Data Mining

Pertemuan 9
Clustering

Tujuan Klastering dan Klasifikasi

- Mengurangi kompleksitas maupun ukuran basis data.
- Klastering : membagi berdasarkan kemiripan dan ketidak miripan
- Klasifikasi : membagi berdasarkan kelas yang telah ditetapkan

Klastering

- Analisis kluster atau pengelompokan adalah tugas mengelompokkan seperangkat objek sedemikian rupa sehingga objek dalam kelompok yang sama (disebut kluster) lebih mirip (dalam arti tertentu) satu sama lain daripada dengan yang ada di kelompok lain (kluster).

Sifat Klastering

- Menentukan jumlah kelompok yang hendak dibagi : (fragmentasi)
- Menentukan variabel yang diprioritaskan sebagai *centroid* dan diabaikan (text, kategori, relasi).
- Centroid : nilai tengah dalam variabel (tinggi, sedang, rendah)

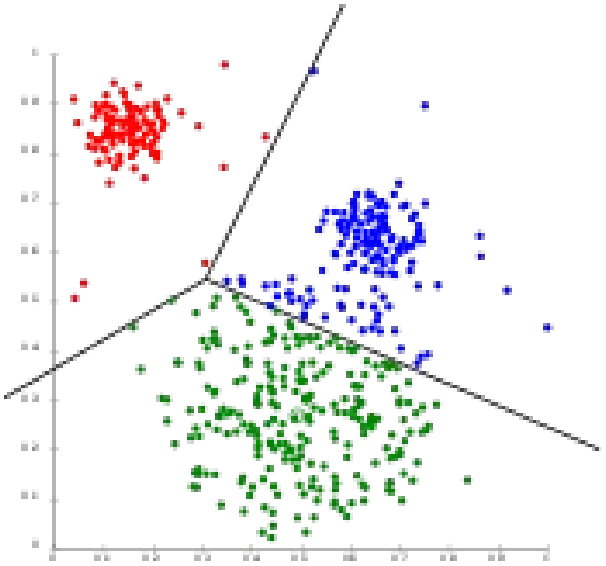
Jenis Klastering

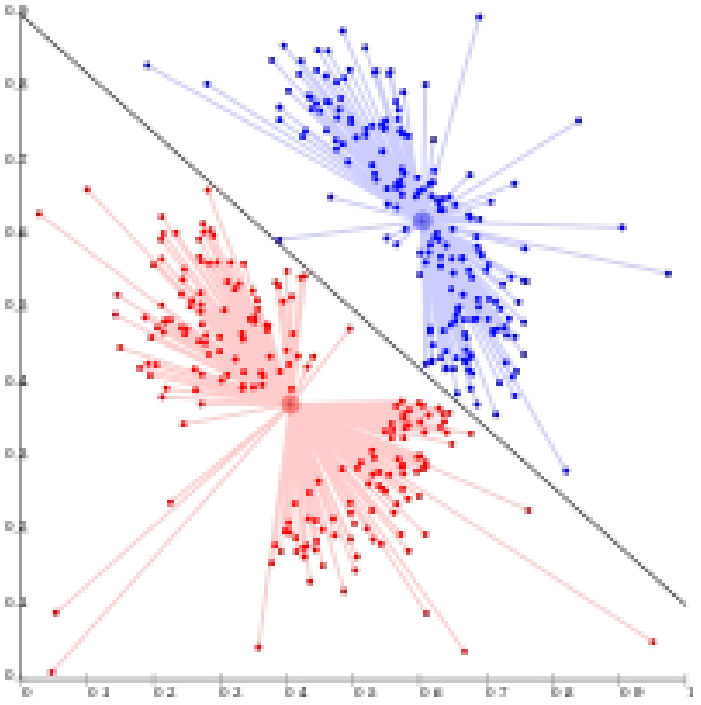
- Centroid-based clustering (Paling umum)
- Connectivity-based clustering (hierarchical clustering)
- Distribution-based clustering
- Density-based clustering

K-means Clustering

- Metode kuantisasi vektor, berasal dari pemrosesan sinyal, yang populer untuk analisis klaster dalam penambangan data.
- Bertujuan untuk mempartisi *n* observasi ke dalam klaster di mana setiap observasi termasuk ke dalam cluster dengan *mean* terdekat, berfungsi sebagai *prototipe* dari *cluster*.

