Pertemuan 10

Pembelajaran Tak Terawasi (Unsupervised Learning)

A. Konsep Clustering dan Segmentasi Data

1. Pengertian Clustering

Clustering adalah teknik pembelajaran tak terawasi (unsupervised learning) yang bertujuan mengelompokkan data ke dalam beberapa kelompok (cluster) berdasarkan kemiripan karakteristik. Melalui proses ini, algoritma secara otomatis mengidentifikasi pola atau struktur alami yang terdapat dalam dataset tanpa memerlukan label kelas. Teknik clustering memungkinkan peneliti memahami hubungan antar objek data secara lebih mendalam dan menyederhanakan kompleksitas informasi. Hasil pengelompokan ini sering digunakan sebagai dasar pengambilan keputusan, segmentasi, atau analisis lanjutan pada berbagai bidang aplikasi.

Pada proses ini, tidak ada label kelas yang diberikan, sehingga model harus menemukan pola secara mandiri dari struktur data.

Tujuan utama clustering:

- Mengidentifikasi struktur atau pola tersembunyi dalam data.
- Memahami karakteristik alami dari kelompok-kelompok yang terbentuk.
- Menyederhanakan kompleksitas data untuk analisis lebih lanjut.

2. Mengapa Clustering Penting dalam AI dan Data Mining?

Clustering digunakan ketika:

- Kita memiliki data dalam jumlah besar tapi tidak mengetahui kategorinya.
- Dibutuhkan segmentasi otomatis tanpa intervensi manual.
- Ingin memahami pola perilaku, preferensi, atau hubungan antar objek.

Contoh penggunaan nyata:

- Segmentasi pelanggan dalam bisnis.
- Pengelompokan dokumen atau berita.

- Identifikasi pola serangan pada jaringan komputer.
- Clustering gambar berdasarkan kemiripan fitur.

3. Konsep Kemiripan (*Similarity*) dan Jarak (*Distance*)

Clustering bekerja berdasarkan intuisi bahwa objek yang mirip harus berada dalam satu cluster, sedangkan objek yang jauh berbeda diletakkan pada cluster lain.

Umumnya ukuran jarak Euclidean digunakan:

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

Beberapa ukuran lain:

- Manhattan Distance
- Cosine Similarity
- Jaccard Similarity (untuk data kategori)

Pemilihan ukuran jarak berdampak langsung pada hasil cluster.

4. Segmentasi Data

Segmentasi data adalah proses membagi data menjadi beberapa segmen yang lebih kecil dan homogen. Clustering adalah salah satu teknik paling populer untuk segmentasi. Segmen yang terbentuk biasanya memiliki karakteristik internal yang serupa sehingga memudahkan analisis perilaku, preferensi, atau pola tertentu dalam data. Dengan segmentasi yang baik, proses pengambilan keputusan dapat menjadi lebih akurat karena setiap kelompok diperlakukan sesuai dengan kebutuhan dan ciri khasnya. Dalam praktiknya, segmentasi banyak digunakan di bidang pemasaran, kesehatan, pendidikan, dan sistem rekomendasi untuk menyusun strategi yang lebih tepat sasaran dan efisien.

Konsep	Definisi	Contoh
Clustering	Proses mengelompokkan data berdasarkan kemiripan nilai	K-Means, Hierarchical Clustering
Segmentasi Data	Hasil akhir berupa pembagian segmen konsumen/data	Segmentasi pelanggan: loyal, reguler, pasif

Segmentasi digunakan dalam:

- Targeted marketing
- Prediksi perilaku pengguna
- Personalization system (rekomendasi konten)

5. Jenis Pendekatan Clustering

a. Partitional Clustering

Membagi data menjadi k cluster secara eksklusif berarti setiap data hanya dapat menjadi anggota dari satu cluster dan tidak boleh berada di lebih dari satu kelompok sekaligus. Pendekatan ini memastikan bahwa pembagian data bersifat tegas sehingga batas antar cluster jelas dan tidak tumpang tindih. Metode ini sering digunakan ketika tujuan analisis adalah memperoleh kelompok yang terpisah secara jelas berdasarkan kemiripan nilai fitur.

Contoh: K-Means, K-Medoids

b. Hierarchical Clustering

Menghasilkan struktur pohon (dendrogram) yang menunjukkan hubungan antar cluster, sehingga setiap proses penggabungan atau pemisahan kelompok dapat ditelusuri secara visual. Dendrogram ini membantu memahami tingkat kemiripan antar objek maupun antar cluster, mulai dari yang paling mirip hingga yang paling berbeda. Pendekatan ini sangat berguna ketika analisis membutuhkan interpretasi hierarkis dan eksplorasi menyeluruh terhadap struktur data.

