BAB 3
METODE PENELITIAN

Penelitian ini merupakan penelitian eksploratif yang bertujuan untuk menganalisis hubungan antara yaitu jumlah penduduk, tindak pidana, persentase penyelesaian kasus, angka melek aksara, persentase penduduk miskin, dan tingkat pengangguran di Indonesia menggunakan teknik klasterisasi.

3.1 Bahan Dan Alat Penelitian

Penelitian ini menggunakan kumpulan data yang berkaitan dengan kondisi demografi dan tingkat kriminalitas di Indonesia selama periode 2021 hingga 2023. Data meliputi jumlah penduduk, tindak pidana, persentase penyelesaian kasus, angka melek aksara, persentase penduduk miskin, dan tingkat pengangguran. Setiap dataset mencakup 34 provinsi di Indonesia. Data yang digunakan dalam penelitian ini mencakup informasi jumlah penduduk (dalam ribu jiwa) per provinsi, jumlah tindak pidana yang dilaporkan di masing-masing wilayah Kepolisian Daerah (Polda), serta persentase penyelesaian tindak pidana sebagai indikator kinerja penegakan hukum. Variabel sosial-ekonomi yaitu angka melek aksara sebagai representasi tingkat literasi, persentase penduduk miskin untuk menggambarkan kondisi kesejahteraan masyarakat, dan tingkat pengangguran sebagai indikator tekanan ekonomi di suatu wilayah. Seluruh data diperoleh dari situs resmi Badan Pusat Statistik (BPS), yang merupakan sumber data nasional terpercaya. Format data yang telah terstruktur dengan baik memudahkan proses pembersihan, analisis, dan penerapan metode klasterisasi guna mengeksplorasi keterkaitan antara aspek demografis, sosial-ekonomi, dan kriminalitas di berbagai wilayah Indonesia.

Alat yang digunakan dalam penelitian ini adalah komputer dengan spesifikasi cukup untuk menjalankan sistem operasi dan perangkat lunak pengembangan serta koneksitas Internet. Sistem Operasi dan program-program aplikasi yang dipergunakan dalam dalam pengembangan aplikasi ini adalah:

1. Sistem Operasi: Windows versi 11,
2. Google Colab versi 3.11.11,

3. Visual Studio Code 1.101.2,
4. Python versi 3.12.0,
5. Framework Flask versi 3.1.1,
6. NumPy versi 1.26.4,
7. SciPy versi 1.14.1,
8. Scikit-learn versi 1.6.1,
9. Pandas versi 2.2.2,
10. Matplotlib versi 3.10.0,
11. Seaborn versi 0.13.2