Visualisasi K-means

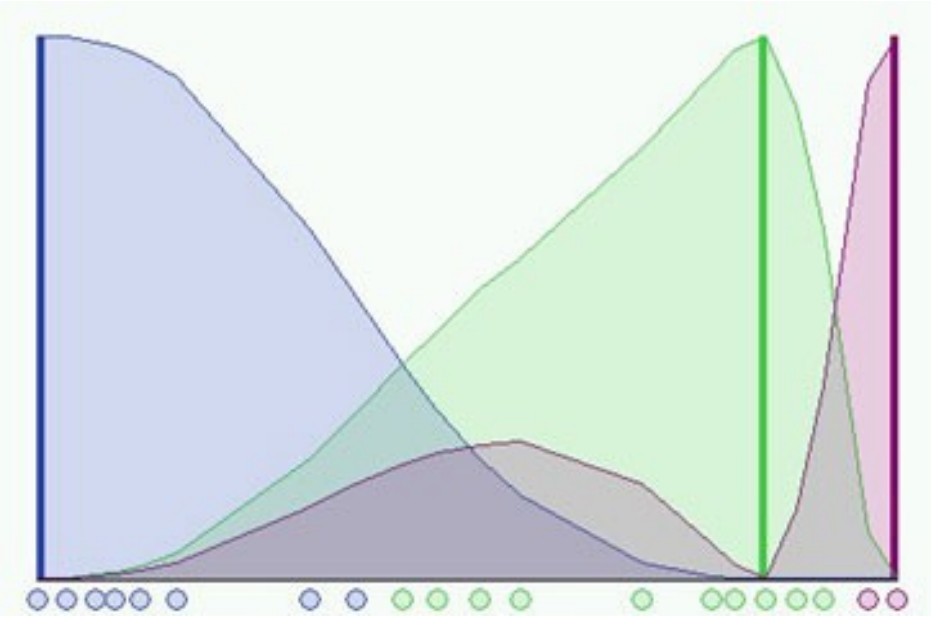


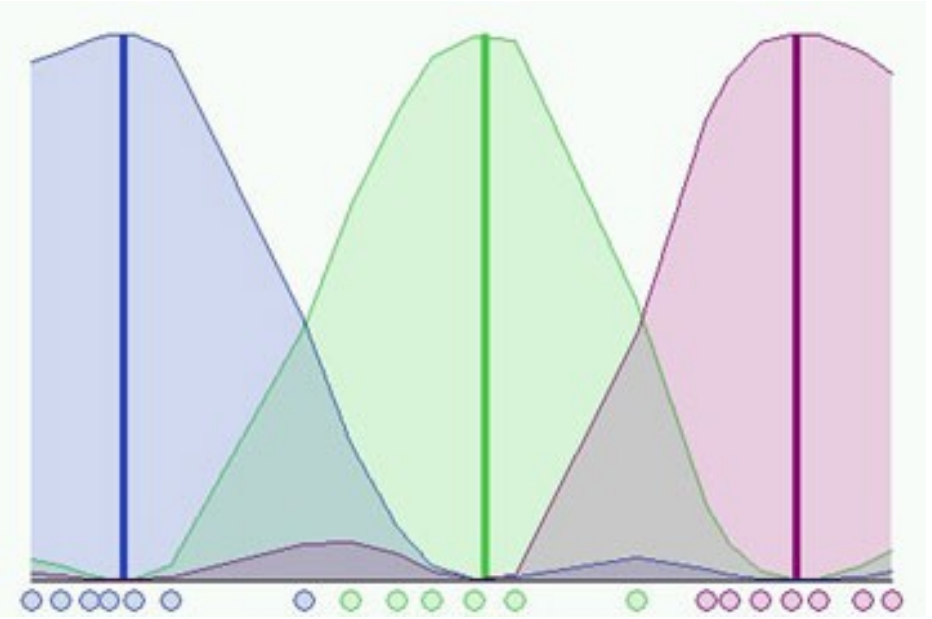


Fuzzy C-Means (Soft - K-means)

- Juga disebut sebagai soft clustering atau soft k-means
- Bentuk pengelompokan di mana setiap titik data dapat menjadi milik lebih dari satu cluster.

