**TUGAS MATA KULIAH**

**VISUALISASI DATA**



**DOSEN:**

Muhammad Ali Akbar, S.Si, M.Kom

**DISUSUN OLEH:**

Muhammad Rafli

41822120005

**PROGRAM STUDI SISTEM INFORMASI**

**FAKULTAS ILMU KOMPUTER**

**Soal 1**

Berikut library untuk memuat dan menganalisa data yang saya gunakan:

1. **Pandas**: Library utama untuk manipulasi dan analisis data

**Keuntungan:**

* 1. Struktur data yang intuitif (DataFrame dan Series)
  2. Integrasi mudah dengan library lain
  3. Performa baik untuk dataset medium

**Kerugian:**

1. Konsumsi memori tinggi untuk dataset besar
2. Tidak optimal untuk data yang sangat besar (>1GB)
3. **NumPy**: Library untuk operasi numerik

**Keuntungan:**

* 1. Operasi matematika yang sangat cepat
  2. Efisien dalam penggunaan memori
  3. Dasar untuk hampir semua library scientific Python

**Kerugian:**

1. Hanya untuk data numerik
2. Sintaks lebih kompleks untuk operasi data tabular
3. **Matplotlib/Seaborn**: Library visualisasi

**Keuntungan:**

* 1. Kontrol penuh atas visualisasi
  2. Integrasi baik dengan pandas
  3. Banyak jenis plot tersedia

**Kerugian:**

1. Sintaks yang verbose (terutama matplotlib)
2. Kurva pembelajaran yang curam

**Soal 2**

# Import library yang diperlukan

import pandas as pd

**# 1. MEMUAT DATA**

# Memuat data dari file CSV

df = pd.read\_csv('nama\_file.csv')

**Soal 3**

Exploratory Data Analysis adalah pendekatan analisis data yang menekankan pada pemahaman mendalam terhadap struktur, pola, dan karakteristik data melalui teknik visualisasi dan statistik deskriptif sebelum melakukan pemodelan formal. EDA bertujuan untuk "mengenal" data secara intuitif dan menemukan insight yang mungkin tidak terlihat dari analisis statistik konvensional.

Filosofi dasar EDA yang dikembangkan oleh John Tukey berlandaskan pada beberapa prinsip:

* **"Let the data speak"** - Membiarkan data menceritakan kisahnya sendiri tanpa asumsi yang terlalu kaku
* **Pendekatan detective work** - Memperlakukan analisis data seperti investigasi detektif yang mencari petunjuk dan pola tersembunyi
* **Skeptisisme yang sehat** - Mempertanyakan asumsi dan tidak langsung menerima hasil pada nilai nominalnya
* **Visualisasi sebagai bahasa** - Menggunakan grafik dan plot sebagai cara utama untuk berkomunikasi dengan data
* **Iterative process** - Proses berulang antara eksplorasi, hipotesis, dan validasi

**Soal 4**

1. **Univariate Analysis** pemeriksaan karakteristik satu variabel secara individual tanpa mempertimbangkan hubungannya dengan variabel lain. Fokus memahami distribusi, central tendency, variabilitas, dan outliers dari satu variabel. Contoh: melihat distribusi umur pelanggan, frekuensi kategori produk, atau statistik deskriptif pendapatan.
2. **Bivariate Analysis** mengeksplorasi hubungan antara dua variabel secara bersamaan. Tujuannya adalah memahami korelasi, asosiasi, atau dependensi antara kedua variabel tersebut. Contoh: hubungan antara tinggi dan berat badan, korelasi antara harga dan kualitas produk, atau perbandingan penjualan antar dua wilayah.
3. **Multivariate Analysis** melibatkan tiga atau lebih variabel secara simultan untuk memahami interaksi kompleks dan pola yang muncul dari kombinasi multiple variabel. Analisis ini mengungkap struktur data yang lebih kompleks dan hubungan yang tidak terlihat dalam analisis univariat atau bivariat. Contoh: analisis cluster untuk segmentasi pelanggan, principal component analysis, atau model prediktif dengan multiple predictors.