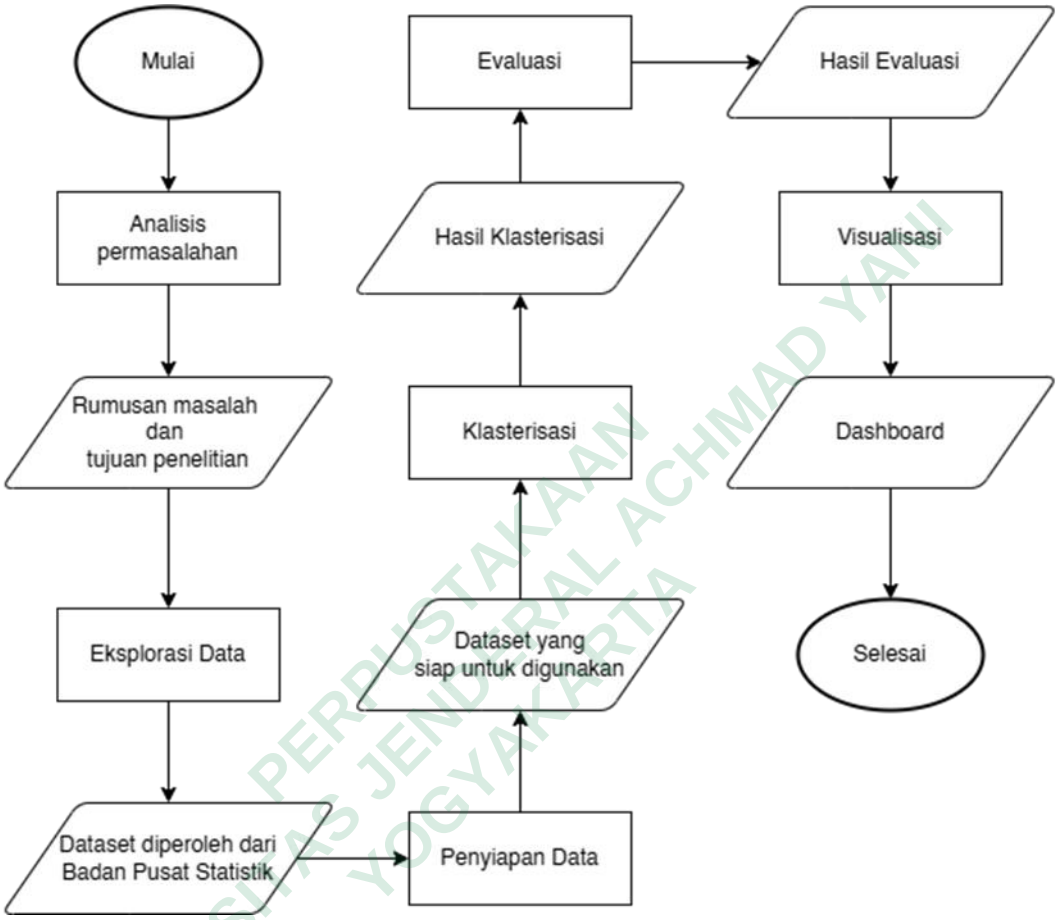
3.2 JALAN PENELITIAN

Penelitian ini menggunakan metode klasterisasi untuk menganalisis hubungan antara jumlah penduduk dan tindak pidana di Indonesia. Pendekatan ini tidak hanya memungkinkan pengelompokan wilayah berdasarkan karakteristik kriminalitasnya, tetapi juga membantu dalam memahami distribusi kejahatan secara lebih komprehensif. Dengan mengelompokkan provinsi-provinsi yang memiliki pola serupa, hasil penelitian ini dapat menjadi dasar dalam perumusan kebijakan keamanan yang lebih efektif serta dalam alokasi sumber daya penegakan hukum.

Dalam metode klasterisasi, data akan diproses melalui beberapa tahapan untuk memastikan hasil yang optimal, ditunjukkan pada Gambar 3.1. Setiap tahap dirancang secara sistematis, dimulai dari pengumpulan data yang valid dan terstruktur, dilanjutkan dengan pra-pemrosesan untuk membersihkan data dari elemen yang dapat mengganggu analisis. Setelah itu, proses klasterisasi dilakukan menggunakan algoritma yang telah ditentukan, dan hasilnya akan dievaluasi untuk memastikan bahwa klaster yang terbentuk benar-benar mencerminkan pola dalam data. Evaluasi klasterisasi dilakukan menggunakan metode *Elbow* dan *Silhouette Coefficient*, sehingga pemilihan jumlah klaster optimal dapat dilakukan secara objektif dan berbasis data.

Hasil dari penelitian ini diharapkan dapat memberikan wawasan yang lebih dalam mengenai hubungan antara jumlah penduduk, tingkat kriminalitas, dan efektivitas penyelesaian kasus di Indonesia. Dengan demikian, penelitian ini dapat

memberikan kontribusi bagi akademisi, aparat penegak hukum, serta pemangku kebijakan dalam upaya meningkatkan efektivitas sistem keamanan dan peradilan di berbagai wilayah.



Gambar 3.1 Jalan Penelitian

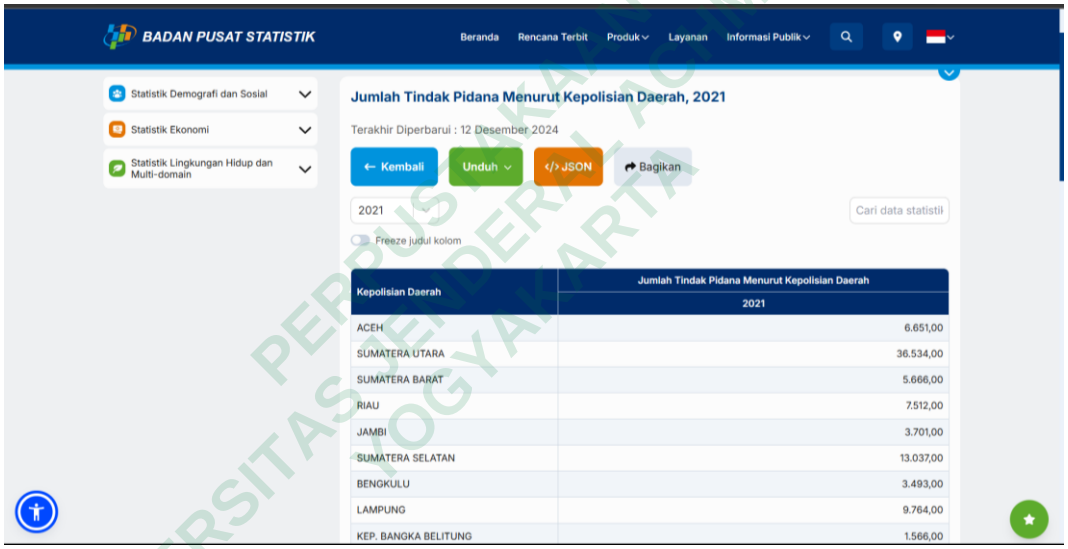
3.2.1 Analisis Permasalahan

Tahap awal dalam penelitian ini dimulai dengan analisis permasalahan yang dilakukan melalui studi literatur terhadap penelitian-penelitian sebelumnya serta kondisi keamanan di Indonesia saat ini. Studi ini mencakup pembahasan mengenai tingkat kriminalitas di Indonesia, teknik klasterisasi dalam analisis data sosial, serta indikator penunjang seperti jumlah penduduk, tindak pidana, persentase penyelesaian kasus, angka melek aksara, persentase penduduk miskin, dan tingkat pengangguran. Dari hasil studi ini diperoleh pemahaman mendalam mengenai gap penelitian yang masih dapat dieksplorasi. Berdasarkan hal tersebut, ditetapkanlah

rumusan masalah dan tujuan penelitian yang menjadi dasar untuk proses-proses selanjutnya.

