LAPORAN TUGAS BESAR MACHINE LEARNING

Implementasi Manual K-Means Clustering untuk Menentukan Pola Perilaku Konsumen

DISUSUN OLEH

AISYA MUFIDAH NAJWA (1206230026)
INAYA REVALINA PUTRI MUJANA (120620038)
MOCHAMMAD FERY ARDIANSYAH (1206230044)

PROGRAM STUDI SAINS DATA

FAKULTAS INFORMATIKA

TELKOM UNIVERSITY SURABAYA

2025

1. **Formulasi Masalah**

   Setiap perusahaan yang bergerak di bidang ritel atau layanan konsumen menghadapi tantangan dalam memahami karakteristik dan perilaku pelanggan yang beragam. Tanpa adanya segmentasi yang tepat, strategi pemasaran, penyusunan promosi, dan pengembangan produk dapat menjadi kurang efektif serta tidak tepat sasaran. Salah satu cara untuk memahami perbedaan karakteristik pelanggan adalah dengan melakukan segmentasi berdasarkan atribut-atribut tertentu seperti umur, pendapatan tahunan, dan tingkat pengeluaran.

   Namun, dalam kenyataannya, hubungan antara variabel-variabel tersebut tidak selalu bersifat linear atau intuitif. Misalnya, tidak semua pelanggan dengan pendapatan tinggi memiliki tingkat pengeluaran yang tinggi, dan tidak semua pelanggan muda bersifat konsumtif. Oleh karena itu, diperlukan metode analisis yang mampu mengelompokkan pelanggan ke dalam klaster yang memiliki kemiripan pola perilaku berdasarkan data yang tersedia.

   Permasalahan utama yang akan diselesaikan dalam penelitian ini adalah:

   - Bagaimana mengelompokkan pelanggan ke dalam beberapa segmen atau klaster berdasarkan atribut-atribut numerik (seperti umur, pendapatan tahunan, dan spending score)?
   - Berapa jumlah klaster yang optimal agar hasil segmentasi dapat merepresentasikan pola perilaku pelanggan secara bermakna?
   - Apa karakteristik dari masing-masing klaster yang terbentuk, dan bagaimana insight yang dihasilkan dapat dimanfaatkan untuk mendukung pengambilan keputusan bisnis?

   Penelitian ini juga bertujuan untuk memahami dan mengimplementasikan algoritma K-Means Clustering secara manual tanpa menggunakan library bawaan, guna memperdalam pemahaman mengenai proses iteratif dalam pembentukan klaster, pemilihan centroid, serta evaluasi hasil clustering melalui metrik seperti Silhouette Score dan Davies-Bouldin Index.

   Dengan menggunakan metode unsupervised learning berupa algoritma K-Means Clustering, diharapkan dapat ditemukan pola segmentasi pelanggan yang dapat membantu perusahaan dalam merancang strategi pemasaran yang lebih efektif dan terarah.

2. **Eksplorasi dan Persiapan Data**

   Dataset:

Dataset diunduh dari Kaggle, berisi kolom:

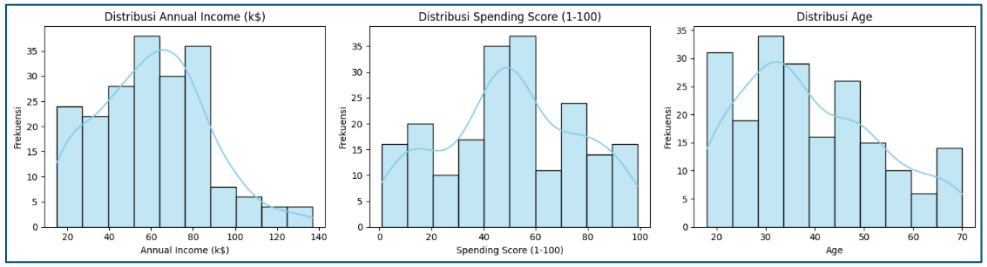
- CustomerID
- Gender
- Age
- Annual Income (k$)
- Spending Score (1–100)