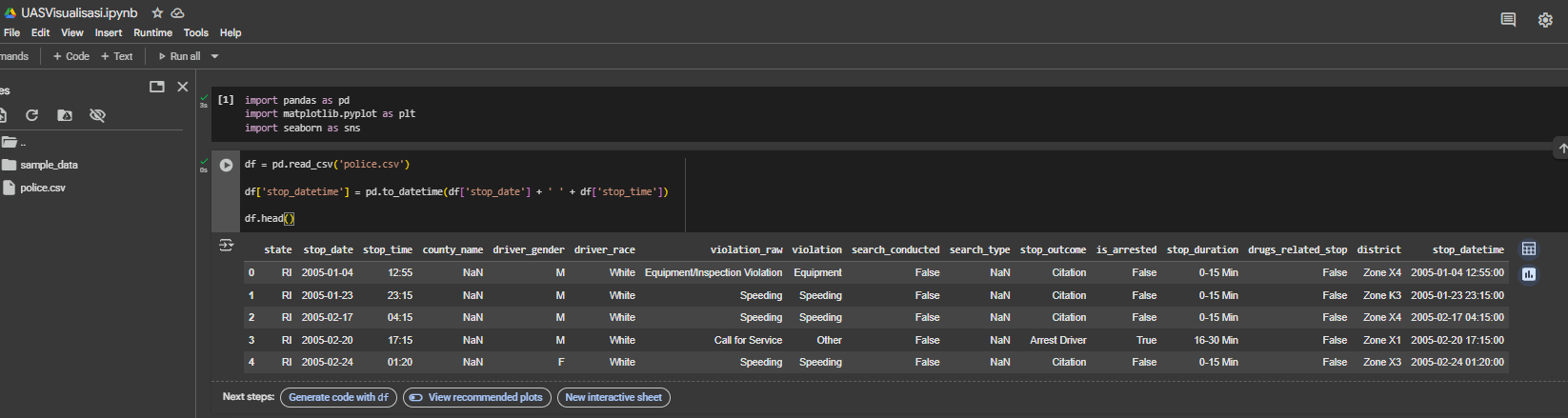
**Soal 5**

**Prinsip-Prinsip Dasar Pemilihan Warna:**

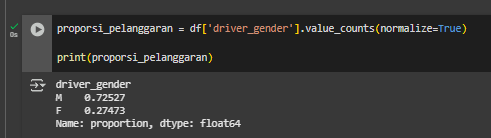
1. **Sequential Colors (Warna Berurutan)** digunakan untuk data ordinal atau numerik yang memiliki urutan dari rendah ke tinggi. Menggunakan gradasi satu atau dua warna dengan intensitas yang bervariasi. Contoh: dari biru muda ke biru tua untuk menunjukkan tingkat intensitas.
2. **Diverging Colors (Warna Divergen)** ideal untuk data yang memiliki titik tengah penting atau nilai netral. Menggunakan dua warna kontras yang bertemu di titik tengah netral. Contoh: merah-putih-biru untuk menunjukkan perubahan positif-netral-negatif.
3. **Categorical Colors (Warna Kategorikal)** untuk membedakan kategori yang tidak memiliki urutan natural. Menggunakan warna-warna yang kontras dan mudah dibedakan. Prinsip utama adalah memastikan setiap kategori memiliki warna yang distinct.