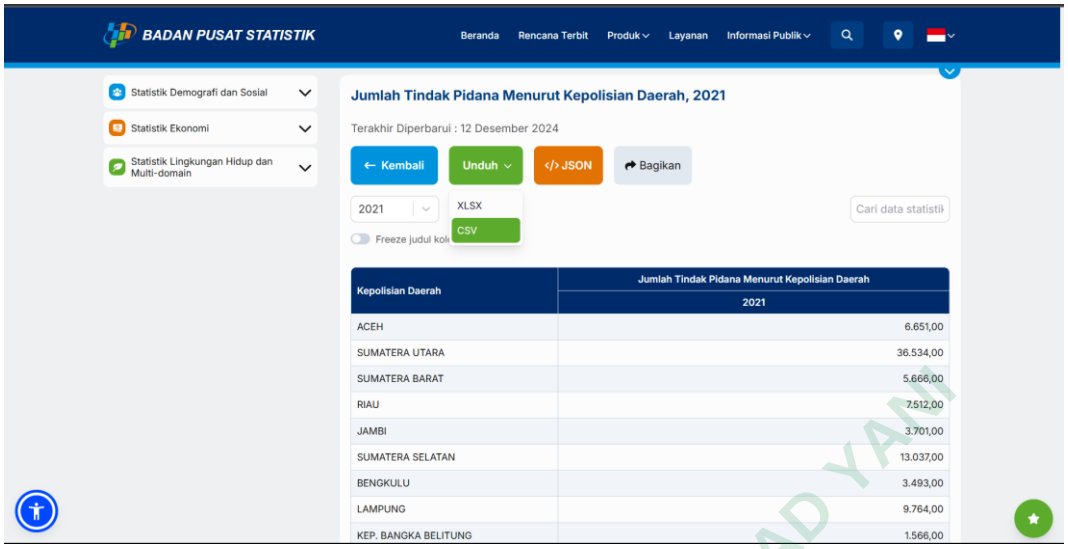
3.2.2 Eksplorasi Data

Setelah merumuskan masalah dan tujuan, langkah berikutnya adalah eksplorasi data yang bertujuan untuk mengidentifikasi jenis data yang dibutuhkan untuk menjawab permasalahan penelitian. Proses eksplorasi ini menghasilkan keputusan untuk menggunakan data resmi dari BPS yang ditujukan pada Gambar 3.2, yaitu data jumlah penduduk, tindak pidana, persentase penyelesaian kasus, angka melek aksara, persentase penduduk miskin, dan tingkat pengangguran menurut provinsi dari tahun 2021 hingga 2023.



Gambar 3.2 Eksplorasi dilakukan pada situs BPS

Data ini dipilih karena memiliki struktur yang rapi, cakupan nasional yang lengkap, serta relevan dengan indikator yang dibutuhkan dalam analisis klasterisasi. Data dikumpulkan dalam format CSV yang ditunjukkan pada Gambar 3.3 untuk diolah lebih lanjut.



Gambar 3.3 Mengunduh dataset dalam bentuk CSV

3.2.3 Penyiapan Data

Dataset yang diperoleh dari BPS kemudian diproses lebih lanjut pada tahap penyiapan data. Proses ini meliputi pembersihan data, penggabungan data antar tahun, penyesuaian format, dan penanganan data yang hilang atau tidak konsisten. Penyiapan ini bertujuan untuk memastikan bahwa seluruh atribut dalam dataset dapat digunakan secara optimal dalam analisis klasterisasi. Hasil dari tahap ini adalah dataset yang telah siap seperti yang ditunjukkan pada Tabel 3.1, untuk digunakan pada proses pengelompokkan data, dengan struktur data yang konsisten dan bebas dari anomali.

Tabel 3.1 Dataset siap untuk dikelompokkan

jumlah_penduduk	*jumlah_pidana*	*persentase_penyelesaian_tindak_pidana*	literasi	*persentase_penduduk_miskin*	*tingkat_pengangguran*
5274.9	6651	56.98	97.39	15.33	6.3
14799.4	36534	68.37	98.84	9.01	6.01
5534.5	5666	100	99	6.63	6.67
6394.1	7512	77.34	98.92	7.12	4.96
2064.6	2481	65.7	98.92	6.12	10.12
3548.2	3701	79.19	97.47	8.09	4.76

8467.4	13037	74	98.27	12.84	5.17
1455.7	1566	69.22	97.08	4.9	5.04
2010.7	3493	46.81	96.83	15.22	3.72
9007.8	9764	64.32	95.98	12.62	4.54