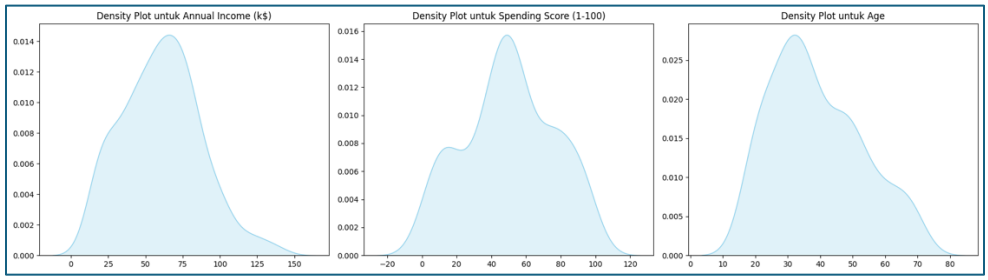
Link dataset: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python

Hanya fitur numerik (Age, Annual Income (k$), dan Spending Score (1-100)) yang digunakan karena relevan untuk clustering.

Langkah eksplorasi dan persiapan data yang dilakukan:

- Menghitung statistik deskriptif membantu lihat distribusi dan skala fitur.
- Memeriksa apakah ada nilai yang kosong dan data duplikat.
- Visualisasi distribusi kolom numerik dengan: histogram dan density untuk melihat distribusi fitur (Age, Income, Spending Score).

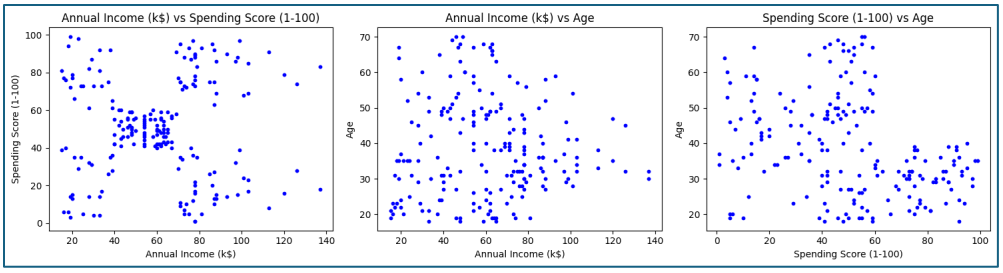




Density Plot menunjukkan distribusi:

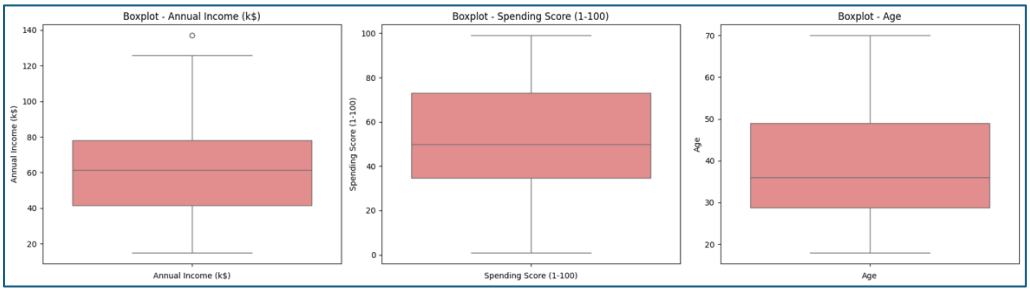
- Age relatif normal, sedikit skewed ke kanan.
- Annual Income terlihat skewed dan memiliki outlier.
- Spending Score hampir uniform, menyebar cukup merata dari 1–100.

- Visualisasi distribusi kolom numerik dengan Scatter plot untuk memeriksa pola hubungan antar variable



Scatter plot antar fitur menunjukkan tidak ada hubungan linear yang jelas antar variabel, yang menegaskan bahwa pendekatan klasterisasi non-linear seperti K-Means cocok.

- Visualisasi Boxplot untuk memeriksa outlier



Minim outlier (hanya ada 1)

Dari eksplorasi dan persiapan kita mendapatkan hasil:

1. Pemilihan fitur sangat berpengaruh terhadap hasil clustering:
   - Kombinasi fitur Annual Income dan Spending Score membentuk pola klaster yang jelas dalam scatterplot.
   - Penambahan fitur Age memberikan dimensi tambahan yang membantu memisahkan segmen berdasarkan siklus hidup pelanggan (misalnya pelajar, dewasa muda, paruh baya).