**Soal Essay**

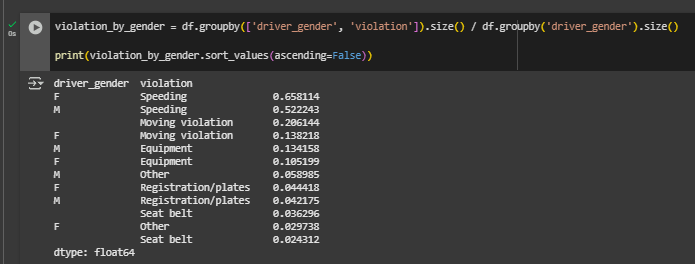
1. Soal No 1 ini menggunakan data police.csv. Data ini adalah data pemberhentian mobil untuk dilakukan pemeriksaan lebih lanjut oleh kepolisian. Apakah terjadi pelanggaran baik pelanggaran berat maupun pelanggaran ringan. Kalian diminta oleh pihak kepolisian USA untuk menganalisa potensi-potensi pelanggaran yang terjadi sampai terjadi kejadian penangkapan oleh pihak kepolisian dikarenakan terjadi pelangaran berat oleh pengendara mobil.
   1. Menampilkan Time Series



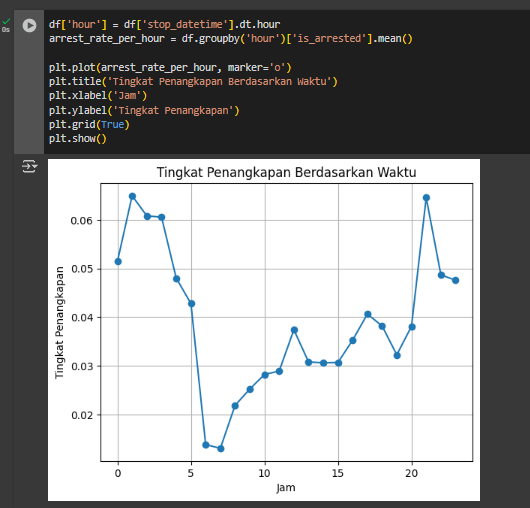
* 1. Lakukan pengecekan jenis kelamin mana yang paling banyak melakukan pelanggaran lalu lintas



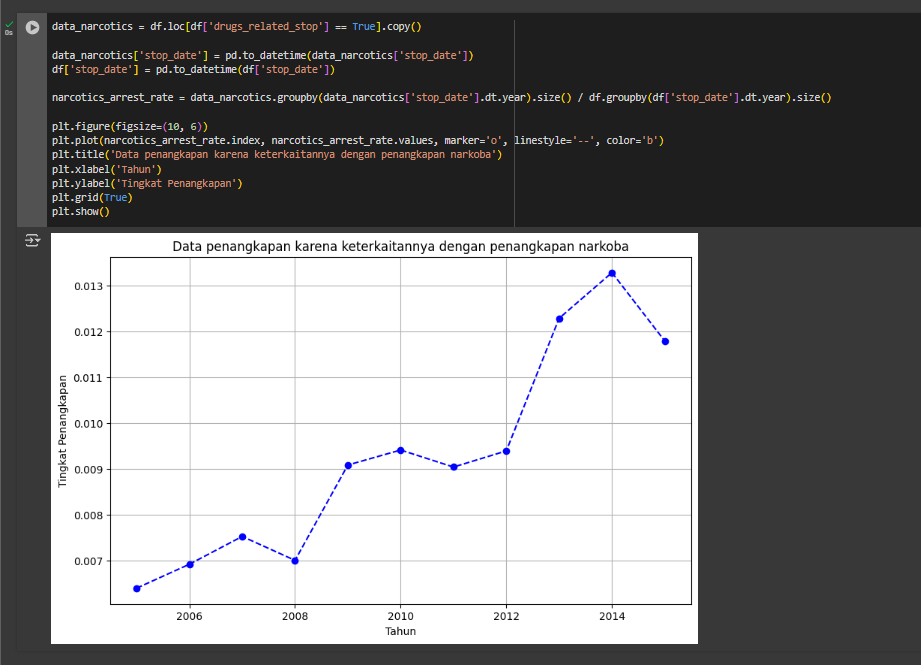
* 1. Dari kedua jenis kelamin tampilkanlah potensi pelanggaran yang banyak dilakukan oleh kedua jenis kelamin tersebut.



