leverages an ensemble of multiple Decision Trees to predict salaries based on years of experience. This approach often delivers more stable predictions than a single tree, as the collective decision-making from multiple trees helps offset individual errors.

Performance Metrics

- Mean Squared Error (MSE)
  - Train: 2,714.12
  - Test: 57,520.00
  - Gap: 54,805.88
- R^2 Score
  - Train: 0.89
  - Test: 0.84

While the R^2 scores show strong performance on both training and test data, the sizeable gap in MSE may hint at slight overfitting or the influence of outliers.



COMPARARISON

Linear Regresion

Mean Squared Error (MSE)
Train: 107.699,85
Test: 128.111,12
Gap: 20.411,27
R^2 Score
Train: 0,77
Test: 0,63

Decision Tree

Mean Squared Error (MSE)
Train: 88,12
Test: 23.627,99
Gap: 23.539,87
R^2 Score
Train: 1,00
Test: 0,93

Random Forest

Mean Squared Error (MSE)
Train: 3.737,44
Test: 21.744,73
Gap: 18.007,29
R^2 Score
Train: 0,99
Test: 0,94

CONCLUSION

The Random Forest model displays robust performance, achieving an R^2 of 0.99 on the training data and 0.94 on the test data. This high coefficient of determination indicates excellent predictive power and minimal overfitting. Although the Mean Squared Error (MSE) is higher on the test set (21,744.73) compared to the training set (3,737.44), the gap of 18,007.29 suggests the model still generalizes effectively to unseen data. These results highlight the Random Forest's capability to capture essential patterns in the dataset while maintaining strong overall stability.

THANK YOU