2. Standarisasi data penting untuk keakuratan model K-Means:
   - Karena K-Means menggunakan jarak Euclidean, fitur dengan skala besar seperti Annual Income bisa mendominasi jika data tidak dinormalisasi.
   - Setelah dilakukan scaling, hasil clustering menjadi lebih seimbang dan representatif.

3. Pemodelan

3.1 Rumus K-Means Clustering

a. Jarak Euclidean

$$d(x, \mu) = \sqrt{\sum_{i=1}^{n} (x_i - \mu_i)^2}$$

b. Update Centeroid

rata-rata dari semua titik yang berada dalam satu klaster:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \epsilon C_j} x_i$$

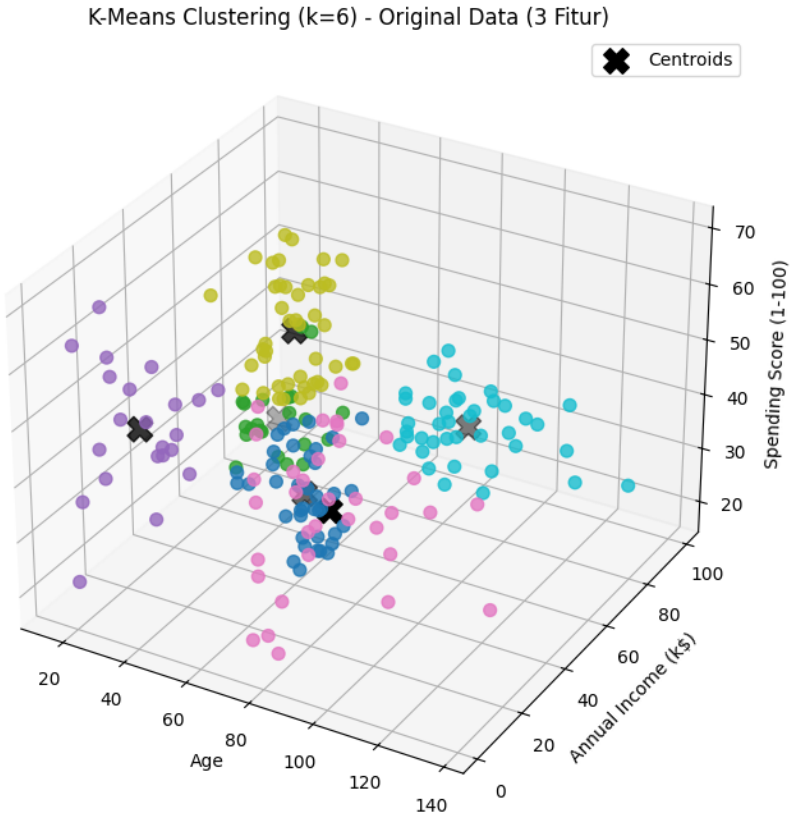
3.2 Implementasi K-means Manual

```python
# Fungsi jarak Euclidean manual defaultdict
def euclidean_distance(p1, p2):
    return math.sqrt(sum((a - b) ** 2 for a, b in zip(p1, p2)))

# Set seed random
def set_seed(seed=42):
    np.random.seed(seed)
    random.seed(seed)

# Inisialisasi centroid acak
def init_centroids(X, k):
    idx = np.random.choice(X.shape[0], k, replace=False)
    return X[idx]

# Assign cluster secara manual dengan euclidean_distance
def assign_clusters(X, centroids):
    labels = []
    for point in X:
        dists = [euclidean_distance(point, centroid) for centroid in centroids]
        labels.append(dists.index(min(dists)))
    return np.array(labels)

# Update centroid berdasarkan rata-rata
def update_centroids(X, labels, k):
    new_centroids = []
    for i in range(k):
        cluster_points = X[labels == i]
        if len(cluster_points) == 0:
            new_centroids.append(X[np.random.randint(len(X))])
        else:
            new_centroids.append(np.mean(cluster_points, axis=0))
    return np.array(new_centroids)

# Fungsi utama K-Means
def kmeans(X, k, max_iter=100, seed=42):
    set_seed(seed)
    centroids = init_centroids(X, k)
    for _ in range(max_iter):
        labels = assign_clusters(X, centroids)
        new_centroids = update_centroids(X, labels, k)
        if np.allclose(centroids, new_centroids):
            break
        centroids = new_centroids
    return labels, centroids
```