Visualisasi Fuzzy C-Means

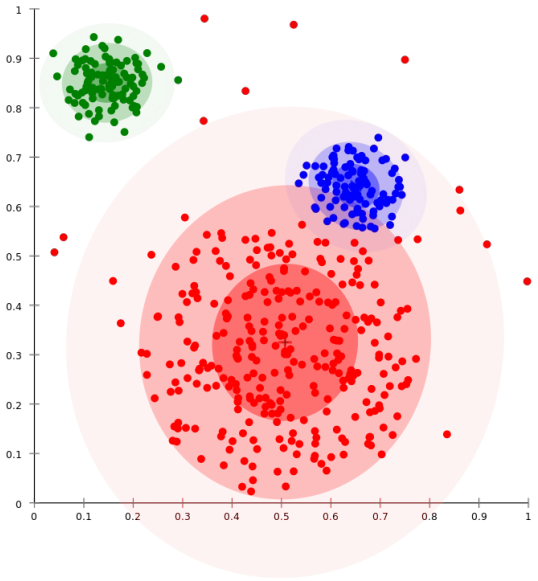


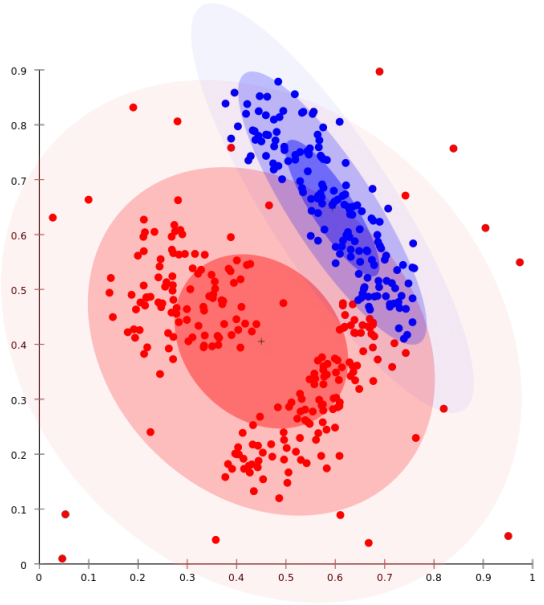


Distribution-based clustering

- Model pengelompokan yang paling terkait erat dengan statistik didasarkan pada model distribusi.
- Cluster kemudian dapat didefinisikan sebagai objek milik yang paling mungkin untuk distribusi yang sama.
- Keunggulan dari metode ini adalah mampu menghasilkan kumpulan data dengan sampling objek acak dari suatu distribusi.

Visualisasi Distribution-based

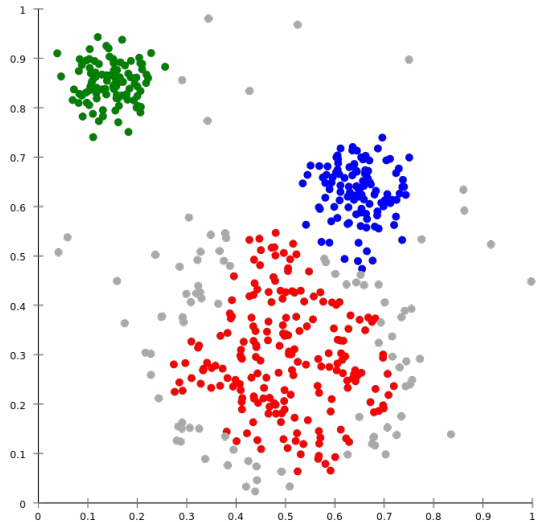


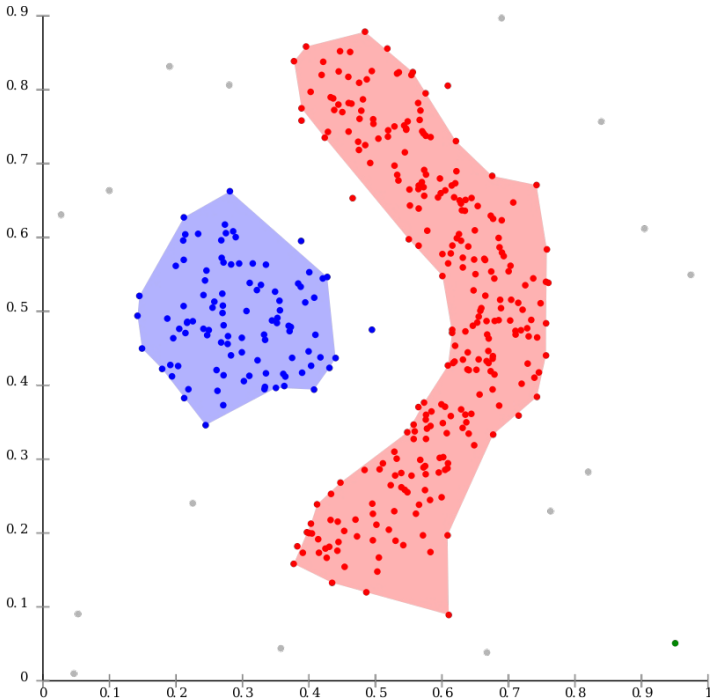


Density-based clustering

- Pengelompokan berbasis kepadatan.
- Kelompok didefinisikan sebagai daerah dengan kepadatan yang lebih tinggi dari sisa kumpulan data.
- Objek *outsider* (memiliki pola langka) - yang diperlukan untuk memisahkan kelompok - biasanya dianggap sebagai titik kebisingan dan perbatasan.

