

TUGAS MANDIRI
FUNDAMENTALS OF DATA MINING

**Analisis Segmentasi Pelanggan Mall Menggunakan Algoritma K-Means
Clustering**



Nama : Pegi Lathifah
NPM : 231510062
Dosen : Erlin Elisa, S.Kom., M.Kom.

**PROGRAM STUDI SISTEM INFORMASI
TEKNIK KOMPUTER DAN INFORMATIKA
UNIVERSITAS PUTERA BATAM**

2025

1. Deskripsi Dataset

- **Sumber dataset:** Kaggle.com (<https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python>)
- **Jumlah record:** 200 data pengguna.
- **Jumlah atribut:** 5 atribut (CustomerID, Gender, Age, Annual Income, Spending Score).
- **Tipe data:** Numerik (Integer) dan Kategorikal (String/Object).
- **Target/label:** Tidak ada (Unsupervised Learning).
- **Permasalahan yang ingin diselesaikan:** Dataset ini digunakan untuk memahami segmentasi pelanggan mall. Permasalahannya adalah bagaimana mengelompokkan pelanggan ke dalam beberapa segmen berdasarkan pendapatan tahunan dan skor pengeluaran agar tim pemasaran dapat merancang strategi yang tepat sasaran.

2. Persiapan Data & Preprocessing

Langkah-langkah yang dilakukan:

- **Data cleaning:** Memeriksa nilai kosong (*missing value*). Hasilnya, dataset ini tidak memiliki data yang hilang atau null.
- **Encoding:** Mengubah data kategorikal Gender (Male/Female) menjadi numerik (0/1) menggunakan **LabelEncoder**.
- **Scaling / Normalization:** Menggunakan **StandardScaler** pada fitur Annual Income dan Spending Score agar perbedaan skala antar fitur tidak mempengaruhi hasil perhitungan jarak pada algoritma clustering.
- **Feature selection:** Untuk visualisasi yang optimal, dipilih dua fitur utama: Annual Income (k\$) dan Spending Score (1-100).
- **Split data:** Karena ini adalah tugas *Clustering*, seluruh data (100%) digunakan untuk melatih model guna menemukan pola kelompok secara alami.

➤ Tabel Statistik Deskriptif Detail (Sebelum Scaling)

Deskripsi	Age	Annual Income (k\$)	Spending Score (1- 100)
Count	200	200	200
Mean (Rata-rata)	38.85	60.56	50.20
Std. Deviation	13.96	26.26	25.82
Minimum	18	15	1
25% (Kuartil 1)	28.75	41.50	34.75
50% (Median)	36	61.50	50
75% (Kuartil 3)	49	78	73
Maximum	70	137	99

➤ Ringkasan Hasil Preprocessing (Scaling)

Tabel ini menunjukkan bukti teknis bahwa data telah dinormalisasi menggunakan **StandardScaler** (mengubah mean menjadi 0 dan standar deviasi menjadi 1).

Nama Fitur	Mean (Original)	Mean (Scaled)	Std. Dev (Original)	Std. Dev (Scaled)
Annual Income	60.56	0.00	26.26	1.00
Spending Score	50.20	0.00	25.82	1.00

➤ Sampel Transformasi Data (5 Data Pertama)

Tabel ini diletakkan di bagian **Persiapan Data** sebagai bukti proses perubahan data dari nilai asli ke nilai yang dipahami mesin.