3.2.4 Klasterisasi

Tahap selanjutnya adalah penerapan teknik klasterisasi terhadap dataset yang telah siap. Pada proses ini, dilakukan pengelompokan provinsi di Indonesia berdasarkan variabel-variabel seperti jumlah penduduk, tindak pidana, persentase penyelesaian kasus, angka melek aksara, persentase penduduk miskin, dan tingkat pengangguran. Tujuannya adalah untuk mengidentifikasi kelompok-kelompok wilayah yang memiliki karakteristik yang serupa. Hasil dari tahap ini adalah sejumlah klaster yang menggambarkan pola-pola kriminalitas di Indonesia secara tersegmentasi.

```
import pandas as pd
from google.colab import drive
drive.mount('/content/drive')

path_base = '/content/drive/MyDrive/SMTR_8/TA/'

dataset_all = pd.read_csv(path_base + 'dataset_all_6.csv')
```

1. Pemanggilan dan Pemuatan Data

   Dataset diunggah ke Google Drive dan kemudian dimuat ke dalam Google Colab menggunakan pustaka pandas. Data ini mencakup enam variabel utama: jumlah penduduk, jumlah tindak pidana, persentase penyelesaian tindak pidana, literasi, persentase penduduk miskin, dan tingkat pengangguran.

```
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt
import numpy as np


df = pd.DataFrame(dataset_all)
```

2. Penyiapan lingkungan dan Data untuk Klasterisasi

Melakukan impor beberapa pustaka penting seperti KMeans dari sklearn.cluster untuk membentuk klaster, StandardScaler dari sklearn.preprocessing untuk menstandarkan skala data agar setiap fitur memiliki rata-rata 0 dan standar deviasi 1, serta *silhouette_score* dari sklearn.metrics yang digunakan untuk mengevaluasi kualitas klaster berdasarkan seberapa baik data berada dalam klasternya masing-masing. Selain itu, digunakan matplotlib.pyplot untuk membuat visualisasi seperti grafik *elbow* atau distribusi klaster, dan numpy untuk melakukan operasi numerik seperti perhitungan rata-rata dan jarak. Kemudian, data yang telah dikumpulkan dan disatukan dalam objek *dataset_all* diubah menjadi struktur tabel DataFrame menggunakan pustaka pandas, yang menjadi dasar untuk analisis dan pemrosesan selanjutnya dalam proses klasterisasi. Keseluruhan proses ini merupakan tahap persiapan penting dalam analisis data, sebelum pola pengelompokkan data K-Means diterapkan dan hasilnya dianalisis lebih lanjut.

```
X=df[['jumlah_penduduk','jumlah_pidana','persentase_penyelesaian
_tindak_pidana','literasi','persentase_penduduk_miskin','tingkat
_pengangguran']]
dd=df[['jumlah_penduduk','jumlah_pidana','persentase_penyelesaia
n_tindak_pidana','literasi','persentase_penduduk_miskin','tingka
t_pengangguran']]

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

corr_matrix = dd.corr()
print(corr_matrix)
```

3. Seleksi Fitur dan Transformasi Data

Melakukan pemilihan fitur yang akan digunakan untuk proses klasterisasi. Kolom-kolom yang dipilih yaitu *jumlah_penduduk*, *jumlah_pidana*, *persentase_penyelesaian_tindak_pidana*, literasi, *persentase_penduduk_miskin*, dan *tingkat_pengangguran*, yang masing-masing merepresentasikan aspek demografis, kriminalitas, dan

sosial-ekonomi tiap provinsi. Nilai-nilai pada kolom tersebut kemudian disalin ke dalam dua variabel, yakni X dan dd, untuk keperluan pemrosesan dan analisis yang berbeda. Variabel X akan digunakan dalam proses normalisasi menggunakan StandardScaler, sedangkan dd dipertahankan dalam bentuk asli untuk analisis korelasi. Proses normalisasi (*X_scaled* = scaler.*fit_transform*(X)) mengubah seluruh fitur menjadi skala yang seragam (mean = 0 dan standar deviasi = 1), yang sangat penting dalam K-Means agar setiap fitur memiliki pengaruh yang setara. Selanjutnya, dihitung matriks korelasi (*corr_matrix* = dd.corr()), yang digunakan untuk melihat hubungan antar fitur dan mendeteksi apakah ada fitur yang terlalu berkorelasi tinggi (redundan), atau justru tidak memiliki korelasi signifikan terhadap yang lain. Hasil ini dapat menjadi pertimbangan dalam analisis lanjutan untuk memastikan kualitas klasterisasi tetap optimal. Hasil dari kode dibawah dapat dilihat pada Tabel 3.2.