Visualisasi Density-based

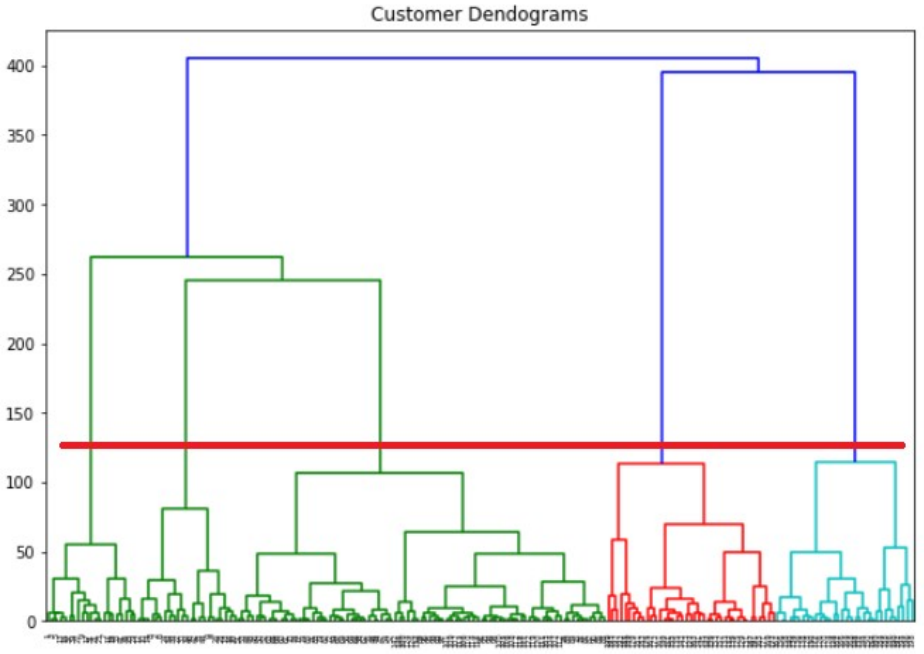


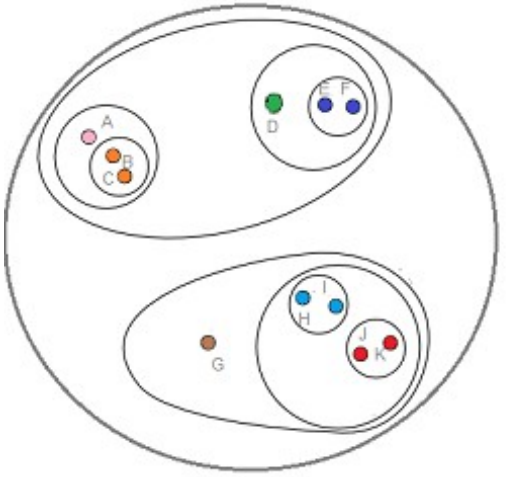


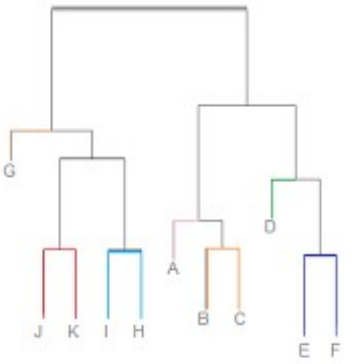
Hierarchical clustering

- Metode analisis klaster yang berusaha membangun hierarki klaster.
  - Agglomerative: Ini adalah pendekatan "bottom-up": setiap observasi dimulai di klasternya sendiri, dan pasangan cluster digabungkan ketika seseorang bergerak naik ke hierarki.
  - Divisive: Ini adalah pendekatan "top-down": semua pengamatan dimulai dalam satu kluster, dan pemisahan dilakukan secara rekursif ketika seseorang bergerak ke hierarki.

Visualisasi Hierarchical clustering







Silhouette

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, if|C_i| > 1$$

$$s(i) = 0, if|C_i| = 1$$

- a = mean intra-cluster distance
- b = mean nearest-cluster distance
- Untuk setiap cluster C sebanyak I
- C_i = |a/b|