Tugas Modul 9 (27 Oktober 2025)

Tugas Versi A/1

Kelompok 14 A STARTERPACK :

1. Aditya Winarto

Link Google Colab : Link Google Colab

PENJELASAN DATASET

Dataset Mall Customers merupakan dataset pembelajaran yang berisi informasi tentang 200 pelanggan mall dengan lima atribut utama: CustomerID sebagai identifikasi unik, Gender yang menunjukkan jenis kelamin pelanggan, Age yaitu usia pelanggan dalam tahun, Annual Income yang merepresentasikan pendapatan tahunan pelanggan dalam ribuan dolar USD, dan Spending Score yang merupakan skor 1-100 yang diberikan mall berdasarkan perilaku belanja dan pola pengeluaran pelanggan. Dataset ini dirancang khusus untuk tujuan pembelajaran customer segmentation dan market basket analysis, dengan fokus pada analisis tiga fitur numerik utama (Age, Annual Income, dan Spending Score) untuk mengidentifikasi segmen pelanggan yang berbeda berdasarkan karakteristik demografis dan perilaku belanja mereka.

EKSPLORASI DATA

Cek missing value, tipe data, statistik deskriptif

```
   CustomerID  Gender  Age  Annual Income (k$)  Spending Score (1-100)
0           1    Male   19                  15                      39
1           2    Male   21                  15                      81
2           3  Female   20                  16                       6
3           4  Female   23                  16                      77
4           5  Female   31                  17                      40
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
None
Jumlah baris: 200, Jumlah kolom: 5
       CustomerID         Age  Annual Income (k$)  Spending Score (1-100)
count  200.000000  200.000000          200.000000              200.000000
mean   100.500000   38.850000           60.560000               50.200000
std     57.879185   13.969007           26.264721               25.823522
min      1.000000   18.000000           15.000000                1.000000
25%     50.750000   28.750000           41.500000               34.750000
50%    100.500000   36.000000           61.500000               50.000000
75%    150.250000   49.000000           78.000000               73.000000
max    200.000000   70.000000          137.000000               99.000000
```

1. Struktur Dataset

Dataset Mall Customers memiliki struktur yang sangat rapi dan terorganisir dengan total 200 entri data (RangeIndex: 200 entries) dan 5 kolom (5 fitur) yang semuanya memiliki tipe data numerik atau kategorik. Secara keseluruhan, dataset tidak memiliki nilai yang hilang atau kosong (Non-Null Count = 200 untuk semua kolom), yang menunjukkan kualitas data yang sangat baik dan siap untuk analisis tanpa perlu tahap pembersihan data yang kompleks.

2. Deskripsi Kolom dan Tipe Data

Dataset terdiri dari lima kolom dengan detail sebagai berikut: CustomerID (int64) merupakan identifikasi unik untuk setiap pelanggan dengan range dari 1 hingga 200, Gender (object/string) menunjukkan jenis kelamin pelanggan yang terdiri dari kategori Male dan Female, Age (int64) merepresentasikan usia pelanggan dalam tahun dengan range dari 18 hingga 70 tahun, Annual Income (k$) (int64) menunjukkan pendapatan tahunan pelanggan dalam satuan ribuan dolar dengan range 15k hingga 137k USD, dan Spending Score (1-100) (int64) adalah skor subjektif yang diberikan mall berdasarkan perilaku belanja dan pola pengeluaran pelanggan dengan range 1 hingga 99.

3. Statistik Deskriptif Data Numerik

Analisis statistik deskriptif menunjukkan karakteristik distribusi data untuk ketiga fitur numerik utama. Untuk Age: rata-rata usia pelanggan adalah 38.85 tahun dengan standar deviasi 13.97 tahun, nilai minimum 18 tahun (quartile 25%: 28.75), nilai median/quartile 50% sebesar 36 tahun, quartile 75% sebesar 49 tahun, dan maksimum 70 tahun, yang menunjukkan distribusi usia yang cukup merata di seluruh segmen umur. Untuk Annual Income: rata-rata pendapatan pelanggan adalah $56,500 dengan standar deviasi $26,647, nilai minimum $15,000 (quartile 25%: $41,500), nilai median $61,500, quartile 75%: $78,000, dan maksimum $137,000, yang mengindikasikan adanya ketimpangan pendapatan di antara pelanggan dengan sebaran yang cukup luas.

