Nama : Annisa Charisma Wijayanti
NIM : 211220122140086
Mata Kuliah : Metode Numerik – Kelas D
Jurusan : Teknik Komputer
Link Github : https://github.com/annisacharisma/Implementasi-Aplikasi-Regresi_Annisa-Charisma_21120122140086
Link Colab : https://colab.research.google.com/drive/1-AAmqFE-8dsPNAH_N1tMU4_4AXdTmLkO?usp=sharing

# APLIKASI REGRESI UNTUK PEMECAHAN PROBLEM

Diinginkan untuk mencari hubungan faktor yang mempengaruhi nilai ujian siswa (NT):

1. Durasi waktu belajar (TB) terhadap nilai ujian (Problem 1)
2. Jumlah latihan soal (NL) terhadap nilai ujian (Problem 2)

Data TB, NL, dan NT diperoleh dari https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression, yaitu kolom Hours Studied, Sample Question Papers Practiced, dan Performance Index.

Implementasikan regresi untuk mencari hubungan tersebut menggunakan metode:

1. Model linear (Metode 1)
2. Model pangkat sederhana (Metode 2)
3. Model eksponensial (Metode 3)
4. Model lainnnya di halaman 24 slide materi (Metode opsional)

Tugas mahasiswa:

1. Mahasiswa membuat kode sumber dengan bahasa pemrograman yang dikuasai untuk mengimplementasikan solusi di atas, dengan ketentuan:
   - NIM terakhir % 4 = 0 mengerjakan Problem 1 dengan Metode 1 dan Metode 2
   - NIM terakhir % 4 = 1 mengerjakan Problem 1 dengan Metode 1 dan Metode 3
   - NIM terakhir % 4 = 2 mengerjakan Problem 2 dengan Metode 1 dan Metode 2
   - NIM terakhir % 4 = 3 mengerjakan Problem 2 dengan Metode 1 dan Metode 3
   - Mahasiswa juga bisa menambah solusi dengan salah satu metode opsional
2. Sertakan kode testing untuk menguji kode sumber tersebut untuk menyelesaikan problem dalam gambar. Plot grafik titik data dan hasil regresinya masing-masing
3. Hitung galat RMS dari tiap metode yang digunakan

4. Mengunggah kode sumber tersebut ke Github dan setel sebagai publik. Berikan deskripsi yang memadai dari project tersebut. Masukkan juga dataset dan data hasil di repositori tersebut.
5. Buat dokumen docx dan pdf yang menjelaskan alur kode dari (1), analisis hasil, dan penjabarannya. Sistematika dokumen: Ringkasan, Konsep, Implementasi Kode, Hasil Pengujian, dan Analisis Hasil.

Berdasarkan ketentuan tugas tersebut, dengan nim terakhir saya yaitu 6, maka akan mengerjakan Problem 2 dengan Metode 1 dan Metode 2. Yaitu problem jumlah latihan soal (NL) terhadap nilai ujian dengan menggunakan metode linear dan metode pangkat sederhana.

## METODE REGRESI LINEAR

Regresi linear adalah teknik analisis data yang digunakan untuk memprediksi nilai dari suatu variabel yang tidak diketahui (variabel dependen) berdasarkan nilai variabel lain yang diketahui (variabel independen). Teknik ini secara matematis memodelkan hubungan antara variabel dependen dan variabel independen dalam bentuk persamaan linier. Model regresi linier bersifat sederhana dan memberikan rumus matematika yang mudah diinterpretasikan untuk menghasilkan prediksi.

Secara matematis, persamaan regresi linier sederhana dapat dinyatakan sebagai:

$$y_i = \beta_0 + \beta_i x_i + \epsilon$$

dengan:

$y_i$ = nilai variabel independen.

$\beta_0$ = intercept (nilai *y* saat *x*=0).

$\beta_i$ = koefisien regresi (mengukur seberapa besar perubahan *y* untuk setiap perubahan satu unit *x*).

$x_i$ = nilai variabel independen.