Tabel 3.2 Hasil *print corr_matrix*

index	A	B	C	D	E	F
A	1.0	0.6717977866670078	0.3204864286794625	-0.10411813832585608	-0.0902532083793 6229	0.3511622581440948
B	0.6717977866670078	1.0	0.013798913367496462	-0.014403308913167566	-0.1379919580379 7923	0.3667497110400259
C	0.3204864286794625	0.013798913367496462	1.0	0.024925928080339686	-0.3374181846801 2524	-0.0970757242145755 3
D	-0.104118	-0.0144033089	0.024925928080339686	1.0	-0.544176	0.3998201882

	13832585608	13167566			469529475 5	4954896
E	-0.0902532083793 6229	-0.13799195803797923	-0.33741818468012524	-0.5441764695294755	1.0	-0.3334276043954192
F	0.3511622581440948	0.3667497110400259	-0.09707572421457553	0.3998201882495489 6	-0.3334276043954192	1.0

index:

1. A = *jumlah_penduduk*
2. B = *jumlah_pidana*
3. C = *persentase_penyelesaian_tindak_pidana*
4. D = literasi
5. E = *persentase_penduduk_miskin*
6. F = *tingkat_pengangguran*

Hasil analisis korelasi antar variabel menunjukkan hubungan yang cukup logis antara beberapa indikator sosial dan demografis yang digunakan dalam proses klasterisasi. Salah satu hubungan terkuat ditunjukkan oleh variabel *jumlah_penduduk* dan *jumlah_tindak_pidana*, dengan nilai korelasi sebesar 0,6718, yang mengindikasikan bahwa provinsi dengan *jumlah_penduduk* lebih tinggi cenderung memiliki *jumlah_tindak_pidana* yang lebih besar. Hal ini dapat dipahami karena wilayah dengan populasi padat biasanya memiliki tingkat interaksi sosial dan tekanan ekonomi yang lebih tinggi, sehingga peluang terjadinya pelanggaran hukum meningkat. Selain itu, terdapat korelasi sedang antara *jumlah_penduduk* dan *tingkat_pengangguran* (0,3512), serta korelasi negatif lemah terhadap tingkat literasi (–0,1041) dan *persentase_penduduk_miskin* (–0,0903), yang mengindikasikan bahwa kepadatan penduduk tidak selalu

berbanding lurus dengan tingkat kesejahteraan atau kualitas pendidikan.

Variabel literasi menunjukkan korelasi negatif cukup kuat terhadap persentase penduduk miskin (–0,5442), menandakan bahwa semakin tinggi tingkat literasi suatu wilayah, semakin rendah tingkat kemiskinannya. Temuan ini sejalan dengan banyak studi yang menekankan peran pendidikan dalam meningkatkan kesejahteraan masyarakat. Adapun variabel *tingkat_pengangguran* memiliki korelasi sedang terhadap *jumlah_pidana* (0,3667) dan *jumlah_penduduk* (0,3512), namun justru berkorelasi negatif terhadap *persentase_penduduk_miskin* (–0,3334), yang sedikit menyimpang dari ekspektasi umum. Hal ini dapat dijelaskan dengan kemungkinan adanya intervensi sosial seperti bantuan pemerintah atau program padat karya di daerah-daerah dengan pengangguran tinggi. Secara keseluruhan, variabel-variabel ini berkontribusi dalam membentuk representasi sosial dan ekonomi tiap provinsi, yang penting untuk diidentifikasi dalam proses klasterisasi.

Sebagian besar variabel menunjukkan hubungan satu sama lain, variabel *persentase_penyelesaian_tindak_pidana* justru memiliki korelasi yang sangat lemah terhadap seluruh variabel lainnya seperti 0,3205 terhadap *jumlah_penduduk*, 0,0138 terhadap *jumlah_pidana*, dan bahkan negatif terhadap *persentase_penduduk_miskin* (–0,3374) dan *tingkat_pengangguran* (–0,0971). Meskipun demikian, variabel ini tetap relevan dan perlu dipertahankan dalam proses klasterisasi karena membawa informasi independen yang tidak tercermin dalam variabel lainnya. Variabel *persentase_penyelesaian_tindak_pidana* merepresentasikan aspek kinerja institusional dari aparat penegak hukum, bukan kondisi sosial masyarakat. Dengan demikian, keberadaannya memberikan dimensi tambahan dalam pengelompokkan data, sehingga hasil klasterisasi tidak hanya menggambarkan tingkat kriminalitas dan kondisi ekonomi, tetapi juga memperlihatkan

efektivitas respons hukum di tiap wilayah. Selain itu, dalam analisis multivariat seperti klasterisasi, kehadiran fitur yang tidak multikolinear (tidak berkorelasi tinggi dengan fitur lain) dapat memperkaya segmentasi dan meningkatkan akurasi dalam membedakan karakteristik wilayah. Oleh karena itu, mempertahankan variabel ini tidak hanya tepat secara teknis, tetapi juga krusial secara substansial untuk mencapai tujuan utama penelitian dalam mengevaluasi pola kriminalitas beserta efektivitas penanganannya di Indonesia.

```
range_n_clusters = range(2,6)
silhouette_scores = []
inertia = []

for n_clusters in range_n_clusters:
  kmeans = KMeans(n_clusters=n_clusters, random_state=42)
  cluster_labels = kmeans.fit_predict(X_scaled)
  silhouette_avg = silhouette_score(X_scaled, cluster_labels)
  silhouette_scores.append(silhouette_avg)
  inertia.append(kmeans.inertia_)

print(f"Silhouette score for {n_clusters} clusters:
{silhouette_avg}")

plt.figure(figsize=(12,5))
plt.subplot(1,2,1)
plt.plot(range_n_clusters, silhouette_scores, marker='o')
plt.xlabel('Number of Clusters')
plt.ylabel('Silhouette score')
plt.title('Silhouette score Analysis')


plt.subplot(1,2,2)
plt.plot(range_n_clusters, inertia, marker='o')
plt.xlabel('Number of Clusters')
plt.ylabel('Inertia')
plt.title('Elbow method')


plt.tight_layout()
plt.show()
```