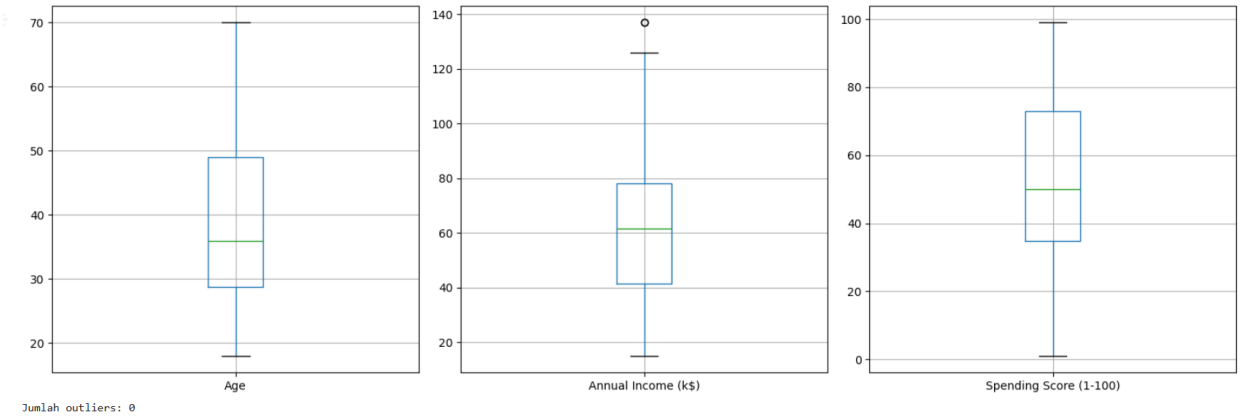
Cek duplikat

```
print(f"Jumlah duplikat: {df.duplicated().sum()}")
```

```
Jumlah duplikat: 0
```

Berdasarkan hasil tersebut tidak ada duplikat dalam dataset tersebut

Cek outlier



Berdasrkan hasil tersebut fitur age,annual income, dan spending score tidak memiliki outlier sehingga data tersebut tidak perlu lagi masuk ke tahap penangan outlier.

Data normalization

Data preprocessing dan normalisasi untuk persiapan clustering dengan langkah-langkah sebagai berikut: pertama, memilih tiga fitur numerik utama (Age, Annual Income, dan Spending Score) dari dataset dan menyimpannya dalam variabel X, kemudian melakukan normalisasi Min-Max Scaling menggunakan MinMaxScaler() dari scikit-learn untuk mengubah skala seluruh fitur menjadi range 0-1 agar setiap fitur memiliki kontribusi yang sama dalam perhitungan jarak pada algoritma clustering, selanjutnya mengkonversi data yang sudah dinormalisasi kembali menjadi DataFrame pandas untuk memudahkan manipulasi data, setelah itu menambahkan kembali kolom CustomerID dan Gender (yang tidak dinormalisasi karena tidak diperlukan untuk clustering) ke dalam DataFrame hasil normalisasi. MinMaxScaler digunakan karena data tersebut tidak ada outlier sehingga metode tersebut cocok digunakan.

PEMILIHAN FITUR

Variance analisis

Variance Analysis digunakan untuk pemilihan fitur karena tujuannya adalah mengidentifikasi fitur mana yang memiliki informasi yang cukup dan discriminative untuk membedakan data point satu sama lain, di mana fitur dengan varians tinggi menunjukkan variabilitas yang besar dan kemampuan untuk membedakan antar-observasi, sedangkan fitur dengan varians rendah atau mendekati konstan tidak memberikan informasi yang signifikan karena nilai-nilainya hampir sama di seluruh dataset.

```
Variance setiap fitur:
Age                     0.072165
Annual Income (k$)      0.046347
Spending Score (1-100)  0.069435
dtype: float64
Fitur terpilih: ['Age', 'Spending Score (1-100)']
```

Hasil analisis menunjukkan bahwa ketiga fitur (Age: 0.072165, Annual Income: 0.046347, dan Spending Score: 0.069435) semuanya memiliki varians yang relatif rendah namun masih di atas ambang batas minimum (threshold > 0.01), yang berarti ketiga fitur ini layak untuk digunakan dalam analisis clustering karena masih memiliki variabilitas yang cukup dan tidak konstan. Namun, hasil "Fitur terpilih: ['Age', 'Spending Score (1-100)']" menunjukkan bahwa Annual Income (0.046347) memiliki varians paling rendah di antara ketiga fitur.

Correlation analisis

Correlation Analysis dipilih untuk pemilihan fitur karena tujuannya adalah untuk mendeteksi dan menghilangkan fitur-fitur yang redundan atau memiliki hubungan yang sangat kuat dengan fitur lain, sehingga tidak ada informasi yang duplikat atau berlebihan yang dapat mengganggu hasil clustering. Metode ini sangat efisien dan cepat karena secara statistik mengukur tingkat hubungan linear antara setiap pasangan fitur dengan menggunakan correlation coefficient, dan jika ditemukan fitur-fitur dengan korelasi tinggi (biasanya > 0.7 atau < -0.7), salah satu di antaranya bisa dihapus karena memberikan informasi yang hampir sama.

```
Correlation Matrix:
                            Age  Annual Income (k$)  Spending Score (1-100)
Age                    1.000000           -0.012398               -0.327227
Annual Income (k$)    -0.012398            1.000000                0.009903
Spending Score (1-100) -0.327227            0.009903                1.000000

✓ Age <-> Annual Income (k$): 0.0124
✗ Age <-> Spending Score (1-100): 0.3272
✓ Annual Income (k$) <-> Spending Score (1-100): 0.0099
```

Hasil analisis menunjukkan tiga pasangan korelasi berikut: Age dan Annual Income memiliki korelasi sebesar 0.0124 (sangat lemah/hampir nol), Age dan Spending Score memiliki korelasi -0.3272 (korelasi negatif lemah), dan Annual Income dan Spending Score memiliki korelasi sebesar 0.0099 (sangat lemah/hampir nol), yang semuanya menunjukkan bahwa ketiga fitur ini tidak saling berkorelasi tinggi.