$\epsilon$ = nilai error term

Atau bisa juga ditulis dengan menggunakan persamaaan:

$$Y = a + bX$$

# METODE PANGKAT SEDERHANA

Metode regresi pangkat sederhana adalah teknik regresi non-linier yang memodelkan hubungan antara variabel dependen dan variabel independen dengan menggunakan fungsi pangkat. Metode ini berguna ketika hubungan antara variabel tidak dapat dijelaskan dengan baik oleh metode linier sederhana. Bentuk umum dari persamaan regresi pangkat sederhana adalah:

$$y = ax^b$$

dengan:

- *y* adalah variabel dependen.
- *x* adalah variabel independen.
- *a* adalah koefisien skala.
- *b* adalah eksponen atau pangkat.

Untuk memudahkan estimasi parameter *a* dan *b*, persamaan tersebut dapat diubah menjadi bentuk linier dengan mengambil logaritma dari kedua sisi:

$$log(y) = log(a) + b\ log(x)$$

Setelah persamaan diubah menjadi bentuk linier, regresi linier biasa dapat digunakan untuk memperkirakan parameter-parameter tersebut.

**Konsep Kode Implementasi Aplikasi Regresi:**

- **Input Data dan Membaca Data**: Kode akan membaca data dari file CSV, memilih kolom yang relevan, dan membersihkan data dari nilai-nilai yang tidak valid.
- **Membangun dan Melatih Metode**: Ada dua metode yang dibangun dan dilatih pada data yang telah dibersihkan yaitu:
  - Metode regresi linier untuk memprediksi indeks performa berdasarkan jumlah latihan soal.
  - Metode regresi pangkat sederhana yang mengubah data menggunakan logaritma untuk memprediksi hubungan non-linier.
- **Evaluasi Metode**: Kedua metode dievaluasi dengan menghitung Root Mean Squared Error (RMS), yang mengukur seberapa baik metode memprediksi data asli.
- **Visualisasi Hasil**: Hasil prediksi dari kedua metode divisualisasikan dengan plot scatter, yang menunjukkan data asli serta garis regresi dari masing-masing metode.

- **Perbandingan Metode**: Menampilkan nilai RMS untuk kedua metode, sehingga kita bisa membandingkan performa kedua metode dan memilih metode yang lebih baik berdasarkan nilai RMS yang lebih rendah.

**Source Code:**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

# input dataset
data = pd.read_csv('Student_Performance.csv')

# Mengambil column NL dan NT
NL = data['Sample Question Papers Practiced'].values
NT = data['Performance Index'].values

# Periksa dan buang data yang mengandung NaN atau Inf
valid_idx = ~np.isnan(NL) & ~np.isnan(NT) & ~np.isinf(NL) & ~np.isinf(NT)
& (NL > 0)
NL = NL[valid_idx].reshape(-1, 1)
NT = NT[valid_idx]

# Metode 1: Model Linear
linear_model = LinearRegression()
linear_model.fit(NL, NT)
NT_pred_linear = linear_model.predict(NL)

# Metode 2: Model Pangkat Sederhana
# y = ax^b => log(y) = log(a) + b*log(x)
NL_log = np.log(NL)
NT_log = np.log(NT)
linear_model_pangkat = LinearRegression()
linear_model_pangkat.fit(NL_log, NT_log)
a_pangkat = np.exp(linear_model_pangkat.intercept_)
b_pangkat = linear_model_pangkat.coef_[0]
NT_pred_pangkat = a_pangkat * NL**b_pangkat

# Menghitung galat RMS
rms_linear = np.sqrt(mean_squared_error(NT, NT_pred_linear))
rms_pangkat = np.sqrt(mean_squared_error(NT, NT_pred_pangkat))

# Plot hasil regresi
plt.figure(figsize=(12, 6))

plt.subplot(1, 2, 1)
plt.scatter(NL, NT, color='blue', label='Data Asli')
plt.plot(NL, NT_pred_linear, color='red', label='Regresi Linear')
plt.xlabel('Jumlah Latihan Soal (NL)')
plt.ylabel('Nilai Ujian (NT)')
plt.title('Regresi Linear')
plt.legend()

plt.subplot(1, 2, 2)
plt.scatter(NL, NT, color='blue', label='Data Asli')
plt.plot(NL, NT_pred_pangkat, color='green', label='Regresi Pangkat')
```