Algoritma K-Means manual diimplementasikan dengan beberapa fungsi utama. Proses dimulai dengan inisialisasi centroid secara acak dari data menggunakan fungsi init_centroids. Selanjutnya, fungsi assign_clusters menghitung jarak Euclidean antara setiap data dan centroid, lalu menentukan klaster berdasarkan jarak terdekat. Setelah semua

data memiliki label klaster, fungsi update_centroids akan menghitung ulang posisi centroid berdasarkan rata-rata data dalam tiap klaster. Jika terdapat klaster kosong, centroid-nya akan diisi ulang dengan data acak. Proses ini dikendalikan oleh fungsi kmeans, yang akan melakukan iterasi pengelompokan dan pembaruan centroid hingga konvergen atau mencapai jumlah iterasi maksimum. Hasil akhir berupa label klaster dan posisi centroid terakhir.

3.1 Menggunakan 3 Fitur:



Visualisasi 3D ini menggambarkan hasil klasterisasi berdasarkan tiga fitur utama: umur (Age), pendapatan tahunan (Annual Income), dan skor pengeluaran (Spending Score). Dengan 6 klaster yang terbentuk, terlihat pola segmentasi pelanggan yang cukup kompleks namun informatif. Berikut adalah insight yang dapat diambil:

1. **Klaster usia muda dengan pendapatan rendah namun pengeluaran tinggi**
   Terdapat satu klaster yang terdiri dari individu berusia muda dengan pendapatan tahunan relatif rendah, namun memiliki skor pengeluaran yang tinggi. Ini mengindikasikan adanya kelompok anak muda yang cenderung konsumtif meskipun tidak memiliki penghasilan besar, mungkin karena gaya hidup atau ketergantungan pada orang tua.

2. **Klaster usia menengah dengan pendapatan dan pengeluaran sedang**
   Klaster ini berisi individu yang berada pada usia menengah dengan pendapatan tahunan dan skor pengeluaran yang tidak terlalu ekstrem. Kelompok ini menunjukkan pola konsumsi yang stabil dan moderat, cocok menjadi target untuk penawaran umum atau program loyalitas.

3. **Klaster usia tua dengan pendapatan tinggi dan pengeluaran rendah**
   Kelompok ini cenderung terdiri dari konsumen yang lebih tua, memiliki pendapatan tinggi, namun spending score rendah. Kemungkinan kelompok ini bersifat hemat atau selektif dalam pengeluaran, cocok untuk pendekatan pemasaran berbasis nilai atau manfaat jangka panjang.

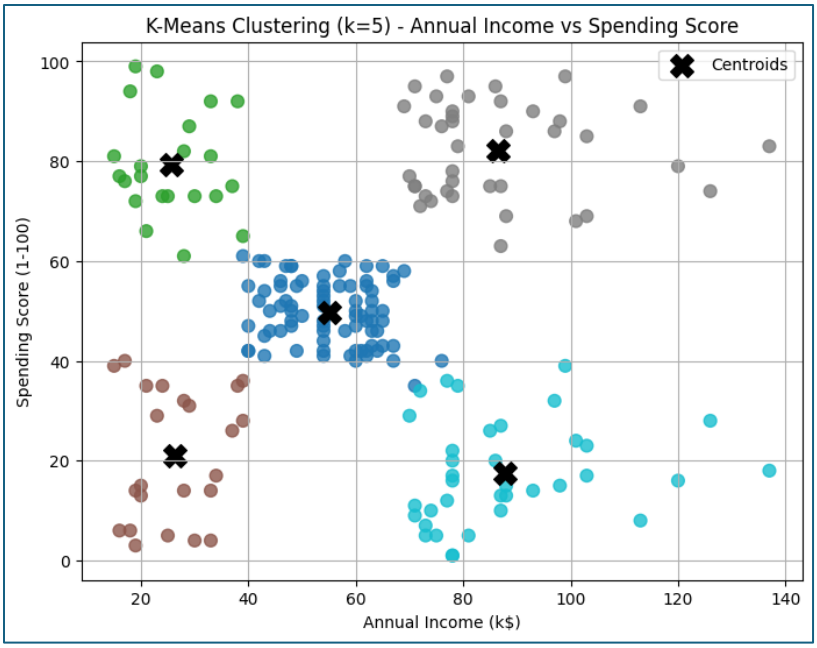
4. **Klaster usia tua dengan pengeluaran tinggi**
   Menariknya, ada juga klaster yang berisi pelanggan usia lebih tua namun dengan spending score tinggi. Ini menunjukkan bahwa tidak semua konsumen tua bersifat hemat, dan masih ada segmen lansia yang aktif dalam berbelanja, mungkin karena kebutuhan gaya hidup atau kebebasan finansial.

