Tugas 3 - Linear Regression

Dataset

Kita akan menggunakan dataset California Housing Prices. Dataset ini berisi informasi tentang berbagai fitur rumah di California dan harganya sesuai dengan sensus California tahun 1990.

- California data - link
- Sample data - link

Untuk tugas pertama, kita akan menggunakan subset dari 200 sampel yang dipilih secara acak dari dataset tersebut melalui tautan diatas.

1. **Linear Regresi dengan OLS di Excel**

   Pada bagian ini, anda akan bekerja dengan menggunakan sebagian data sampel yang sudah di tentukan dan lakukan perhitungan manual dengan menggunakan Excel.

   Dari data sampel tersebut dengan jumlah data sebesar 200 data, terdapat 2 kolom yaitu:

   - **median_income**: Pendapatan median untuk rumah tangga dalam satu blok perumahan (diukur dalam puluhan ribu Dolar AS)
   - **medianHouse_value:** Nilai median rumah untuk rumah tangga dalam satu blok (diukur dalam Dolar AS)

   Tugas anda adalah:

   1. Hitung rata-rata dari “**median_income”** dan “**medianHouse_value**”.
   2. Hitung varians dari “**median_income”**.
   3. Hitung kovarians antara '**median_income**' dan '**medianHouse_value**'.
   4. Tentukan kemiringan / slope (β1) dan intercept (β0) dari garis regresi linear menggunakan rumus OLS:
      - $\beta_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
      - $\beta_0 = \bar{y} - \beta_1\bar{x}$
   5. Tuliskan persamaan regresi linearnya.
   6. Gunakan persamaan yang Anda peroleh untuk memprediksi nilai median rumah untuk lingkungan dengan nilai “**median_income”** sebesar 3.0147.
   7. Hitung nilai R-kuadrat untuk mengevaluasi kesesuaian model.

8. Berdasarkan hasil yang anda peroleh, tuliskan 2-3 kalimat mengenai hasil anda, yang meliputi:
    - Arti dari kemiringan (slope) dan titik potong (intercept) yang anda temukan terkait dengan dataset.
    - Apa yang nilai R-squared katakan kepada Anda tentang kinerja model Anda.

2. **Linear Regresi dengan Python**

   Pada bagian ini, Anda akan mengimplementasikan regresi linear menggunakan Python dan memilih fitur yang berbeda dari dataset California Housing Prices untuk memprediksi harga rumah. Gunakan seluruh data bukan sampel seperti pada nomor 1.

   Tugas anda adalah:

   1. Load dataset dari link yang diberikan.
   2. Pilih satu fitur dari beberapa kolom fitur yang tersedia selain dari "**median_income**" yang menurut anda memiliki hubungan terbaik dengan target. Berikan alasannya.
   3. Hitung / implementasikan:
      a. Mean
      b. Variance
      c. Covariance
      d. OLS Coefficients (β0 and β1)
      e. Hasil prediksi dengan 1 contoh data
      f. R-squared
   4. Gunakan fungsi fungsi diatas untuk membuat regresi linear anda fit dengan dataset.
   5. Plot data dan garis regresi linear dengan matplotlib.
   6. Print persamaan regresinya beserta nilai dari R-squared.

Pedoman Pengumpulan

1. Kumpulkan spreadsheet Excel Anda yang menunjukkan semua perhitungan untuk Bagian 1.
2. Kumpulkan kode Python Anda untuk Bagian 2.

3. Sertakan laporan yang mencakup:
    - Interpretasi Anda dari Bagian 1
    - Proses pemilihan fitur dan alasan Anda untuk Bagian 2
    - Perbandingan hasil antara Bagian 1 dan 2
    - Jelaskan tentang tantangan yang dihadapi selama mengerjakan dan bagaimana Anda mengatasinya

Semoga sukses!