```
plt.xlabel('Jumlah Latihan Soal (NL)')
plt.ylabel('Nilai Ujian (NT)')
plt.title('Regresi Pangkat Sederhana')
plt.legend()

plt.show()

print(f"RMS Metode Linear: {rms_linear}")
print(f"RMS Metode Pangkat Sederhana: {rms_pangkat}")
```

**Analisis Kode:**

Kode tersebut digunakan untuk melakukan analisis regresi pada data yang berisi jumlah latihan soal yang dikerjakan oleh siswa (NL) dan indeks performa mereka (NT). Tujuan utamanya adalah untuk membangun model prediksi yang dapat memprediksi indeks performa siswa berdasarkan jumlah latihan soal yang mereka kerjakan. Untuk alur kode tersebut yaitu:

- Import library yang diperlukan untuk analisis data, regresi linear, dan visualisasinya.
- Membaca atau input dataset menggunakan kode `data = pd.read_csv('Student_Performance.csv')`
- Mengambil kolom 'Sample Question Papers Practiced' sebagai variabel independen *X* dan 'Performance Index' sebagai variabel dependen *Y* dengan kode:

```
NL = data['Sample Question Papers Practiced'].values
NT = data['Performance Index'].values
```

- Menghapus data yang mengandung nilai NaN atau Inf dan memastikan bahwa nilai *X* positif. Data kemudian diubah menjadi bentuk yang sesuai untuk regresi.

```
valid_idx = ~np.isnan(NL) & ~np.isnan(NT) & ~np.isinf(NL) & ~np.isinf(NT) & (NL > 0)
NL = NL[valid_idx].reshape(-1, 1)
NT = NT[valid_idx]
```

- Selanjutnya adalah membangun dan melatih metode regresi linear. Pada langkah ini, sebuah model regresi linier dibangun dan dilatih menggunakan kelas `LinearRegression`. Dengan menggunakan metode `fit()`, model disesuaikan dengan data latih, yang terdiri dari variabel independen `NL` (jumlah latihan soal) dan variabel dependen `NT` (indeks performa). Setelah pelatihan, metode `predict()` digunakan untuk membuat prediksi nilai `NT` berdasarkan data `NL`, dan hasil prediksi disimpan dalam variabel `NT_pred_linear`.

```
linear_model = LinearRegression()
linear_model.fit(NL, NT)
NT_pred_linear = linear_model.predict(NL)
```

- Metode regresi pangkat sederhana dibangun dengan mengasumsikan hubungan non-linier berbentuk $y = ax^b$. Untuk memudahkan estimasi parameter *a* dan *b*, data diubah menjadi bentuk linier dengan mengambil logaritma dari kedua variabel independen dan dependen, yaitu `NL` dan `NT`. Setelah itu, model regresi linier diterapkan pada data yang telah

ditransformasi ini menggunakan kelas `LinearRegression`. Parameter *a* dan *b* diperoleh dari model tersebut, dan hasil prediksi `NT` dalam bentuk regresi pangkat dihitung menggunakan parameter-parameter ini.

```
NL_log = np.log(NL)
NT_log = np.log(NT)
linear_model_pangkat = LinearRegression()
linear_model_pangkat.fit(NL_log, NT_log)
a_pangkat = np.exp(linear_model_pangkat.intercept_)
b_pangkat = linear_model_pangkat.coef_[0]
NT_pred_pangkat = a_pangkat * NL**b_pangkat
```

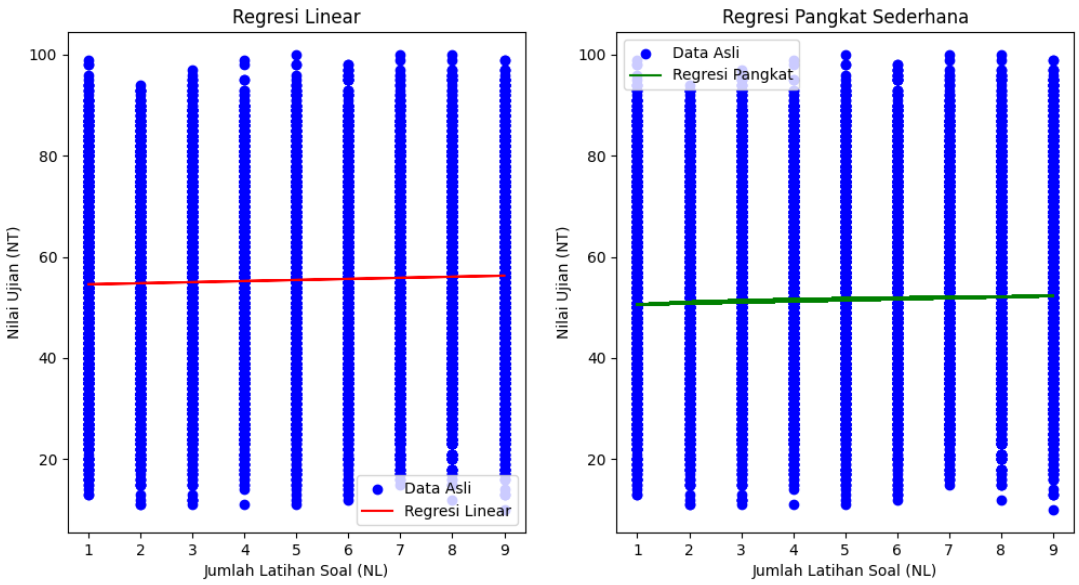
- Menghitung Root Mean Squared Error (RMS) untuk kedua model sebagai ukuran performa prediksi menggunakan kode:

```
rms_linear = np.sqrt(mean_squared_error(NT, NT_pred_linear))
rms_pangkat = np.sqrt(mean_squared_error(NT, NT_pred_pangkat))
```

- Selanjutnya adalah membuat plot untuk membandingkan hasil prediksi kedua model dengan data asli. Plot tersebut akan ditampilkan menggunakan kode `plt.show()`
- Terakhir, menampilkan hasil Galat RMS dari kedua metode untuk menilai performanya.

```
print(f"RMS Metode Linear: {rms_linear}")
print(f"RMS Metode Pangkat Sederhana: {rms_pangkat}")
```

**Hasil:**




```
[34] print(f"RMS Metode Linear: {rms_linear}")
     print(f"RMS Metode Pangkat Sederhana: {rms_pangkat}")

RMS Metode Linear: 19.196567736370024
RMS Metode Pangkat Sederhana: 19.565035257712687
```

**Analisis Hasil:**

Terdapat dua metode yang dibangun dan dibandingkan dalam kode ini:

1. **Metode Regresi Linier**. Metode ini mengasumsikan hubungan linier antara variabel independen (jumlah latihan soal) dan variabel dependen (indeks performa). Metode ini memberikan garis lurus yang paling sesuai dengan data.
2. **Metode Regresi Pangkat Sederhana**. Metode ini mengasumsikan hubungan non-linier berbentuk y=axby = ax^by=axb antara variabel independen dan dependen. Untuk mempermudah estimasi, data diubah menggunakan logaritma, sehingga regresi linier dapat diterapkan pada data yang telah ditransformasi.

Grafik sebelah kiri yaitu metode regresi linear menunjukkan garis merah yang hampir datar di tengah data asli yang diwakili oleh titik-titik biru. Hal tersebut menunjukkan bahwa model linear tidak mampu menangkap variasi dalam data dengan baik. Berdasarkan nilai RMS yang masih cukup tinggi, menunjukkan bahwa rata-rata kesalahan kuadrat akar dari prediksi model linear adalah sekitar 19.20 dari nilai sebenarnya. Garis regresi yang hampir datar menunjukkan bahwa model linear tidak efektif dalam menjelaskan variasi data, mungkin karena hubungan antara jumlah latihan soal dan nilai ujian tidak bersifat linear.

Grafik sebelah kanan yaitu metode regresi pangkat sederhana juga menunjukkan garis hijau yang hampir datar di tengah data asli. Hasil ini serupa dengan model linear, dengan nilai RMS sebesar 19.57, sedikit lebih tinggi dibandingkan model linear. Garis regresi yang datar menunjukkan bahwa model pangkat sederhana juga tidak mampu menangkap hubungan yang lebih kompleks dalam data.

Berdasarkan hasil tersebut, kedua metode menunjukkan kinerja yang kurang memuaskan dalam memprediksi nilai ujian berdasarkan jumlah latihan soal. Hal ini terlihat dari garis regresi yang datar dan nilai RMS yang relatif tinggi, menunjukkan bahwa mungkin ada faktor-faktor lain yang lebih berpengaruh terhadap nilai ujian siswa, atau hubungan antara jumlah latihan soal dan nilai ujian bersifat lebih kompleks daripada yang dapat ditangkap oleh model linear dan pangkat sederhana.