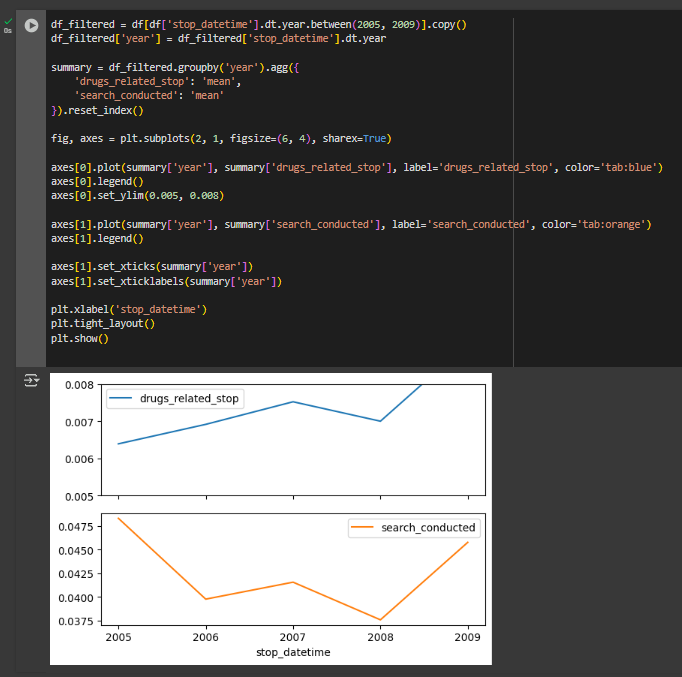
* 1. Buatlah grafik yang menampilkan tingkat penangkapan yang dilakukan oleh pihak kepolisian seperti gambar berikut:



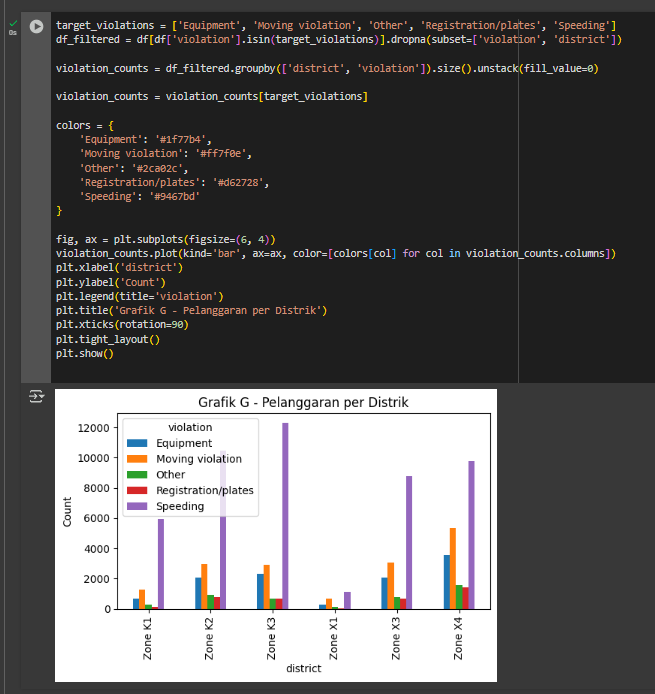
* 1. Tampilkanlah data penangkapan karena keterkaitannya dengan penangkapan narkoba seperti data berikut