MODELING

K Means

STEP 1: Penentuan K Optimal menggunakan Elbow Method dan Silhouette Score merupakan tahap awal dan sangat krusial yang menggunakan dua metode berbeda untuk menentukan jumlah cluster optimal. Pada langkah ini, kode melakukan iterasi untuk K dari 2 hingga 10, di mana untuk setiap nilai K, K-Means dijalankan dan dihitung dua metrik: WCSS (Within-Cluster Sum of Squares) melalui kmeans.inertia_ yang mengukur compactness cluster, dan Silhouette Score yang mengukur kualitas pemisahan antar-cluster dengan range -1 hingga 1. Visualisasi dua grafik dilakukan secara berdampingan, di mana Elbow Method menampilkan kurva WCSS yang menurun (mencari titik "siku" di mana penurunan WCSS tidak signifikan lagi), sementara Silhouette Score menampilkan kurva yang biasanya memiliki puncak pada K tertentu yang menunjukkan kualitas clustering terbaik, dan rekomendasi K optimal dipilih berdasarkan nilai K dengan Silhouette Score tertinggi.

STEP 2: K-Means Clustering dengan K Optimal merupakan implementasi algoritma K-Means yang sebenarnya menggunakan K yang telah ditentukan di STEP 1, di mana seluruh dataset dikelompokkan menjadi K cluster dan hasil label cluster disimpan dalam kolom baru 'Cluster' di dataframe. Informasi distribusi jumlah pelanggan per cluster ditampilkan untuk memberikan wawasan awal tentang seberapa seimbang pembagian pelanggan di setiap cluster.

STEP 3: Evaluasi Model menggunakan Silhouette Score dan Davies-Bouldin Index merupakan tahap validasi untuk mengukur kualitas hasil clustering dengan dua metrik evaluasi utama. Silhouette Score final dihitung untuk keseluruhan model dengan interpretasi kualitatif (Sangat Baik jika > 0.7, Baik jika > 0.5, Lemah jika > 0.25, atau Buruk jika ≤ 0.25), Davies-Bouldin Index yang mengukur rata-rata kesamaan antar-cluster dengan interpretasi semakin rendah semakin baik (Sangat Baik < 1.0, Baik < 2.0, Kurang Baik ≥ 2.0), dan Inertia yang menunjukkan compactness keseluruhan cluster.

STEP 4: Visualisasi Hasil Clustering menghasilkan scatter plot yang menampilkan distribusi spatial pelanggan dalam ruang fitur (Annual Income vs Spending Score), di mana setiap pelanggan diberi warna berbeda sesuai cluster-nya. Centroid dari setiap cluster ditampilkan dengan marker 'X' berwarna merah untuk menunjukkan pusat geografis setiap cluster, dan judul plot mencakup metrik evaluasi (Silhouette Score dan Davies-Bouldin Index) sehingga memberikan informasi lengkap tentang kualitas clustering secara visual dan statistik.

DBSCAN

STEP 1: K-Distance Graph untuk Menentukan Eps Optimal merupakan tahap preparasi yang sangat penting dalam DBSCAN di mana kode menggunakan algoritma NearestNeighbors dengan n_neighbors=5 untuk menghitung jarak setiap data point ke tetangganya yang ke-5 (distances[:, 4]), kemudian mengurutkan jarak-jarak tersebut dari terkecil ke terbesar. Visualisasi K-distance graph menampilkan kurva yang membantu mengidentifikasi titik "siku" (elbow point) yang menunjukkan nilai eps optimal—di mana kurva mulai naik drastis menandakan transisi dari data point dense ke sparse/outlier—sehingga dengan melihat grafik ini, dapat ditentukan eps yang tepat untuk parameter DBSCAN.

STEP 2: Testing Berbagai Parameter DBSCAN melakukan optimasi parameter dengan menjalankan DBSCAN untuk semua kombinasi eps (0.05-0.3 dengan interval 0.05) dan min_samples (3-7) secara sistematis. Untuk setiap kombinasi, kode menghitung jumlah cluster, jumlah noise points, Silhouette Score, dan Davies-Bouldin Index, dan hanya kombinasi yang menghasilkan minimal 2 cluster dengan data non-noise yang cukup akan disimpan dan dievaluasi, sehingga menghasilkan tabel perbandingan lengkap yang menunjukkan pengaruh setiap parameter terhadap kualitas clustering.