5. **Klaster dengan pendapatan menengah tinggi dan pengeluaran sangat rendah**
   Klaster ini berisi konsumen yang memiliki penghasilan cukup tinggi, namun sangat jarang melakukan pembelanjaan. Ini mungkin menunjukkan pelanggan pasif atau mereka yang kurang tertarik terhadap produk/jasa yang ditawarkan saat ini.

6. **Klaster beragam usia dengan perilaku belanja menengah**
   Salah satu klaster terlihat menggabungkan pelanggan dari berbagai usia dan tingkat penghasilan, namun dengan skor pengeluaran menengah. Ini bisa jadi merupakan klaster yang paling heterogen dan mewakili konsumen umum.

   Hasil klasterisasi ini menunjukkan bahwa perilaku belanja tidak hanya dipengaruhi oleh pendapatan, tapi juga oleh faktor usia dan mungkin preferensi atau gaya hidup. Segmentasi semacam ini sangat penting dalam merancang strategi pemasaran yang terpersonalisasi. Misalnya, kelompok muda konsumtif bisa ditarget dengan promosi intensif dan media sosial, sementara kelompok usia tua berpendapatan tinggi bisa ditawarkan produk premium berbasis kualitas.

3.2 Menggunakan 2 Fitur:



Hasil klasterisasi pada fitur Annual Income vs Spending Score menghasilkan temuan yang cukup menarik dan memberikan gambaran segmentasi konsumen yang jelas, meskipun hubungan antara pendapatan dan pengeluaran tidak linear. Beberapa pola penting yang dapat diamati:

1. **Pendapatan Rendah** (sekitar 15–40 ribu dolar):
   Terdapat dua klaster yang kontras:
   - Klaster 1: Konsumen dengan spending score rendah (sekitar 0–40), mengindikasikan frekuensi atau intensitas belanja yang rendah.
   - Klaster 2: Konsumen dengan spending score tinggi (sekitar 60–100), menunjukkan kelompok yang berbelanja secara aktif meskipun memiliki pendapatan rendah.

   Tidak ditemukan klaster dengan spending score menengah pada segmen ini. Hal ini mengindikasikan perilaku belanja yang cenderung ekstrem pada kelompok berpendapatan rendah.

2. **Pendapatan Menengah** (sekitar 40–70 ribu dolar):
   - Klaster 3: Seluruh konsumen dalam segmen ini terkonsentrasi dalam satu klaster, dengan spending score menengah (sekitar 40–60). Ini mengisyaratkan bahwa kelompok berpendapatan menengah memiliki perilaku belanja yang lebih moderat dan homogen.

3. **Pendapatan Tinggi** (sekitar 70–140 ribu dolar):
   Terbentuk dua klaster berbeda:

- Klaster 4: Klaster dengan spending score rendah (sekitar 0–40), mengindikasikan adanya konsumen berpendapatan tinggi namun cenderung tidak menghabiskan banyak untuk konsumsi.
- Klaster 5: Klaster dengan spending score tinggi (sekitar 60–100), menunjukkan kelompok affluent spender.
  Seperti pada segmen pendapatan rendah, tidak ada klaster dengan spending score menengah, sehingga perilaku belanja konsumen berpendapatan tinggi juga tampak terbagi secara ekstrem.