4. Pemilihan Nilai K dalam Klasterisasi K-Means

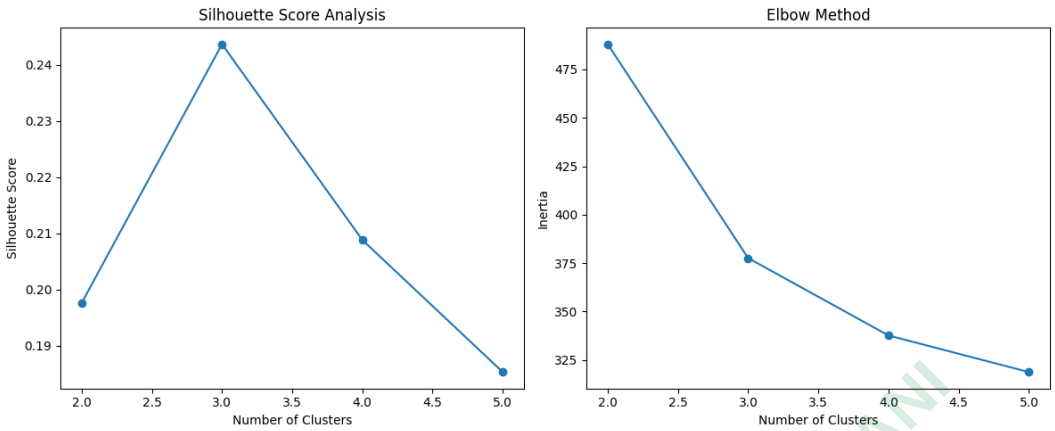
Untuk menentukan jumlah klaster yang optimal dalam proses klasterisasi, digunakan dua metode evaluasi internal, yaitu *Silhouette score* dan *Elbow method*. Kode yang digunakan mengevaluasi performa klasterisasi K-Means untuk jumlah klaster antara 2 hingga 5. Pada setiap

iterasi, pola pengelompokkan K-Means dibentuk menggunakan data yang telah dinormalisasi, dan kemudian dilakukan pengelompokan data ke dalam klaster berdasarkan centroid yang dihasilkan. Kualitas pemisahan antar klaster dievaluasi menggunakan *Silhouette score*, yaitu metrik yang mengukur seberapa baik suatu data cocok dengan klaster-nya sendiri dibandingkan dengan klaster lain. Nilai *silhouette* yang tinggi (mendekati 1) menunjukkan bahwa objek dikelompokkan dengan baik.

Selain itu, juga dihitung nilai inertia, yaitu jumlah kuadrat jarak antara data dengan pusat klasternya, yang mencerminkan kepadatan internal suatu klaster. Nilai inertia secara alami menurun seiring bertambahnya jumlah klaster, namun penurunan tersebut akan melambat setelah titik tertentu; titik perlambatan ini dikenal sebagai *elbowpoint*, yang dianggap sebagai indikator jumlah klaster optimal. Hasil evaluasi kemudian divisualisasikan dalam dua grafik: grafik *Silhouette score* terhadap jumlah klaster untuk menilai kualitas pemisahan, dan grafik *Elbow method* terhadap jumlah klaster untuk mengidentifikasi titik tekukan. Pemilihan jumlah klaster dilakukan secara objektif berdasarkan metrik evaluatif yang ditunjukkan oleh Gambar 3.3 yang menampilkan *Silhouette score* dan Gambar 3.4 yang menamilkan grafik *silhouette score* dan grafik *elbow*. Dengan menggunakan kode berikut:

```
Silhouette Score for 2 clusters: 0.19757850311513217
Silhouette Score for 3 clusters: 0.24361957169924112
Silhouette Score for 4 clusters: 0.20880684684223416
Silhouette Score for 5 clusters: 0.18531171856606124
```

Gambar 3.3 *Silhouette Score*



Gambar 3.4 Grafik *Silhouette score* dan Grafik *Elbow*

Berdasarkan kedua grafik tersebut, dapat disimpulkan bahwa jumlah klaster optimal dalam penelitian ini adalah 3. Hasil ini konsisten secara kuantitatif (melalui nilai *Silhouette* tertinggi) maupun visual (melalui titik tekukan dalam *Elbow method*). Oleh karena itu, pola pengelompokkan data klasterisasi K-Means akan menggunakan K=3 sebagai jumlah klaster utama untuk proses selanjutnya.

```
best_n_clusters = range_n_clusters[np.argmax(silhouette_scores)]
  print(f"Best number of clusters: {best_n_clusters}")

kmeans = KMeans(n_clusters=best_n_clusters, random_state=42)
clusters_all = kmeans.fit_predict(X_scaled)

dd['cluster'] = clusters_all

final_silhouetter = silhouette_score(X_scaled, clusters_all)
print(f"Final Silhouette score: {final_silhouetter}")

dataset_all['cluster'] = clusters_all
print(dataset_all['cluster'].tolist())
```