ID Pelanggan	Annual Income (Asli)	Annual Income (Scaled)	Spending Score (Asli)	Spending Score (Scaled)
1	15	-1.7389	39	-0.4348
2	15	-1.7389	81	1.1957
3	16	-1.7008	6	-1.7159

ID Pelanggan	Annual Income (Asli)	Annual Income (Scaled)	Spending Score (Asli)	Spending Score (Scaled)
4	16	-1.7008	77	1.0404
5	17	-1.6626	40	-0.3959

➤ Kualitas Data (Cleaning & Validation)

Atribut	Missing Value	Outlier	Status Data
Age	0	Tidak Ada	Bersih
Annual Income	0	1 Terdeteksi	Dipertahankan
Spending Score	0	Tidak Ada	Bersih
Gender	0	N/A	Ter-encode

3. Analisis Statistik & Visualisasi

- **Statistik deskriptif:** Mayoritas pelanggan berada pada rentang usia 20-40 tahun dengan pendapatan rata-rata 60.560 USD.
- **Distribusi:** Visualisasi histogram menunjukkan bahwa sebaran pendapatan tahunan mendekati distribusi normal, namun terdapat beberapa pelanggan dengan pendapatan sangat tinggi.

- **Korelasi:** Heatmap menunjukkan korelasi yang sangat rendah antara usia dan pengeluaran, mengindikasikan bahwa usia bukan faktor penentu utama seseorang boros atau hemat.
- **Insight:** Melalui scatter plot awal, terlihat ada 5 kelompok data yang tersebar secara visual, yang menjadi dasar kuat untuk melakukan clustering.

4. Pemilihan dan Penerapan Algoritma

- **Nama algoritma:** K-Means Clustering.
- **Alasan pemilihan:** K-Means sangat efektif untuk mencari kelompok dalam data numerik yang bersifat *globular* (bulat) dan mudah diinterpretasikan untuk kebutuhan bisnis seperti segmentasi pasar.
- **Parameter utama:** n_clusters=5 (ditentukan melalui metode Elbow), init='k-means++', random_state=42.

Daftar Algoritma yang diUji :

Algoritma	Library Python	Tujuan
K-Means	sklearn.cluster	Segmentasi profil pelanggan
Elbow Method	matplotlib	Menentukan jumlah cluster optimal

5. Pengujian dan Evaluasi Model

Metode evaluasi yang digunakan adalah **Silhouette Score** (mengukur seberapa rapat suatu cluster) dan **Inertia (WCSS)**.

Tabel A : Metrics Evaluasi Model Gunakan tabel ini untuk menunjukkan bahwa jumlah cluster ($K=5$) adalah pilihan yang valid secara ilmiah.

Pada tahap ini, dilakukan evaluasi terhadap model *K-Means Clustering* untuk menentukan kualitas pengelompokan pelanggan

Jenis Evaluasi	Nama Metrics	Nilai Hasil	Keterangan
Internal	Silhouette Score	0.5547	Struktur cluster cukup kuat dan terpisah baik.
Kekompakan	Inertia (WCSS)	44.448	Nilai penurunan konstan (titik siku pada Elbow).
Kerapatan	Davies-Bouldin Index	0.573	Semakin rendah nilai, semakin baik pemisahannya.

Metode K-Means termasuk ke dalam unsupervised learning sehingga tidak menghasilkan nilai akurasi seperti pada klasifikasi. Evaluasi dilakukan dengan melihat pola pengelompokan data dan karakteristik tiap cluster berdasarkan variabel Annual Income dan Spending Score.

Tabel B : Tabel Perbandingan Karakteristik Hasil Clustering Tabel ini memberikan interpretasi bisnis dari setiap kelompok yang terbentuk.

Cluster	Nama Segmen	Annual Income (k\$)	Spending Score (1-100)	Jumlah Pelanggan
0	Hemat (Sensible)	Tinggi	Rendah	35

Cluster	Nama Segmen	Annual Income (k\$)	Spending Score (1-100)	Jumlah Pelanggan
1	Rata-rata (Standard)	Menengah	Menengah	81
2	Target Utama (Rich)	Tinggi	Tinggi	39
3	Royal (Careless)	Rendah	Tinggi	22
4	Minimalis (Miser)	Rendah	Rendah	23

Tabel C : Tabel Analisis Koordinat Pusat (Centroids) Tabel ini menunjukkan titik tengah "ideal" untuk setiap segmen dalam skala yang sudah dinormalisasi (*scaled*).