Distribusi klaster menunjukkan bahwa baik konsumen dengan pendapatan rendah maupun tinggi memiliki perilaku belanja yang bervariasi (tinggi atau rendah), sementara konsumen dengan pendapatan menengah cenderung memiliki perilaku belanja yang seragam dan moderat. Hal ini bisa menjadi dasar segmentasi pemasaran yang lebih tepat sasaran, misalnya dengan menawarkan promosi intensif pada kelompok spender aktif, dan pendekatan personal untuk kelompok spender pasif.

4. Evaluasi

Untuk mengetahui pengaruh jumlah fitur terhadap hasil clustering, dilakukan evaluasi menggunakan dua skenario:

1. **Menggunakan 3 Fitur**
   - Fitur: Age, Annual Income (k$), dan Spending Score (1-100)
   - Silhouette Score: 0.4679
   - Davies-Bouldin Index: 0.7542
2. **Menggunakan 2 Fitur**
   - Fitur: Annual Income (k$) dan Spending Score (1-100) saja
   - Silhouette Score: 0.5646
   - Davies-Bouldin Index: 0.5711

Hasil menunjukkan bahwa menggunakan 2 fitur memberikan performa clustering yang lebih baik dibandingkan 3 fitur. Hal ini ditunjukkan oleh:

- Silhouette Score yang lebih tinggi (0.5646), yang berarti objek dalam klaster lebih dekat satu sama lain dan lebih jauh dari klaster lainnya.
- DBI yang lebih rendah (0.5711), yang mengindikasikan pemisahan klaster yang lebih baik dan kompak.

Penggunaan 2 fitur (Annual Income dan Spending Score) lebih optimal dalam membentuk klaster yang jelas dan terpisah, dibandingkan dengan menambahkan fitur Age.

Penambahan fitur yang tidak relevan atau tidak berkontribusi signifikan pada pemisahan klaster justru dapat menurunkan kualitas hasil clustering.

5. Eksperimen

Kami melakukan beberapa eksperimen clustering dengan:

- 2 fitur: Annual Income & Spending Score
- 3 fitur: Age, Annual Income, Spending Score. Lalu membandingkan hasilnya.
- Membandingkan hasil clustering dengan dan tanpa standarisasi/normalisasi data.

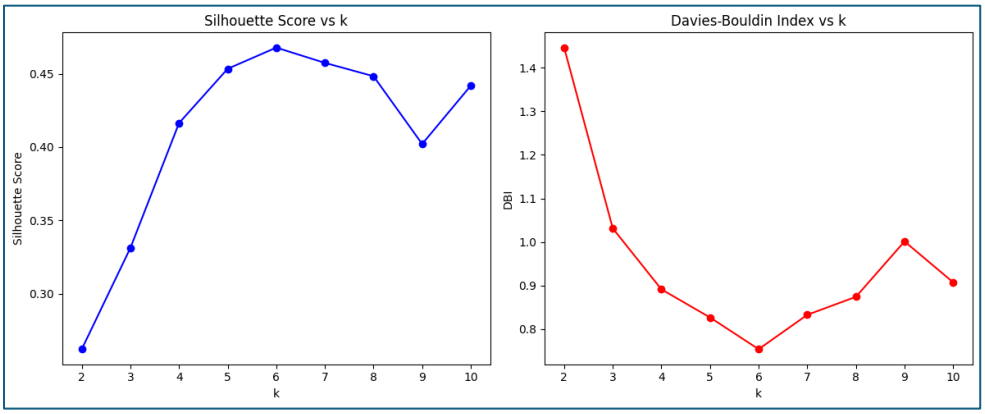
Alasan:

- Scaling: K-Means menghitung jarak Euclidean antar data. Jika fitur memiliki skala berbeda, fitur dengan skala besar (misalnya Annual Income) akan mendominasi perhitungan jarak. Ini bisa menyebabkan pembentukan cluster yang bias.
- Kombinasi Fitur:
  - Menggunakan ketiga fitur untuk melihat apakah memberikan cluster yang signifikan.
  - Mencoba 2 fitur: Income & Score dipilih karena ada pola cluster yang jelas. (di scatterplot antar fitur)

Kami melakukan standarisasi karena fitur-fitur tersebut cukup simetris dan tidak ekstrem secara distribusi, dan standardisasi membuat kontribusi tiap fitur setara terhadap perhitungan jarak.

