CLUSTERING

1. Introduction

Clustering merupakan teknik dalam unsupervised learning yang digunakan untuk mengelompokkan data berdasarkan kemiripan karakteristik tanpa menggunakan label. Tujuan utama clustering adalah menemukan struktur atau pola tersembunyi dalam data.

Dalam konteks bisnis, clustering sering digunakan untuk:

- segmentasi pelanggan
- pengelompokan produk
- analisis perilaku pengguna

Berbeda dengan klasifikasi, clustering tidak memiliki “label”, sehingga termasuk unsupervised learning.

2. k-Means Clustering

Konsep

k-Means adalah algoritma clustering yang membagi data menjadi k kelompok berdasarkan jarak ke centroid.

Langkah k-Means

1. Tentukan jumlah cluster (k)
2. Inisialisasi centroid
3. Hitung jarak ke centroid
4. Kelompokkan data
5. Update centroid
6. Ulangi sampai konvergen

3. Agglomerative Hierarchical Clustering

Konsep

Metode ini menggabungkan data dari bawah ke atas (bottom-up).

Karakteristik

- Tidak perlu menentukan jumlah cluster di awal
- Menghasilkan dendrogram

Metode Linkage

- Single linkage → jarak terdekat
- Complete linkage → jarak terjauh
- Average linkage → rata-rata jarak

STUDI KASUS

Sebuah perusahaan e-commerce ingin melakukan segmentasi pelanggan berdasarkan:

- umur
- pendapatan
- jumlah kunjungan

Yang Harus Dianalisis

1. Menentukan jumlah cluster terbaik
2. Mengelompokkan pelanggan
3. Membandingkan metode clustering
4. Visualisasi hasil clustering
5. Interpretasi karakteristik tiap cluster

LANGKAH PENYELESAIAN

LANGKAH 1: IMPORT LIBRARY

Penjelasan:

Digunakan library untuk clustering dan visualisasi.

```
import numpy as np
import matplotlib.pyplot as plt

from sklearn.cluster import KMeans, AgglomerativeClustering
from sklearn.preprocessing import StandardScaler
from scipy.cluster.hierarchy import dendrogram, linkage
```

LANGKAH 2: INPUT DATA

Penjelasan:

Dataset sama seperti modul sebelumnya agar konsisten.

```
X = np.array([
[22,4,5],[25,5,6],[28,6,7],[30,6,8],[32,7,10],
[35,7,12],[38,8,13],[40,8,15],[42,9,16],[45,9,18],
[23,4,4],[26,5,7],[29,6,6],[31,6,9],[34,7,11],
[37,8,12],[39,8,14],[43,9,15],[46,10,17],[49,10,19]
])
```

LANGKAH 3: NORMALISASI DATA

Penjelasan:

Normalisasi penting agar skala fitur tidak mempengaruhi hasil clustering.

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

LANGKAH 4: MENENTUKAN JUMLAH CLUSTER (ELBOW METHOD)

Penjelasan:

Elbow method digunakan untuk mencari jumlah cluster optimal.

```
inertia = []

for k in range(1, 6):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X_scaled)
    inertia.append(kmeans.inertia_)

plt.figure()
plt.plot(range(1,6), inertia, marker='o')
plt.xlabel("Jumlah Cluster")
plt.ylabel("Inertia")
plt.title("Elbow Method")
plt.show()
```

LANGKAH 5: K-MEANS CLUSTERING

Penjelasan:

Mengelompokkan data berdasarkan jumlah cluster terbaik.

```
kmeans = KMeans(n_clusters=3, random_state=42)
labels_kmeans = kmeans.fit_predict(X_scaled)
```

VISUALISASI K-MEANS

```
plt.figure()
plt.scatter(X_scaled[:,0], X_scaled[:,1], c=labels_kmeans)

plt.xlabel("Umur (scaled)")
plt.ylabel("Pendapatan (scaled)")
plt.title("K-Means Clustering")

plt.show()
```

LANGKAH 6: HIERARCHICAL CLUSTERING

Penjelasan:

Menggunakan metode agglomerative untuk membentuk cluster.

```
agglo = AgglomerativeClustering(n_clusters=3)
labels_agglo = agglo.fit_predict(X_scaled)
```

VISUALISASI HIERARCHICAL

```
plt.figure()
plt.scatter(X_scaled[:,0], X_scaled[:,1], c=labels_agglo)

plt.xlabel("Umur (scaled)")
plt.ylabel("Pendapatan (scaled)")
plt.title("Hierarchical Clustering")

plt.show()
```

LANGKAH 7: DENDROGRAM

Penjelasan:

Dendrogram menunjukkan proses penggabungan cluster.

```
linked = linkage(X_scaled, method='ward')

plt.figure()
dendrogram(linked)
plt.title("Dendrogram")
plt.show()
```

INTERPRETASI HASIL

- k-Means:
    - cepat dan efisien
    - cocok untuk data besar
- Hierarchical:
    - lebih informatif (ada dendrogram)
    - cocok untuk analisis struktur data

Cluster dapat diinterpretasikan sebagai:

- Cluster 1 → pelanggan aktif
- Cluster 2 → pelanggan menengah
- Cluster 3 → pelanggan pasif

KESIMPULAN

Clustering merupakan metode penting dalam analisis data untuk menemukan pola tanpa label. k-Means unggul dalam kecepatan dan efisiensi, sedangkan hierarchical clustering memberikan visualisasi struktur data yang lebih jelas melalui dendrogram. Pemilihan metode tergantung pada kebutuhan analisis dan karakteristik data. Kombinasi keduanya dapat memberikan insight yang lebih komprehensif dalam segmentasi pelanggan.