Tugas 4 - Analisa Regresi

> **Topic: Linear Regression, Multiple Linear Regression, and Polynomial Regression**

Seluruh dataset yang diperlukan ada di link berikut:

https://github.com/jagustinus/ml-class/tree/8b7e06d8b82b4a134b81709bacac0638931722c2/Tugas 4

Bagian 1: Analisis Fitur Menggunakan Regresi Linear Sederhana

Menggunakan california house pricing dataset secara keseluruhan dari tugas sebelumnya, lakukan hal-hal berikut:

1. Untuk setiap fitur dalam dataset, implementasikan regresi linear sederhana dengan variabel target (median_house_price).
2. Hitung dan catat nilai R-squared untuk setiap fitur.
3. Buat tabel ringkasan yang mengurutkan fitur berdasarkan nilai R-squared-nya.
4. Visualisasikan hubungan antara setiap fitur dan target menggunakan scatter plot.
5. Tulislah analisis singkat 1-2 paragraf yang menjelaskan fitur mana yang menunjukkan hubungan paling kuat dengan harga rumah dan jelaskan alasannya mengapa.

Bagian 2: Implementasi Manual Regresi Linear Berganda

Menggunakan dataset sampel yang disediakan (5 data point dengan 2 fitur dan 1 target):

1. Implementasikan regresi linear berganda TANPA menggunakan pustaka sklearn atau statsmodels.
2. Implementasi Anda harus mencakup:
    a. Pembuatan matriks fitur
    b. Implementasi persamaan normal (β = (X^T X)^(-1) X^T y)
    c. Fungsi prediksi
    d. Perhitungan R-squared
    e. Tunjukkan perhitungan langkah demi langkah Anda
3. Data Sampel:

Size (sq ft)	Age (years)	Price ($)
1500	10	300,000
1600	8	320,000
1700	5	340,000
1800	2	360,000
1900	1	380,000

Bagian 3: Regresi Linear Berganda & Polynomial (Bisa Berkelompok - 2 Orang)

Menggunakan dataset advertising.csv dari link diatas:

1. Menggunakan sklearn, implementasikan:
    a. Regresi linear berganda
    b. Regresi polinomial (hingga derajat 2-3)
2. Untuk setiap model:
    a. Bagi data menjadi data latih (70%) dan data pengujian (30%) menggunakan train_test_split
    b. Lakukan feature scaling menggunakan StandardScaler atau Standarization bila diperlukan
    c. Pasang model dan buat prediksi
    d. Hitung dan laporkan:
        i. R-squared untuk kedua set pelatihan dan pengujian
        ii. Mean Squared Error (MSE) untuk kedua set pelatihan dan pengujian
        iii. Root Mean Squared Error (RMSE) untuk kedua set pelatihan dan pengujian
3. Buat visualisasi:
    a. Feature correlation heatmap
    b. Plot scatter dari prediksi vs. nilai aktual
    c. Plot residual untuk kedua model
4. Analisis Model (maksimal 500 kata):
    a. Bandingkan metrik kinerja kedua model
    b. Diskusikan model mana yang lebih sesuai untuk dataset ini dan mengapa
    c. Analisis potensi masalah overfitting atau underfitting
    d. Sarankan kemungkinan perbaikan untuk model

e. Diskusikan dampak fitur polinomial pada kompleksitas dan kinerja model

Catatan

- Anda dapat menggunakan **numpy** untuk operasi matematika
- Anda dapat menggunakan **matplotlib/seaborn** untuk visualisasi
- Bagian 1 & 2: Tidak diperkenankan menggunakan **sklearn** untuk implementasi regresi
- Bagian 3: Anda diperkenankan menggunakan **sklearn** atau statsmodels
- Anda dapat menggunakan pandas untuk manipulasi data
- Untuk Bagian 1 & 2 → Hari Jumat 25 Oktober
- Untuk Bagian 3 → Hari Jumat 1 November