Untuk menentukan jumlah klaster yang optimal serta mengevaluasi performa hasil clustering, digunakan dua metrik evaluasi yaitu Silhouette Score dan Davies-Bouldin Index (DBI). Evaluasi dilakukan terhadap tiga versi data:

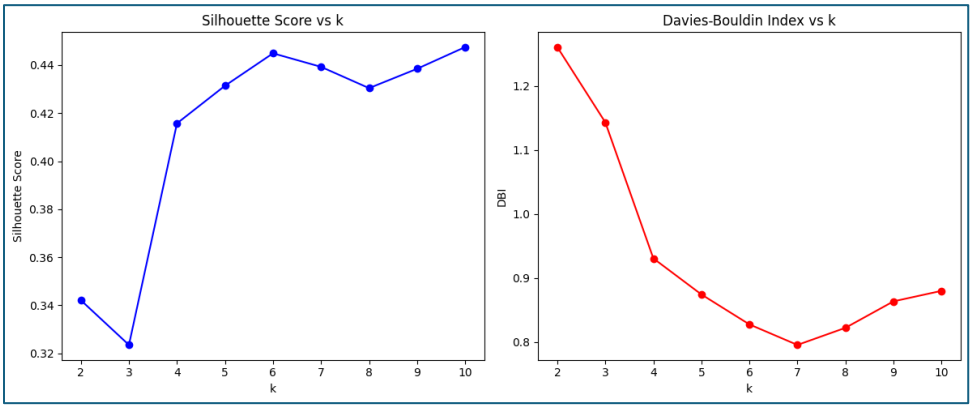
1. Data Asli (tanpa praproses)



k=2 → Silhouette=0.262, DBI=1.4458

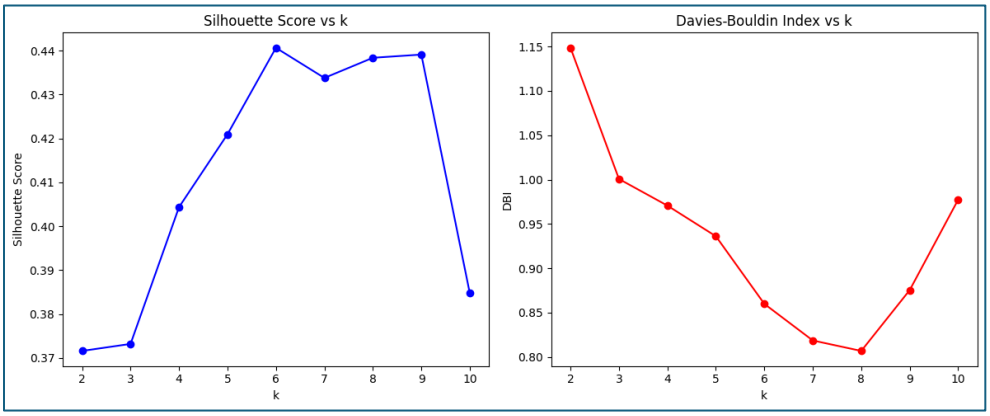
k=3 → Silhouette=0.3312, DBI=1.0313
k=4 → Silhouette=0.4164, DBI=0.8915
k=5 → Silhouette=0.4534, DBI=0.827
k=6 → Silhouette=0.4679, DBI=0.7542
k=7 → Silhouette=0.4575, DBI=0.8331
k=8 → Silhouette=0.4483, DBI=0.8743
k=9 → Silhouette=0.4021, DBI=1.0014
k=10 → Silhouette=0.4419, DBI=0.9076