5. Penerapan Klasterisasi

Setelah dilakukan evaluasi untuk menentukan jumlah klaster optimal menggunakan *silhouette score*, tahap selanjutnya adalah membangun klasterisasi menggunakan K-Means akhir berdasarkan nilai terbaik tersebut. Jumlah klaster terbaik (*best_n_clusters*) diperoleh dengan menggunakan fungsi np.argmax() untuk mengambil indeks tertinggi dari daftar skor *Silhouette*, kemudian digunakan sebagai

parameter jumlah klaster. Hasi klasterisasi K-Means kemudian dibangun ulang dengan nilai K optimal, *random_state* untuk memastikan reprodusibilitas. Data yang telah dinormalisasi (*X_scaled*) digunakan untuk mendapatkan pola pengelompokkan data dan menghasilkan label klaster untuk setiap entri dalam dataset. Label ini disimpan dalam variabel *clusters_all* dan ditambahkan ke DataFrame (dd) sebagai kolom cluster untuk analisis lanjutan. Sebagai bentuk validasi akhir, nilai *Silhouette score* juga dihitung kembali terhadap seluruh dataset gabungan untuk menilai kualitas pola kelompok data secara keseluruhan. Skor yang ditunjukkan pada Gambar 3.5 *Final Silhouette* memastikan bahwa pemisahan antar klaster tetap baik setelah pola pengelompokkan data diterapkan secara final. Terakhir, hasil klasterisasi disisipkan ke dalam *dataset_all*, sehingga masing-masing entitas data kini memiliki atribut klaster. Berikut merupakan kodenya:

```
Best number of clusters: 3
Final Silhouette Score: 0.24361957169924112
[2, 0, 2, 2, 2, 2, 2, 2, 1, 2, 0, 0, 2, 0, 2, 0, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 1, 1, 1, 2, 1, 1, 2, 0, 2,
```

Gambar 3.5 *Final Silhouette*

6. Visualisasi Hasil Klasterisasi Menggunakan PCA

Interpretasi hasil klasterisasi, dilakukan visualisasi data dalam ruang dua dimensi menggunakan *Principal Component Analysis* (PCA) seperti Gambar 3.7 PCA. PCA merupakan teknik reduksi dimensi yang bertujuan untuk mengubah data berdimensi tinggi menjadi bentuk dua dimensi (2D) sambil tetap mempertahankan sebanyak mungkin variansi dari data asli. Dalam kode ini, objek PCA diinisialisasi dengan *n_components*=2, lalu diterapkan pada data yang telah dinormalisasi (*X_scaled*) untuk menghasilkan representasi baru dalam dua komponen utama, yang disimpan dalam *X_pca_all*. Selanjutnya, posisi centroid dari masing-masing klaster hasil K-Means juga diproyeksikan ke ruang PCA menggunakan pca.transform(kmeans.*cluster_centers_*), agar dapat divisualisasikan bersama dengan data observasi.