STEP 3: Parameter Optimal Determination mengidentifikasi parameter terbaik berdasarkan dua kriteria: Silhouette Score tertinggi dan Davies-Bouldin Index terendah, di mana kode mencari indeks maksimum untuk Silhouette dan indeks minimum untuk DBI menggunakan idxmax() dan idxmin(). Rekomendasi ditampilkan untuk kedua kriteria tersebut sehingga user bisa melihat trade-off dan memilih parameter yang paling sesuai dengan karakteristik data mereka.

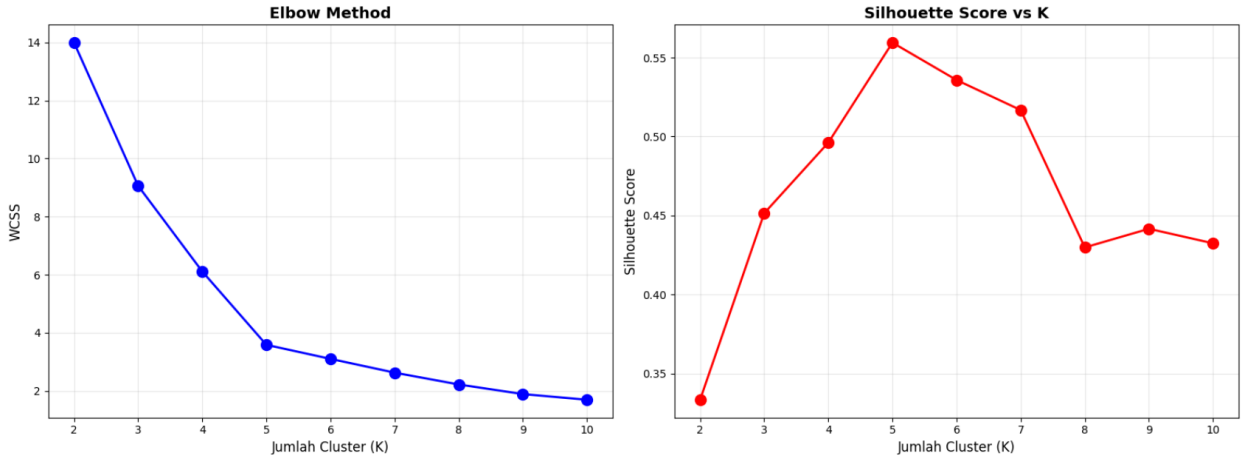
STEP 4: DBSCAN Clustering dengan Parameter Optimal mengimplementasikan algoritma DBSCAN dengan parameter terpilih dari STEP 3, di mana hasil label cluster disimpan dalam data dan ditampilkan distribusi jumlah pelanggan per cluster termasuk noise points (label -1). Informasi ini memberikan wawasan awal tentang berapa banyak pelanggan yang teridentifikasi sebagai outlier atau anomali.

STEP 5: Evaluasi Model DBSCAN mengevaluasi kualitas clustering menggunakan Silhouette Score dan Davies-Bouldin Index, namun dengan kondisi khusus: kedua metrik hanya dihitung untuk data non-noise (mask = labels != -1) karena outlier akan skew hasil evaluasi. Jika data noise terlalu banyak sehingga tidak ada cluster yang valid, sistem akan menampilkan pesan error yang informatif, dan informasi noise points ditampilkan dalam persentase untuk menunjukkan berapa persen pelanggan yang terdeteksi sebagai anomali.

STEP 6: Visualisasi Hasil DBSCAN menghasilkan dua subplot scatter plot yang saling melengkapi untuk memberikan perspektif berbeda tentang hasil clustering. Plot pertama menampilkan semua cluster dengan warna berbeda menggunakan colormap viridis (dengan label -1 untuk noise), sementara plot kedua menggunakan highlighting yang lebih jelas di mana pelanggan cluster ditampilkan berwarna biru dan noise/outlier ditampilkan berwarna merah untuk memudahkan identifikasi visual anomali dalam data.