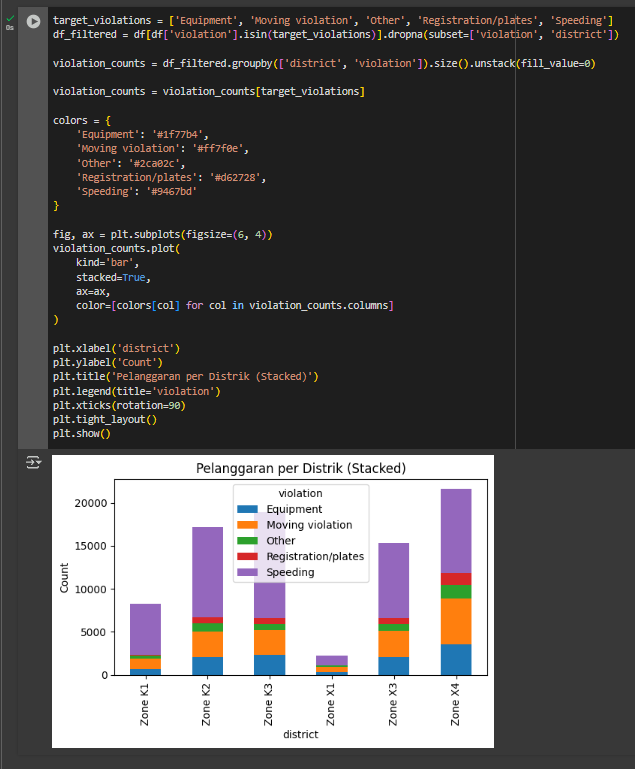
* 1. Tampilkanlah data yang menunjukan pemberhentian kendaraan dengan keterkaitannya dengan penangkapan narkoba



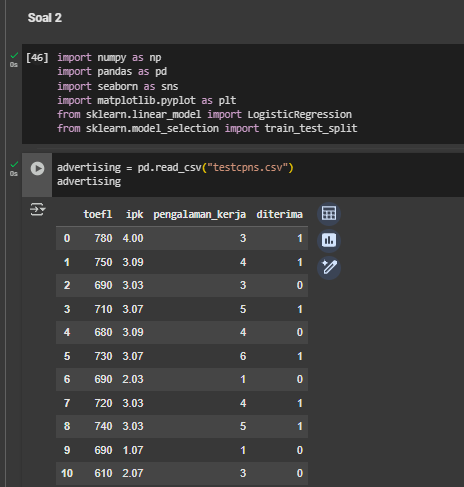
* 1. Tampilkanlah grafik batang yang menunjukan pelanggaran yang berhubungan dengan distrik pengendara



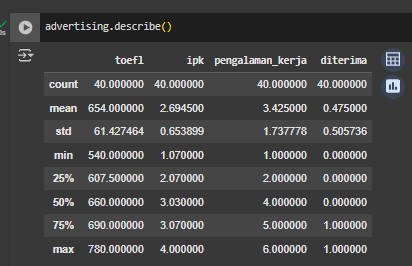
* 1. Tampilkan grafik batang yang menampilkan hubungan distrik dari pengendara dengan pelanggaran yang dilakukan



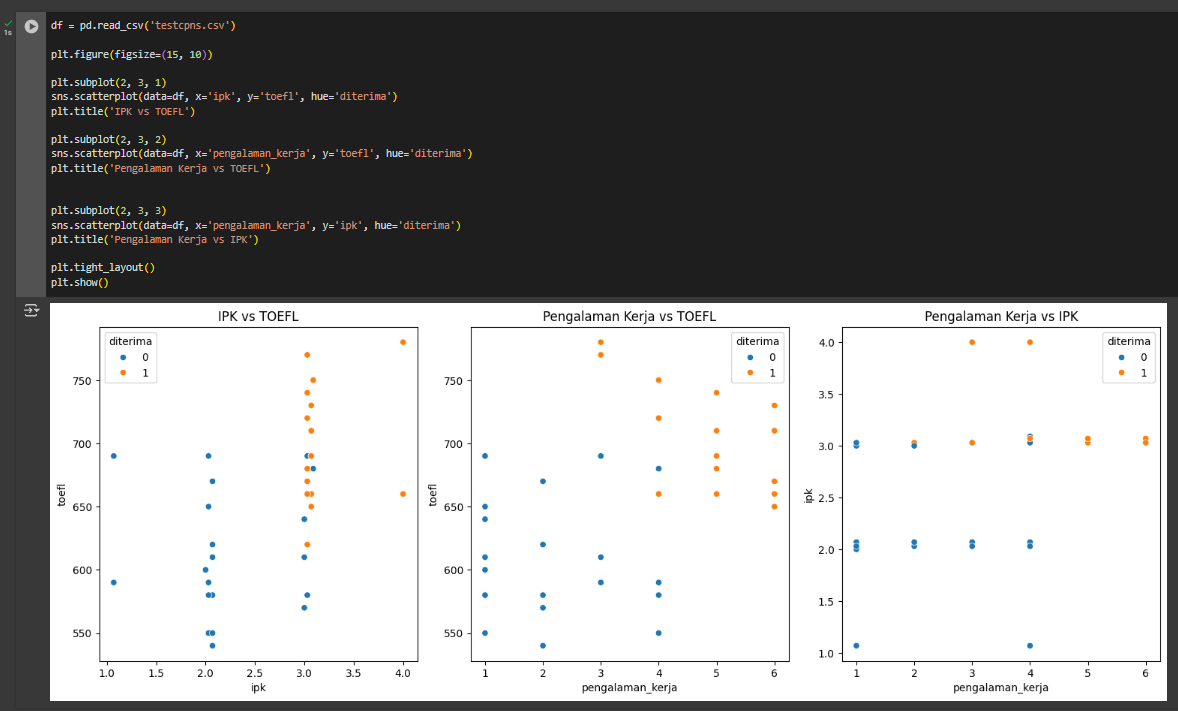
1. Sebuah perusahaan multinasional yang bergengsi menerima seleksi masuk perusahaan dengan data sebagai berikut, data selengkapnya terdapat pada google drive perkuliahan; testcpns.csv:



* 1. Tampilkanlah ringkasan data



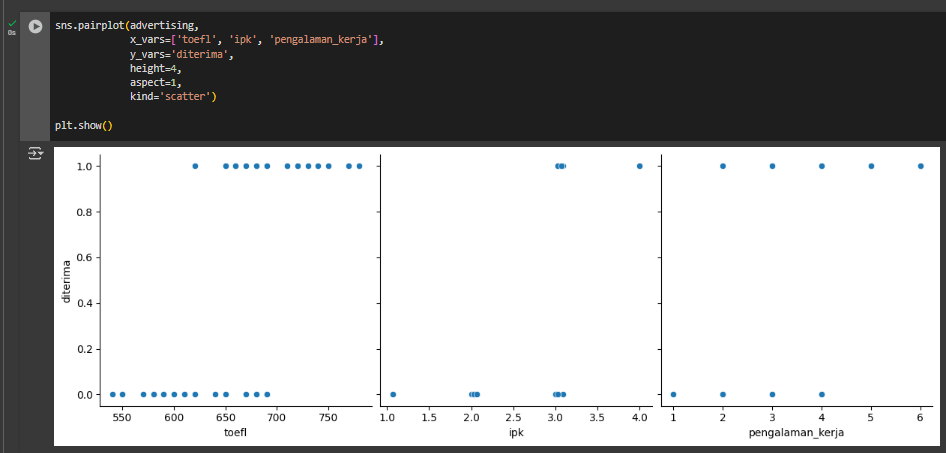
* 1. Sebelum menganalisa data tampilkanlah data dalam bentuk scatter, dengan menggunakan library seaborn



