





































**1. Import Library**

Pertama, kita mengimpor beberapa library penting yang akan digunakan dalam proses analisis dan pemodelan data. Library yang digunakan adalah:

* pandas untuk manipulasi data.
* matplotlib.pyplot dan seaborn untuk visualisasi data.
* numpy untuk operasi matematika.

**2. Eksplorasi Data Awal (Exploratory Data Analysis)**

Kita memuat data dan melakukan analisis awal dengan melihat struktur data menggunakan data.info() yang menunjukkan bahwa data memiliki 2 kolom (YearsExperience dan Salary) dengan 30 entri yang semuanya memiliki nilai (non-null). Informasi statistik dasar dari data juga ditampilkan menggunakan data.describe() yang menunjukkan beberapa metrik statistik seperti rata-rata (mean), standar deviasi (std), nilai minimum (min), dan maksimum (max) dari masing-masing kolom.

**3. Visualisasi Data**

Menggunakan pairplot dari seaborn, kita memvisualisasikan hubungan antara variabel YearsExperience dan Salary dalam bentuk scatter plot. Namun, peringatan muncul karena parameter size telah diganti dengan height dalam versi seaborn yang lebih baru. Visualisasi ini menunjukkan penyebaran data sehingga kita bisa mengamati korelasi antara pengalaman kerja dan gaji.

**4. Feature Engineering**

Pada tahap ini, kita melakukan pemeriksaan lebih lanjut pada fitur-fitur dalam data untuk melihat apakah ada data duplikat, nilai yang hilang, atau outlier.

* **Pemeriksaan Data Duplikat:** Menggunakan df.duplicated(), kita menemukan bahwa tidak ada data duplikat dalam dataset ini, karena bentuk data tetap sama setelah proses penghapusan duplikat (df.drop\_duplicates()).
* **Pemeriksaan Missing Value (Nilai yang Hilang):** Menggunakan df.isna().sum(), diperoleh bahwa tidak ada nilai yang hilang pada kolom YearsExperience dan Salary.
* **Analisis Outlier:** Menggunakan boxplot untuk melihat apakah ada outlier pada kolom YearsExperience. Plot ini menunjukkan bahwa tidak ada nilai outlier yang signifikan dalam data.

Karena tidak ada masalah yang ditemukan pada fitur, kita bisa melanjutkan ke pemodelan tanpa penanganan lebih lanjut pada fitur.

**5. Pemisahan Data**

Data dibagi menjadi dua bagian:

* **X (Predictor):** kolom YearsExperience yang berfungsi sebagai prediktor.
* **y (Target):** kolom Salary yang merupakan target variabel.

Pembagian ini memungkinkan kita untuk melatih model pada data latih dan menguji model pada data uji.

**6. Model Linear Regression**

Model regresi linier dibentuk dengan persamaan:

y=25478.12+9731.016×xy = 25478.12 + 9731.016 \times xy=25478.12+9731.016×x

dengan x adalah jumlah tahun pengalaman kerja (YearsExperience). Persamaan ini menyatakan hubungan linear antara pengalaman kerja dan gaji.

**7. Model Decision Tree**

Selain regresi linier, kita juga menggunakan model lain yaitu **Decision Tree Regressor**. Model ini dilatih menggunakan DecisionTreeRegressor dari sklearn.tree. Setelah melatih model, prediksi gaji (y\_pred\_dt) dihasilkan dan dibandingkan dengan nilai aktual (y\_test). Visualisasi plot perbandingan antara nilai aktual dan prediksi menunjukkan seberapa akurat model Decision Tree dalam melakukan prediksi.

* **R-Square**: Menghitung nilai R-squared yang mengukur seberapa baik prediksi model sesuai dengan data aktual. Semakin tinggi nilai R-squared, semakin baik model dalam memprediksi data.

**8. Model Random Forest**

Selain Decision Tree, model **Random Forest Regressor** juga digunakan untuk membandingkan performa model prediksi. Random Forest adalah kombinasi dari beberapa decision tree untuk meningkatkan akurasi prediksi.