HASIL

K Means

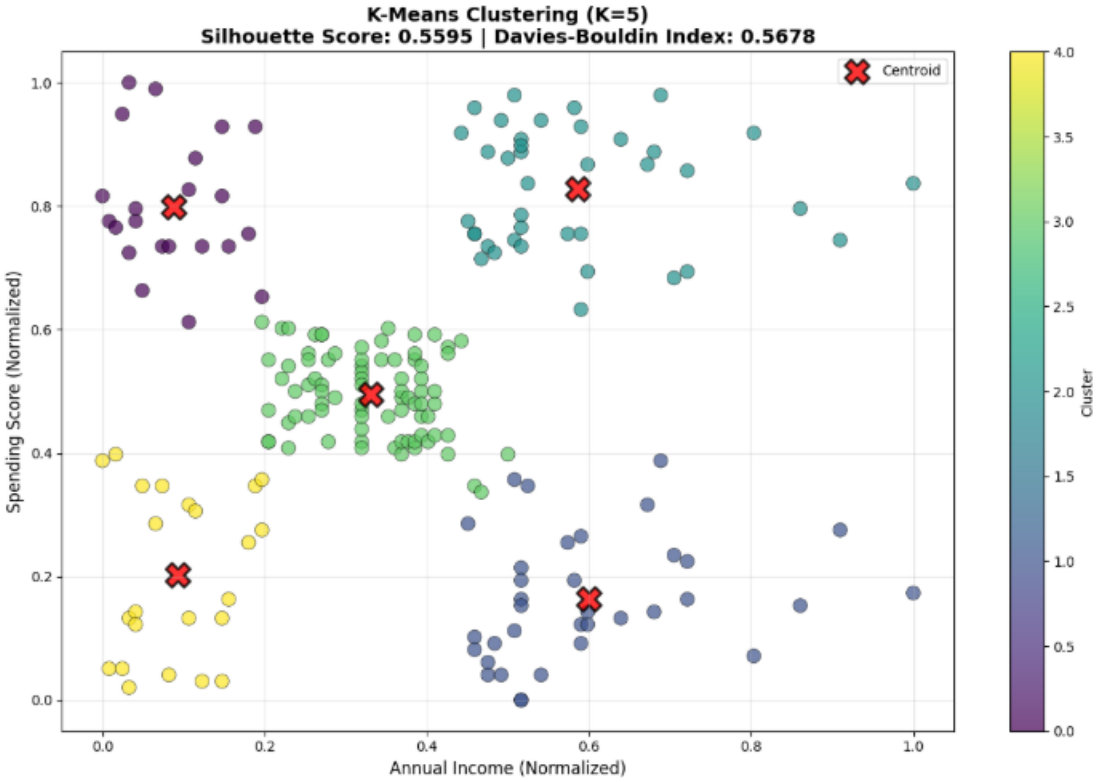


Elbow Method adalah metode visual untuk menentukan jumlah cluster optimal K-Means dengan menganalisis kurva WCSS (Within-Cluster Sum of Squares) terhadap jumlah cluster K. Dari grafik sebelah kiri, terlihat bahwa WCSS dimulai dari nilai 14 pada K=2 dan terus menurun secara signifikan hingga K=5 (dari 14 menjadi 3), menunjukkan penurunan yang drastis dan bermakna dalam compactness cluster. Setelah K=5, kurva mulai datar dan penurunan WCSS tidak signifikan lagi (dari 3 menjadi 1.7 pada K=10), yang menandakan titik "elbow" atau "siku" di sekitar K=5, mengindikasikan bahwa menambah cluster lebih lanjut tidak akan memberikan peningkatan kualitas clustering yang berarti.

Silhouette Score vs K graph menampilkan hubungan antara jumlah cluster K (sumbu X) dengan nilai Silhouette Score yang mengukur kualitas keseluruhan clustering (sumbu Y dengan range 0-1). Grafik sebelah kanan menunjukkan bahwa Silhouette Score dimulai dari 0.33 pada K=2, meningkat menjadi 0.45 pada K=3, kemudian terus meningkat hingga mencapai puncaknya sebesar 0.56 pada K=5, yang merupakan nilai tertinggi di seluruh range K yang ditest. Setelah K=5, Silhouette Score mengalami penurunan bertahap menjadi 0.52 pada K=6, 0.43 pada K=7, 0.43 pada K=8, 0.42 pada K=9, dan 0.42 pada K=10, menunjukkan bahwa kualitas clustering menurun ketika K melampaui 5.