Contoh: Agglomerative, Divisive

c. Density-Based Clustering

Mengenali cluster berdasarkan kepadatan titik berarti algoritma mengidentifikasi area dalam ruang data yang memiliki konsentrasi titik tinggi sebagai sebuah cluster. Titik-titik yang berada di wilayah dengan kepadatan rendah dianggap sebagai noise atau outlier. Pendekatan ini sangat efektif untuk menemukan cluster dengan bentuk yang tidak beraturan dan ukuran yang bervariasi, terutama ketika data tidak mengikuti pola spherical seperti pada K-Means.

Contoh: DBSCAN, OPTICS

d. Model-Based Clustering

Menggunakan model probabilistik berarti algoritma membentuk cluster berdasarkan distribusi peluang dari data, bukan sekadar jarak antar titik. Setiap cluster dipandang sebagai komponen probabilistik, misalnya distribusi Gaussian, yang mewakili pola tertentu dalam data. Pendekatan ini memungkinkan satu data memiliki tingkat keanggotaan (probability) terhadap beberapa cluster sekaligus, sehingga hasil pengelompokan menjadi lebih fleksibel dan realistis pada data yang memiliki struktur kompleks.

Contoh: Gaussian Mixture Model (GMM)

6. Karakteristik Clustering yang Baik

Cluster yang baik memiliki ciri:

- Intra-cluster similarity tinggi → anggota dalam cluster mirip satu sama lain.
- Inter-cluster similarity rendah → antar cluster berbeda signifikan.
- Stabil dan konsisten terhadap perubahan kecil pada data.
- Memiliki interpretasi yang jelas sesuai konteks domain.

7. Tantangan pada Clustering

- Menentukan jumlah cluster optimal (k).
- Data berdimensi tinggi sulit dipetakan pola jaraknya.
- Cluster tidak selalu berbentuk bulat/spherical.
- Sensitif terhadap outlier (terutama K-Means).
- Interpretasi hasil yang bersifat subjektif.

8. Contoh Ilustrasi Konseptual

Misalkan perusahaan retail ingin memahami perilaku pembeli.

Tanpa label, data berikut dianalisis:

- Jumlah pembelian bulanan
- Total pengeluaran
- Frekuensi kunjungan
- Jenis produk yang paling sering dibeli

Clustering akan membagi pelanggan menjadi segmen, seperti:

- **Cluster 1: High-Value Customer**
- **Cluster 2: Regular Customer**
- **Cluster 3: Low-Activity Customer**

Hasil segmentasi kemudian digunakan untuk strategi pemasaran yang lebih tepat sasaran.

B. Metode K-Means dan Hierarchical Clustering

1. Metode K-Means

a. Konsep Dasar K-Means

K-Means merupakan metode partitional clustering yang membagi data menjadi *k* cluster berdasarkan jarak terhadap pusat cluster (*centroid*). Pendekatan ini bekerja secara iteratif dengan tujuan meminimalkan variasi internal (*intra-cluster variance*) sehingga data dalam satu cluster memiliki kemiripan yang tinggi.

b. Langkah-Langkah Algoritma K-Means

- Menentukan jumlah cluster (*k*) yang diinginkan.
- Menginisiasi centroid awal secara acak atau menggunakan teknik seperti k-means++.
- Menghitung jarak setiap data ke centroid terdekat menggunakan Euclidean Distance:

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

- Mengelompokkan data sesuai centroid terdekat.
- Menghitung centroid baru berdasarkan rata-rata anggota cluster:

$$C_j = \frac{1}{|S_j|}\sum_{x_i \in S_j} x_i$$

- Mengulangi langkah 3–5 hingga centroid stabil atau perubahan cluster tidak lagi terjadi.

c. Kelebihan K-Means

- Sangat cepat dan efisien pada dataset besar.
- Mudah diimplementasikan.
- Hasil cluster mudah diinterpretasikan.

d. Kelemahan K-Means

- Harus menentukan jumlah *k* di awal.
- Sensitif terhadap outlier dan pemilihan centroid awal.
- Kurang efektif untuk cluster berbentuk non-spherical atau memiliki ukuran berbeda.

2. Hierarchical Clustering

Hierarchical clustering adalah metode yang membangun struktur berbentuk hierarki atau pohon pengelompokan tanpa menentukan jumlah cluster di awal. Hasil akhirnya digambarkan dalam bentuk dendrogram.

a. Dua Pendekatan Utama

1) Agglomerative (Bottom-Up)

- Memulai dari tiap data sebagai cluster tunggal.
- Menggabungkan dua cluster paling mirip secara bertahap.
- Berlanjut hingga seluruh data berada dalam satu cluster besar.

2) Divisive (Top-Down)

- Memulai dari satu cluster besar berisi semua data.
- Memecah cluster menjadi kelompok yang lebih kecil secara bertahap.
- Berlanjut hingga diperoleh jumlah cluster yang diinginkan.

b. Linkage Criteria

Penentuan cluster mana yang digabungkan bergantung pada metode *linkage*:

1. Single Linkage

   Jarak antar cluster ditentukan oleh jarak minimum antar anggota.