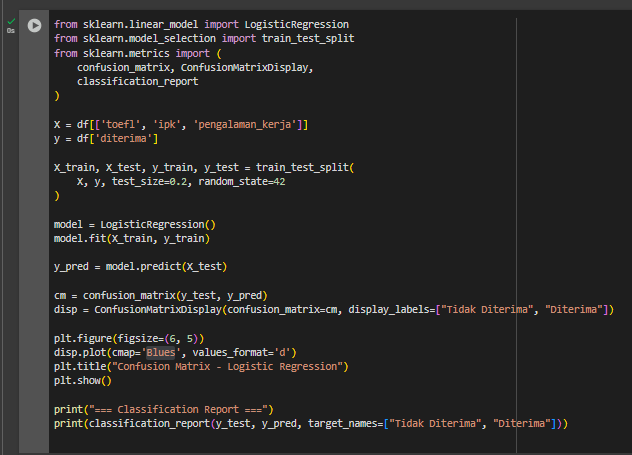
**Analisis Singkat**

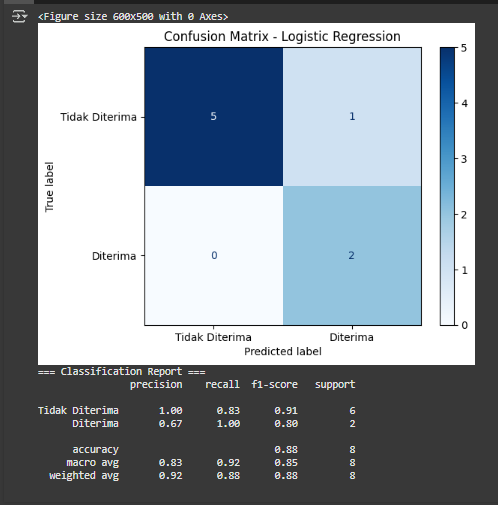
Nilai TOEFL, IPK, dan pengalaman kerja berpengaruh pada diterimanya kandidat. Secara visual, IPK dan TOEFL tampak lebih berpengaruh dengan pola yang lebih jelas. Namun, karena ada tumpang tindih data, model prediktif tetap diperlukan untuk analisis lebih akurat.

* 1. Buatlah visualisasi data seperti gambar berikut



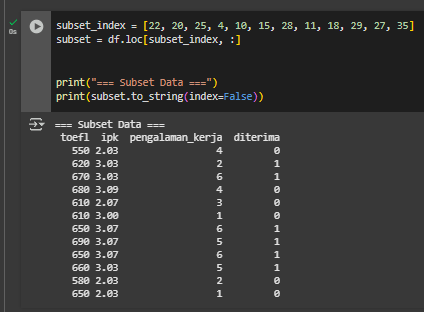
* 1. Buatlah confusion matrix dari data tersebut, dan berikan penjelasan singkat dari hasil confusion matrix tersebut:





**Penjelasan Singkat:**

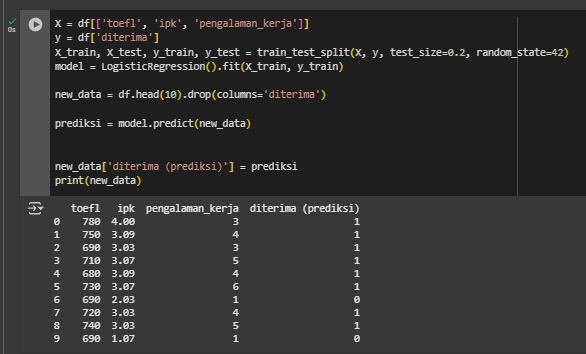
* Data dibagi menjadi data training dan testing (80:20).
* Model **Logistic Regression** dilatih dengan fitur TOEFL, IPK, dan pengalaman kerja.
* Prediksi dilakukan pada data test.
* Ditampilkan **confusion matrix** untuk melihat performa model.
* Dicetak laporan klasifikasi lengkap (precision, recall, f1-score).
  1. Buatlah plot data seperti table, Analisalah hasil di atas berikan penjelasan singkat dengan menggunakan Bahasa Anda sendiri.



**Analisa Singkat**

IPK adalah faktor paling berpengaruh terhadap diterimanya kandidat. Kandidat dengan IPK rendah (2.0–2.5) biasanya tidak diterima meskipun punya pengalaman kerja banyak. TOEFL juga penting, dengan skor ≥600 dan IPK ≥3.0 meningkatkan peluang diterima. Pengalaman kerja kurang berpengaruh dibanding IPK dan TOEFL.

* 1. Dengan menggunakan model yang telah Anda bentuk pada soal 2.e maka prediksilah apakah pelamar baru akan diterima ditempat kerja baru atau tidak diterima.



Link Github: <https://github.com/raflistay/UASVD>