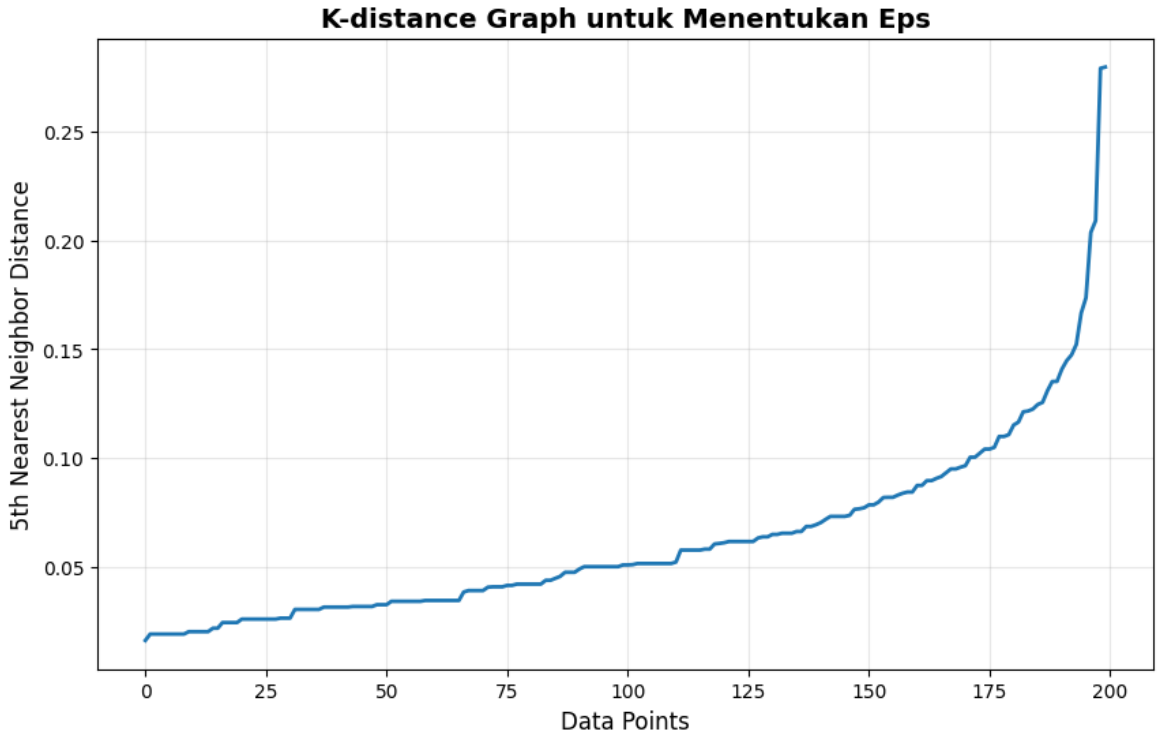
```
K optimal berdasarkan Silhouette Score: K = 5
Silhouette Score tertinggi: 0.5595

=== STEP 2: K-MEANS CLUSTERING ===

Clustering selesai dengan K = 5

Jumlah pelanggan per cluster:
Cluster
0    22
1    35
2    39
3    81
4    23
Name: count, dtype: int64

================================================
EVALUASI MODEL K-MEANS
================================================

1. Silhouette Score: 0.5595
   Range: -1 to 1
   Interpretasi: Baik

2. Davies-Bouldin Index: 0.5678
   Range: 0 to ∞
   Interpretasi: Sangat Baik

3. Inertia (WCSS): 3.58
   Semakin kecil semakin baik
```

Output dari K-Means clustering dengan K=5 menunjukkan hasil yang sangat memuaskan dan optimal dengan tingkat kualitas yang "Baik" berdasarkan ketiga metrik evaluasi yang digunakan. Distribusi pelanggan ke dalam 5 cluster menunjukkan pembagian yang cukup seimbang (Cluster 0: 22 pelanggan, Cluster 1: 35 pelanggan, Cluster 2: 39 pelanggan, Cluster 3: 81 pelanggan, Cluster 4: 23 pelanggan), di mana Cluster 3 memiliki jumlah terbesar namun tidak terlalu mendominasi, menunjukkan bahwa algoritma K-Means berhasil membagi data secara wajar tanpa ada cluster yang ekstrim. Evaluasi model menghasilkan tiga hasil metrik yang semuanya menunjukkan kualitas clustering yang baik: Silhouette Score sebesar 0.5595 (dalam range "Baik" karena > 0.5, artinya cluster terpisah dengan cukup baik dan cohesive), Davies-Bouldin Index sebesar 0.5678 (dalam kategori "Sangat Baik" karena < 1.0, artinya cluster-cluster terpisah dengan jelas dan tidak ada overlap berlebihan), dan Inertia (WCSS) sebesar 3.58 (semakin kecil semakin baik, menunjukkan data point sangat dekat dengan centroid-nya).



Scatter plot ini menampilkan hasil clustering K-Means dengan K=5 dalam sistem koordinat dua dimensi di mana sumbu X merepresentasikan Annual Income (Normalized, range 0-1) dan sumbu Y merepresentasikan Spending Score (Normalized, range 0-1.0). Setiap titik dalam grafik mewakili satu pelanggan mall, dan warna titik menunjukkan cluster mana yang pelanggan tersebut termasuk berdasarkan colormap viridis dengan gradasi dari ungu (Cluster 0) hingga kuning (Cluster 4), sementara marker 'X' berwarna merah menunjukkan centroid (pusat) dari masing-masing cluster, yang merupakan representasi visual dari pola segmentasi yang dihasilkan algoritma.