2. Complete Linkage

   Menggunakan jarak maksimum antar titik dua cluster.

3. Average Linkage

   Berdasarkan rata-rata jarak antar semua pasangan titik.

4. Ward's Method

   Menggabungkan cluster yang menghasilkan peningkatan variansi total paling kecil.

c. Kelebihan Hierarchical Clustering

- Tidak perlu menentukan jumlah cluster di awal.
- Dapat divisualisasikan melalui dendrogram untuk memahami struktur data.
- Bekerja baik pada dataset kecil hingga menengah.

d. Kelemahan Hierarchical Clustering

- Tidak efisien untuk dataset yang sangat besar.
- Proses penggabungan/pemisahan bersifat permanen (tidak bisa di-*undo*).
- Sensitif terhadap pemilihan jarak dan metode linkage.

3. Perbandingan K-Means vs Hierarchical Clustering

Aspek	K-Means	Hierarchical
Penentuan jumlah cluster	Harus ditentukan di awal	Dapat ditentukan setelah melihat dendrogram
Waktu komputasi	Cepat (efisien untuk dataset besar)	Lambat pada dataset besar
Sensitivitas outlier	Tinggi	Tinggi
Interpretasi	Lebih sederhana	Lebih kaya (melalui dendrogram)
Bentuk cluster	Cenderung spherical	Tidak wajib spherical

C. Evaluasi hasil clustering (Silhouette Coefficient)

Evaluasi dalam clustering sangat penting karena metode ini tidak menggunakan label kelas. Oleh karena itu, kita membutuhkan metrik yang dapat menilai kualitas cluster berdasarkan konsistensi internal dan pemisahan antar cluster. Salah satu metrik yang paling umum digunakan adalah *Silhouette Coefficient*.

1. Konsep Dasar Silhouette Coefficient

Silhouette Coefficient mengukur seberapa baik suatu data ditempatkan dalam cluster-nya, dengan membandingkan kedekatannya terhadap anggota cluster sendiri dan cluster terdekat lainnya. Nilai yang dihasilkan memberikan gambaran apakah data tersebut benar-benar berada pada kelompok yang tepat atau justru lebih mirip dengan cluster lain. Dengan demikian, metrik ini membantu menilai kualitas pemisahan dan kekompakan cluster secara objektif.

Nilai silhouette untuk setiap data berada pada rentang:

$$-1 \leq s(i) \leq 1$$

Interpretasi nilai:

- s(i) mendekati 1 → data berada pada cluster yang sangat tepat (kompak dan terpisah dengan baik)
- s(i) sekitar 0 → data berada di batas antara dua cluster
- s(i) mendekati -1 → data salah penempatan, lebih mirip dengan cluster lain

2. Rumus Silhouette Coefficient

Untuk setiap data (i), perhitungan dilakukan dalam tiga langkah:

a. Menghitung rata-rata jarak dalam cluster sendiri (a(i))

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i,j)$$

Mencerminkan seberapa dekat data dengan anggota cluster-nya.

b. Menghitung jarak minimum ke cluster lain (b(i))

$$b(i) = \min_{k \neq C_i} \left(\frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \right)$$

Mencerminkan kedekatan data terhadap cluster tetangga terdekat.

c. Menghitung silhouette score untuk data i

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Nilai ini menggambarkan kualitas penempatan data dalam cluster.

d. Silhouette Coefficient keseluruhan cluster

$$S = \frac{1}{n} \sum_{i=1}^{n} s(i)$$

Semakin besar nilai (S), semakin baik hasil clustering.

3. Interpretasi Nilai Silhouette Coefficient

Nilai	Kualitas Cluster
0.71 – 1.00	Struktur cluster sangat kuat
0.51 – 0.70	Struktur cluster baik
0.26 – 0.50	Struktur cluster lemah
≤ 0.25	Tidak ada struktur cluster yang jelas

Metrik ini membantu memutuskan apakah:

- Jumlah *k* sudah tepat pada K-Means
- Metode clustering menghasilkan kelompok yang valid

- Perlu dilakukan preprocessing ulang (normalisasi, penghilangan outlier)

4. Keunggulan Silhouette Coefficient

- Tidak memerlukan ground truth atau label.
- Dapat digunakan untuk membandingkan hasil clustering antar metode.
- Memberikan informasi pada tingkat:
    - per individu (kejelasan posisi data),
    - per cluster,
    - keseluruhan model.

5. Keterbatasan Silhouette Coefficient

- Tidak selalu cocok untuk data berdimensi sangat tinggi.
- Tidak optimal untuk cluster dengan bentuk kompleks (misalnya cluster non-konveks).
- Perhitungan dapat mahal untuk dataset sangat besar karena membutuhkan perhitungan jarak berpasangan.