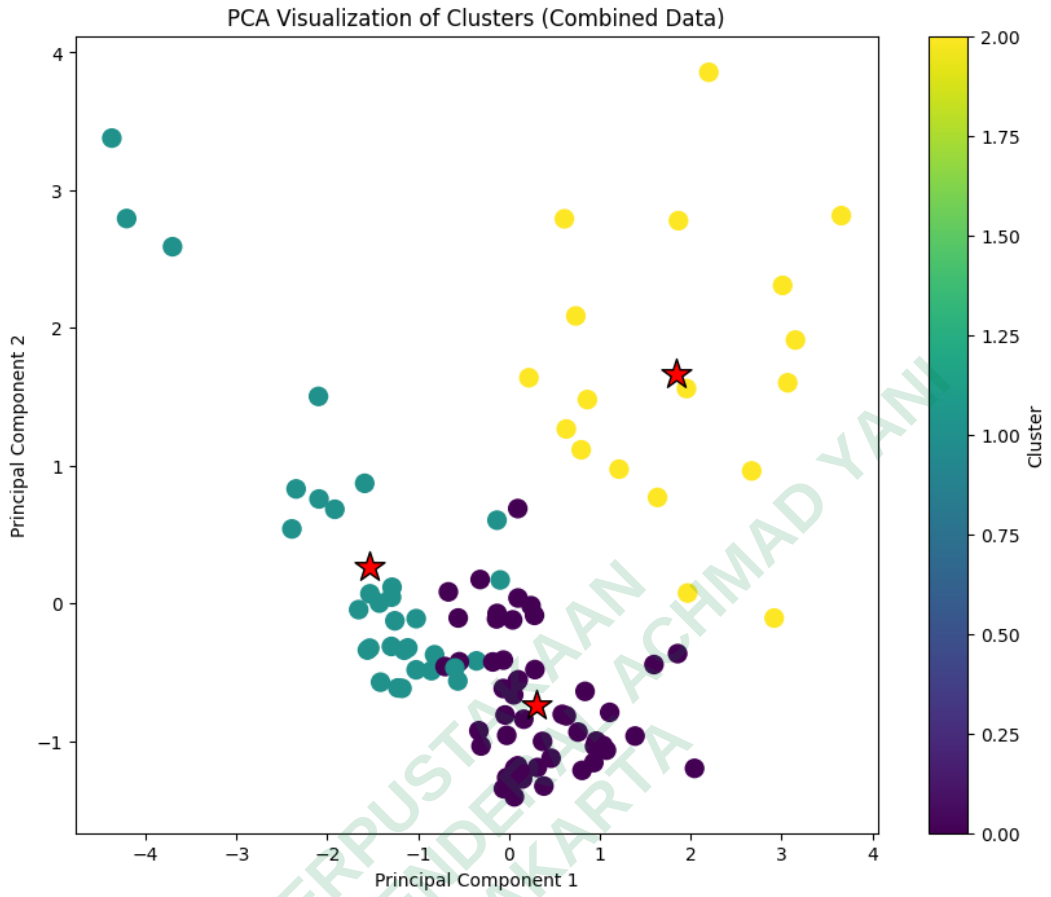
Visualisasi dilakukan menggunakan matplotlib, di mana setiap titik data ditampilkan dalam scatter plot dan diwarnai berdasarkan label klaster (*clusters_all*) dengan colormap 'viridis'. Titik centroid masing-masing klaster digambarkan dengan simbol bintang berwarna merah (marker='*'), berukuran lebih besar (s=300), dan diberi garis tepi hitam untuk menonjolkan perannya sebagai pusat klaster. Visualisasi ini memberikan gambaran intuitif mengenai distribusi dan pemisahan antar klaster, serta seberapa jauh setiap klaster tersebar terhadap pusatnya. Selain itu, skor *Silhouette* akhir (*final_silhouetter*) dicetak kembali untuk memperkuat validasi hasil pola pengelompokkan data seperti pada Gambar 3.6 Validasi Hasil Pengelompokkan Data. Berikut merupakan kodenya:

```
Final Silhouette Score: 0.24361957169924112
    jumlah_penduduk  jumlah_pidana  persentase_penyelesaian_tindak_pidana  \
1           14799.4          36534                                  68.37
15          40665.7          19257                                  55.35
10          10562.1          29103                                  97.99
11          48274.2           7502                                  73.06
13          36516.0           8909                                  70.78
..              ...            ...                                    ...
88            720.1           1701                                  95.47
89           2660.8          14265                                  20.74
91           3051.2           8944                                  57.88
84           5549.7           6028                                  65.15
99           1318.5           2334                                  45.07

    literasi  persentase_penduduk_miskin  tingkat_pengangguran  cluster
1      98.84                        9.01                  6.01        0
15     90.06                       11.40                  5.17        0
10     99.67                        4.72                  8.51        0
11     97.93                        8.40                  8.92        0
13     91.39                       11.79                  5.96        0
..       ...                         ...                   ...      ...
88     97.78                        6.45                  4.10        2
89     99.79                        7.38                  6.19        2
91     98.14                       12.41                  3.49        2
84     94.79                        6.71                  4.52        2
99     98.81                        6.46                  4.60        2

[102 rows x 7 columns]
```

Gambar 3.6 Validasi Hasil Pengelompokkan Data



Gambar 3.7 Visualisasi PCA Klasterisasi Data Gabungan

Setelah proses klasterisasi dilakukan pada data gabungan, hasil pola pengelompokkan data kemudian digunakan untuk mengelompokkan data tahun 2021, 2022, dan 2023 secara terpisah guna mengamati konsistensi dan perubahan pola tiap tahun.

3.2.5 Evaluasi

Hasil pengelompokan tersebut dievaluasi untuk menilai kualitas pengelompokkan data. Evaluasi dilakukan menggunakan metrik tertentu seperti *Silhouette Coefficient*, guna menilai seberapa baik objek-objek dalam satu klaster menyerupai satu sama lain, dan seberapa berbeda mereka dari klaster lainnya. Dari proses ini diperoleh hasil evaluasi yang menjadi dasar dalam menentukan validitas dan efektivitas pola pengelompokkan data klasterisasi yang telah dibangun seperti yang ditunjukkan pada Gambar 3.3 *Silhouette Score*.

3.2.6 Visualisasi

Hasil evaluasi klaster yang telah diverifikasi selanjutnya digunakan sebagai dasar dalam proses visualisasi. Tahap ini bertujuan untuk menyajikan hasil klasterisasi ke dalam bentuk yang lebih mudah dipahami. Visualisasi ini kemudian diimplementasikan dalam bentuk *dashboard* berbasis web.