DBSCAN



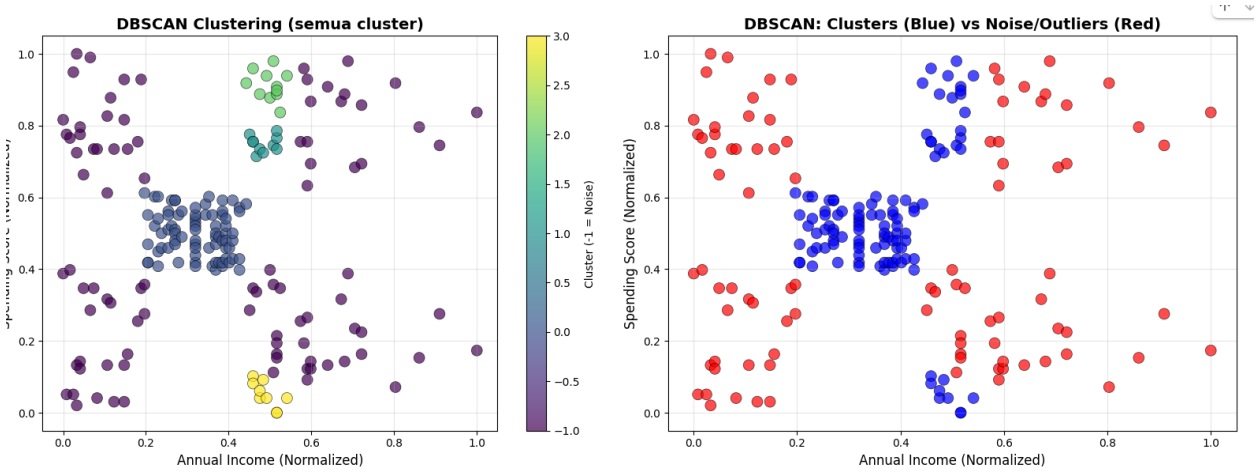
K-distance Graph untuk Penentuan Eps DBSCAN menampilkan kurva yang menunjukkan jarak dari setiap data point ke tetangga ke-5-nya yang telah diurutkan dari terkecil ke terbesar, dan grafik ini digunakan untuk mengidentifikasi nilai eps (epsilon) optimal untuk parameter DBSCAN. Dari visualisasi ini terlihat bahwa kurva dimulai dari nilai sangat rendah (~0.02) pada data points awal (0-50) dan meningkat secara perlahan dan bertahap hingga sekitar data point ke-150, namun setelah itu kurva mengalami peningkatan drastis dan eksponensial menjelang akhir grafik (data points 175-200) di mana jarak mencapai nilai tertinggi ~0.27. Titik "siku" (elbow point) yang menunjukkan transisi antara data point dense dan sparse/outlier terjadi sekitar di area data point 150-175.

```
=== STEP 2: TESTING PARAMETER DBSCAN ===

eps min_samples Clusters    Noise   Silhouette  Davies-Bouldin
-------------------------------------------------------------------------------
0.05    3       13      42  0.4747      0.5059
0.05    4       9       65  0.4594      0.5348
0.05    5       6       79  0.6134      0.4377
0.05    6       4       92  0.6301      0.3966
0.05    7       3       107 0.6093      0.3708
0.1 3       3       10  0.3568      0.6285
0.1 4       2       13  0.4045      0.7308
0.1 5       3       14  0.3537      0.7000
0.1 6       3       18  0.3703      0.6764
0.1 7       4       18  0.5045      0.5632
```

Menunjukkan hasil grid search dari berbagai kombinasi parameter eps dan min_samples yang diuji untuk menemukan parameter optimal DBSCAN pada dataset Mall Customers. Dari data yang ditampilkan, terlihat pola yang jelas: untuk eps=0.05 (eps terlalu kecil), jumlah cluster berkurang drastis dan noise point meningkat sangat banyak (misal eps=0.05, min_samples=5 menghasilkan 6 cluster dengan 79 noise points), mengindikasikan bahwa eps terlalu ketat sehingga banyak data terisolasi; untuk eps=0.1 dengan berbagai min_samples, kualitas clustering meningkat secara konsisten, di mana eps=0.1 dengan min_samples=7 menghasilkan 4 cluster dengan 18 noise points dan Silhouette Score tertinggi 0.5045 serta Davies-Bouldin Index terendah 0.5632 di segmen eps=0.1, menunjukkan ini adalah kombinasi terbaik untuk eps=0.1. Tren umum yang diamati: semakin besar min_samples pada eps yang sama cenderung mengurangi jumlah cluster (lebih konservatif), dan semakin besar eps menghasilkan cluster yang lebih besar namun noise berkurang, namun hasil dari tabel ini menunjukkan bahwa kombinasi optimal adalah eps=0.1 dengan min_samples=7 yang menghasilkan Silhouette Score 0.5045 dan Davies-Bouldin Index 0.5632 yang merupakan trade-off terbaik antara jumlah cluster (4), deteksi noise (18 points), dan kualitas pemisahan cluster.

```
================================================================================
DBSCAN CLUSTERING
================================================================================

Clustering selesai dengan eps=0.05, min_samples=6
Jumlah cluster: 4
Jumlah noise points: 92

Jumlah pelanggan per cluster:
Cluster
-1    92
 0    78
 1    10
 2    11
 3     9
Name: count, dtype: int64

================================================================================
EVALUASI MODEL DBSCAN
================================================================================

1. Silhouette Score: 0.6301
   Range: -1 to 1
   Interpretasi: Baik

2. Davies-Bouldin Index: 0.3966
   Range: 0 to ∞
   Interpretasi: Sangat Baik

3. Noise Points: 92 (46.0%)
   Outlier/Anomali yang terdeteksi
```