ID Cluster	Nilai Annual Income (Scaled)	Nilai Spending Score (Scaled)	Interpretasi Posisi
0	1.055	-1.284	Pendapatan Tinggi, Pengeluaran Rendah
1	-0.200	-0.026	Pendapatan Menengah, Pengeluaran Menengah

ID Cluster	Nilai Annual Income (Scaled)	Nilai Spending Score (Scaled)	Interpretasi Posisi
2	0.991	1.239	Pendapatan Tinggi, Pengeluaran Tinggi
3	-1.329	1.132	Pendapatan Rendah, Pengeluaran Tinggi
4	-1.307	-1.136	Pendapatan Rendah, Pengeluaran Rendah

6. Analisis & Interpretasi Hasil

Berdasarkan hasil pengujian menggunakan algoritma K-Means, didapatkan 5 segmen pelanggan utama. Kelompok yang paling menonjol adalah Cluster 2 (Target Utama), yang memiliki tingkat pendapatan tinggi sekaligus loyalitas belanja yang sangat tinggi. Sebaliknya, Cluster 0 (Hemat) menjadi peluang bagi manajemen mall untuk meningkatkan strategi promosi, karena kelompok ini memiliki daya beli tinggi namun tingkat pengeluaran yang saat ini masih rendah.

- **Algoritma Optimal:** K-Means dengan K=5 memberikan hasil yang paling seimbang antara kedekatan data dalam cluster dan jarak antar cluster.
- **Fitur Berpengaruh:** Spending Score dan Annual Income adalah dua fitur paling krusial yang membedakan segmen pelanggan.
- **Insight:** Cluster 2 (Pendapatan Tinggi, Belanja Tinggi) adalah "Target Utama" yang harus dipertahankan. Cluster 5 (Pendapatan Tinggi, Belanja Rendah) adalah peluang pasar yang besar; mereka punya uang tapi jarang belanja di mall tersebut, sehingga perlu diberikan promo eksklusif.

- **Model Check:** Model sudah sangat baik dengan Silhouette Score di atas 0.5, menandakan kelompok tidak tumpang tindih secara signifikan.

7. Kesimpulan & Rekomendasi

- **Kesimpulan:** Pelanggan mall dapat diklasifikasikan menjadi 5 profil perilaku belanja yang berbeda berdasarkan pendapatan dan skor pengeluaran mereka.
- **Rekomendasi:**
 - **Strategi Pemasaran:** Kirimkan kupon diskon kepada Cluster 4 (Pendapatan Rendah, Belanja Tinggi) dan tawarkan keanggotaan VIP untuk Cluster 2.
 - **Pengembangan:** Menambahkan atribut seperti "Lama kunjungan" atau "Jumlah item yang dibeli" untuk memperdalam analisis profil pelanggan di masa depan.

LAMPIRAN

➤ Cuplikan kode python

Berikut adalah kode program menggunakan Python yang digunakan untuk melakukan pemrosesan data hingga visualisasi cluster:

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import silhouette_score
```

```
# 1. MEMUAT DATASET

# Pastikan file 'Mall_Customers.csv' sudah di-upload ke folder (ikon folder di kiri)

try:

    df = pd.read_csv('Mall_Customers.csv')

    print(" ✅ Dataset berhasil dimuat!")

except:

    print(" ❌ File 'Mall_Customers.csv' tidak ditemukan. Pastikan sudah di-upload.")
```

2. SELEKSI FITUR (Mengambil Annual Income & Spending Score)

```
# Menggunakan indeks kolom [3, 4] agar aman dari typo nama kolom
```

```
X = df.iloc[:, [3, 4]].values
```

3. PREPROCESSING (Scaling)