2. Data Distandarisasi (StandardScaler)



k=2 → Silhouette=0.3421, DBI=1.2607
k=3 → Silhouette=0.3235, DBI=1.1431
k=4 → Silhouette=0.4157, DBI=0.9308
k=5 → Silhouette=0.4314, DBI=0.8746
k=6 → Silhouette=0.4449, DBI=0.8277
k=7 → Silhouette=0.4393, DBI=0.7957
k=8 → Silhouette=0.4304, DBI=0.8224
k=9 → Silhouette=0.4385, DBI=0.8637
k=10 → Silhouette=0.4475, DBI=0.8801

3. Data Dinormalisasi (Min-Max Scaling)



k=2 → Silhouette=0.3716, DBI=1.1485
k=3 → Silhouette=0.3732, DBI=1.0007
k=4 → Silhouette=0.4043, DBI=0.9706
k=5 → Silhouette=0.4209, DBI=0.9361
k=6 → Silhouette=0.4406, DBI=0.8599
k=7 → Silhouette=0.4338, DBI=0.8186
k=8 → Silhouette=0.4384, DBI=0.8069
k=9 → Silhouette=0.4391, DBI=0.8755
k=10 → Silhouette=0.3847, DBI=0.9774

Hasilnya:

- Data Asli memberikan hasil yang paling baik, ditunjukkan oleh nilai Silhouette Score tertinggi sebesar 0.4697 dan DBI terendah sebesar 0.7542 pada k = 6. Ini menunjukkan pembentukan klaster yang paling kompak dan terpisah dengan baik dibandingkan dua versi data lainnya.
- Data Distandarisasi memiliki performa yang cukup baik, tetapi nilai DBI dan Silhouette-nya tidak melebihi data asli. Nilai Silhouette tertinggi hanya 0.4475, dan DBI terendah 0.7957.
- Data Dinormalisasi menunjukkan performa paling rendah di antara ketiganya, dengan Silhouette tertinggi 0.4406 dan DBI terbaik 0.7925.

Proses clustering bekerja paling optimal pada data asli tanpa praproses, di mana pola antar fitur tetap terjaga dan menghasilkan pemisahan klaster yang paling jelas dan stabil menurut kedua metrik evaluasi tersebut.

6. Kesimpulan

Beberapa eksperimen dilakukan dengan membandingkan performa clustering pada data asli (tanpa transformasi), data yang telah dinormalisasi (Min-Max Scaling), dan data yang telah distandarisasi (Z-Score). Berdasarkan hasil evaluasi:

Untuk data asli dengan dua fitur (Annual Income dan Spending Score), nilai silhouette score dan DBI terbaik diperoleh saat k = 5, yaitu Silhouette = 0.5646 dan DBI = 0.5711. Ini menunjukkan pemisahan klaster yang baik dan kompak.

Untuk data asli dengan tiga fitur (Age, Annual Income, Spending Score), nilai terbaik diperoleh saat k = 6, dengan Silhouette = 0.4679 dan DBI = 0.7542. Hasil ini menunjukkan bahwa penambahan fitur age memberikan dimensi baru dalam segmentasi yang tetap memberikan struktur klaster yang cukup kuat.

Meskipun normalisasi dan standarisasi membantu mengurangi skala ketimpangan antar fitur, hasil eksperimen menunjukkan bahwa data asli memberikan performa yang lebih baik secara konsisten, terutama pada kombinasi metrik evaluasi tersebut.

Dengan demikian, dapat disimpulkan bahwa formulasi masalah dalam segmentasi pelanggan ini berhasil dijawab melalui pendekatan clustering dengan K-Means, menggunakan data asli tanpa transformasi, dengan pemilihan jumlah klaster berdasarkan kombinasi nilai Silhouette Score dan DBI terbaik. Segmentasi ini tidak hanya memberikan wawasan mendalam terhadap perilaku pelanggan, tetapi juga membuka peluang untuk merancang strategi pemasaran yang lebih terpersonalisasi dan efektif.

Lampiran

Dataset:
https://www.google.com/url?q=https%3A%2F%2Fwww.kaggle.com%2Fdatasets%2Fvjchoudhary7%2Fcustomer-segmentation-tutorial-in-python
Colab:https://colab.research.google.com/drive/1CfH0Abga-N71Z03TlhasC7avpwmlWERD?usp=sharing
Youtube: https://youtu.be/Dcd_aJCWVQE
Github: GitHub - FeryArdi/ML