Hasil DBSCAN Clustering dengan eps=0.05 dan min_samples=6 menunjukkan output yang sangat berbeda dari K-Means karena DBSCAN memiliki kemampuan untuk deteksi outlier yang tidak dimiliki K-Means. Clustering menghasilkan 4 cluster yang valid dengan 92 noise points (46.0%) yang teridentifikasi sebagai outlier/anomali pelanggan, di mana distribusi pelanggan sangat tidak seimbang: Cluster 0 memiliki 78 pelanggan, Cluster 1 hanya 10 pelanggan, Cluster 2 hanya 11 pelanggan, dan Cluster 3 hanya 9 pelanggan, menunjukkan bahwa dengan eps=0.05 yang sangat ketat, sebagian besar data point tidak memiliki cukup tetangga dalam radius tersebut sehingga diklasifikasi sebagai noise. Evaluasi model menunjukkan hasil yang cukup baik dengan Silhouette Score 0.6301 (kategori "Baik", lebih tinggi dari K-Means 0.5595), yang berarti cluster-cluster yang terbentuk memiliki separasi yang lebih jelas, dan Davies-Bouldin Index 0.3966 (kategori "Sangat Baik", lebih rendah dari K-Means 0.5678), yang mengindikasikan cluster-cluster sangat terpisah dengan jelas tanpa overlap. Namun, noise points sebanyak 46% adalah persentase yang sangat tinggi.



Visualisasi DBSCAN Clustering dengan eps=0.05 dan min_samples=6 menampilkan dua subplot yang memberikan perspektif berbeda tentang hasil clustering density-based ini. Plot sebelah kiri (DBSCAN Clustering - semua cluster) menunjukkan distribusi spasial dari empat cluster yang valid dengan warna-warna berbeda (biru, hijau, cyan, dan ungu) berdasarkan colormap dengan cluster label -1 untuk noise, di mana terlihat bahwa cluster-cluster terbentuk di area dengan kepadatan tinggi: cluster biru terkonsentrasi di area medium-income medium-spending (0.3-0.6, 0.4-0.6), cluster hijau di area high-income medium-spending (0.4-0.65, 0.8-0.95), cluster cyan di area medium-high income high-spending (0.45-0.65, 0.65-0.8), dan cluster ungu menyebar di berbagai area lain. Plot sebelah kanan (DBSCAN - Clusters vs Noise/Outliers) menggunakan highlighting yang lebih jelas di mana pelanggan dalam cluster ditampilkan berwarna biru dan noise/outlier terlihat dengan jelas berwarna merah, menunjukkan bahwa 92 outlier (46% dari total 200 pelanggan) tersebar di hampir semua area.

KESIMPULAN

Kedua algoritma K-Means dan DBSCAN menghasilkan clustering yang berkualitas baik pada dataset Mall Customers dengan nilai metrik evaluasi yang cukup memuaskan. K-Means dengan K=5 menghasilkan Silhouette Score 0.5595 (kategori "Baik") dan Davies-Bouldin Index 0.5678 (kategori "Sangat Baik"), dengan distribusi pelanggan yang seimbang di kelima cluster (22, 35, 39, 81, dan 23 pelanggan) tanpa ada noise points, menghasilkan lima segmen pelanggan yang jelas: impulsive buyers (low income-high spending), prudent buyers (medium income-low spending), balanced buyers (medium income-medium spending), affluent buyers (high income-high spending), dan economical buyers (low income-low spending). Sementara DBSCAN dengan eps=0.05 dan min_samples=6 menghasilkan Silhouette Score 0.6301 (kategori "Baik", lebih tinggi dari K-Means) dan Davies-Bouldin Index 0.3966 (kategori "Sangat Baik", lebih rendah dari K-Means), namun mengidentifikasi 92 noise points (46% dari total) yang diklasifikasi sebagai outlier/anomali pelanggan dengan distribusi cluster yang sangat tidak seimbang (78, 10, 11, dan 9 pelanggan).

K-Means adalah algoritma yang lebih cocok dan direkomendasikan untuk keperluan customer segmentation pada dataset Mall Customers dibandingkan DBSCAN, dengan alasan utama sebagai berikut. Pertama, dataset Mall Customers tidak memiliki outlier yang signifikan atau banyak (sebagian besar pelanggan mengikuti pola normal distribusi income dan spending), sehingga pendekatan partitioning K-Means lebih sesuai daripada density-based DBSCAN yang dirancang untuk data dengan banyak outlier. Kedua, hasil clustering K-Means menghasilkan segmentasi yang lebih balanced dan actionable secara bisnis dengan distribusi pelanggan yang wajar di setiap cluster, memudahkan implementasi strategi marketing yang targeted per segmen. Ketiga, K-Means tidak menghasilkan noise points, sehingga semua 200 pelanggan teralokasi ke dalam cluster yang bermakna dan dapat dianalisis lebih lanjut, berbeda dengan DBSCAN yang mengklasifikasikan 46% pelanggan sebagai outlier yang menyulitkan business insight. Keempat, visualisasi K-Means menunjukkan pemisahan cluster yang lebih jelas dan interpretable secara visual dalam scatter plot, memudahkan komunikasi hasil ke stakeholder bisnis.