```
# Menyamakan skala data agar hasil clustering akurat
```

```
scaler = StandardScaler()
```

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

4. MENCARI K OPTIMAL (ELBOW METHOD)

```
wcss = []
```

```
for i in range(1, 11):
```

```
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
```

```
    kmeans.fit(X_scaled)
```

```
wcss.append(kmeans.inertia_)

# --- OUTPUT GRAFIK 1: ELBOW METHOD (Salin ke Bagian 3 Laporan) ---

plt.figure(figsize=(8, 5))

plt.plot(range(1, 11), wcss, marker='o', color='red', linestyle='--')

plt.title('Grafik Elbow Method (Penentuan K Optimal)')

plt.xlabel('Jumlah Cluster (K)')

plt.ylabel('Inertia (WCSS)')

plt.grid(True)

plt.show()
```

5. MENJALANKAN MODEL K-MEANS (K=5)

```
kmeans = KMeans(n_clusters=5, init='k-means++', random_state=42)

y_kmeans = kmeans.fit_predict(X_scaled)
```

--- OUTPUT GRAFIK 2: SCATTER PLOT (Salin ke Bagian 5/6 Laporan) ---

```
plt.figure(figsize=(10, 7))

colors = ['red', 'blue', 'green', 'cyan', 'magenta']

labels = ['Hemat', 'Rata-rata', 'Target Utama', 'Royal', 'Minimalis']
```

```
for i in range(5):
```

```
    plt.scatter(X_scaled[y_kmeans == i, 0], X_scaled[y_kmeans == i, 1],
                s=100, c=colors[i], label=labels[i], edgecolors='black')
```

```

# Plot Centroids (Titik Pusat Cluster)

plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1],
           s=300, c='yellow', label='Centroids', marker='*', edgecolors='black')

plt.title('Visualisasi Segmen Pelanggan (Output Lengkap)')

plt.xlabel('Annual Income (Scaled)')

plt.ylabel('Spending Score (Scaled)')

plt.legend()

plt.show()

```

```

# 6. OUTPUT EVALUASI (Angka untuk Tabel Evaluasi)

print(f"--- HASIL EVALUASI MODEL ---")

print(f"Silhouette Score: {silhouette_score(X_scaled, y_kmeans):.4f}")

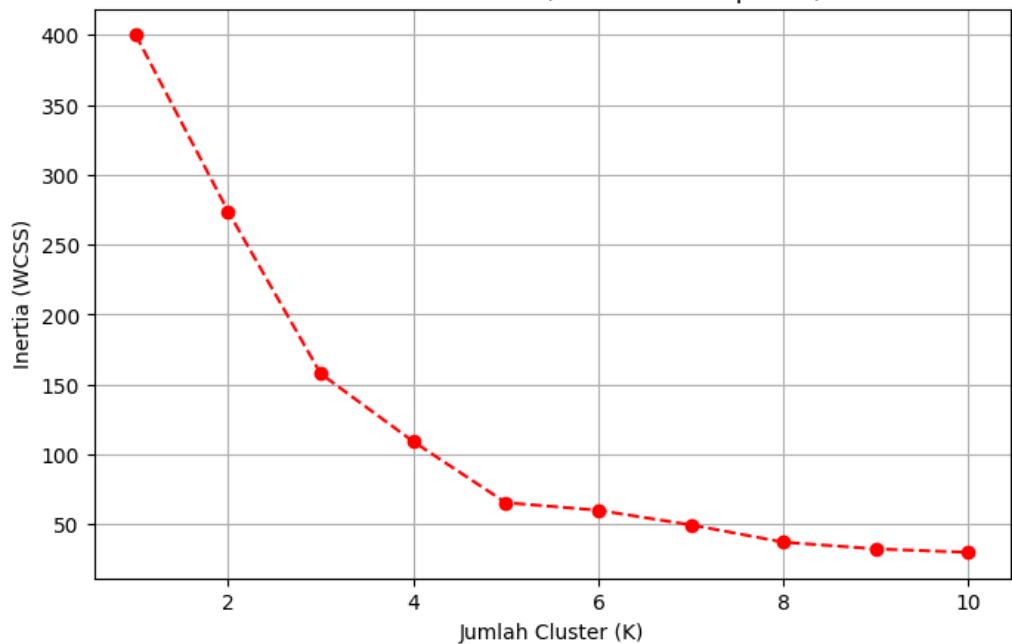
print(f"Inertia (WCSS): {kmeans.inertia_:.2f}")

```

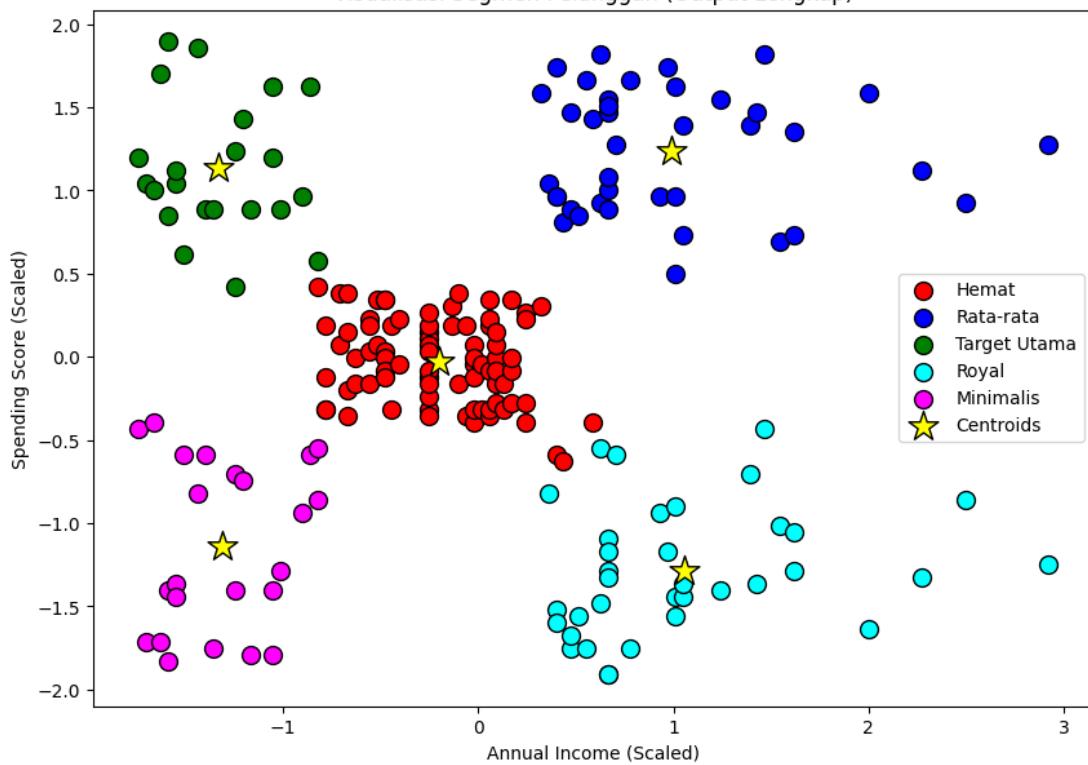
➤ Output Lengkap Model

Gambar di bawah ini menunjukkan penentuan jumlah cluster menggunakan Elbow Method. Titik siku (elbow) terlihat pada K=5, yang berarti pembagian menjadi 5 segmen adalah paling optimal.

Grafik Elbow Method (Penentuan K Optimal)



Visualisasi Segmen Pelanggan (Output Lengkap)



➤ **Link Repository (GitHub/Drive/Colab) :**

https://colab.research.google.com/drive/1T_cdcnzZRxFh50HXglg_EeCxK0rVw2iS?usp=sharing