Pengantar *Data Mining* #6: *Clustering*

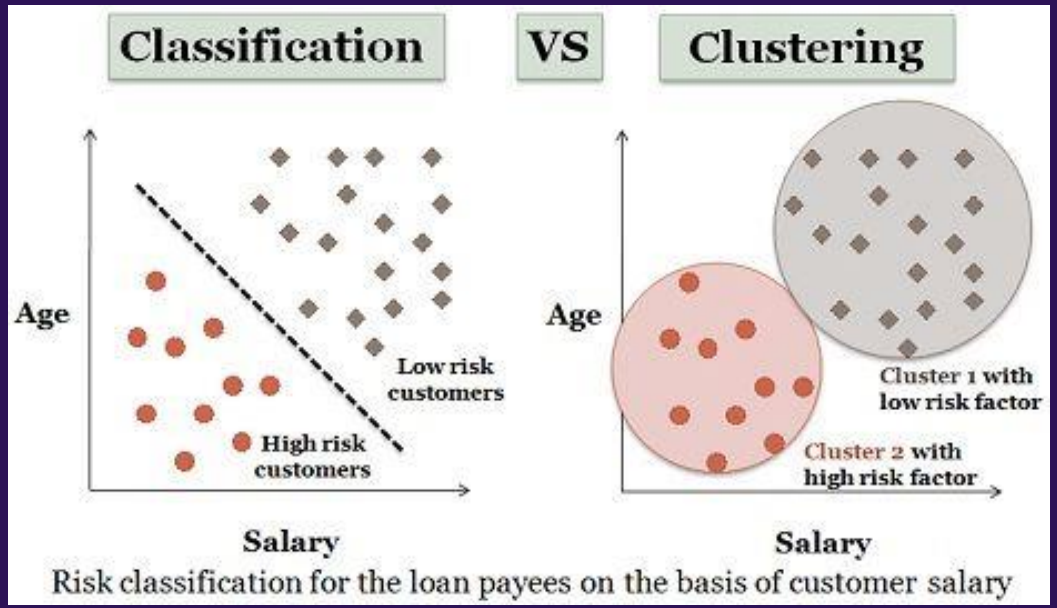
Isnan Mulia, S.Komp, M.Kom

Apa Itu *Clustering/Cluster Analysis*?

- Proses mengelompokkan kumpulan objek data ke dalam beberapa kelompok/*cluster* sedemikian sehingga objek-objek di dalam masing-masing *cluster* saling mirip satu dengan yang lain, serta saling tidak mirip dengan objek-objek di *cluster* lain
- Proses untuk meminimalkan jarak intra *cluster* & memaksimalkan jarak antar *cluster*
- Penentuan kemiripan/ketidakmiripan berdasarkan atribut yang dimiliki data & melibatkan ukuran jarak
- Menggunakan data yang tidak memiliki label kelas

  ➔ Disebut *unsupervised learning* & *learning by observation*

Clustering vs Klasifikasi



Sumber: https://techdifferences.com/wp-content/uploads/2018/01/Untitled.jpg

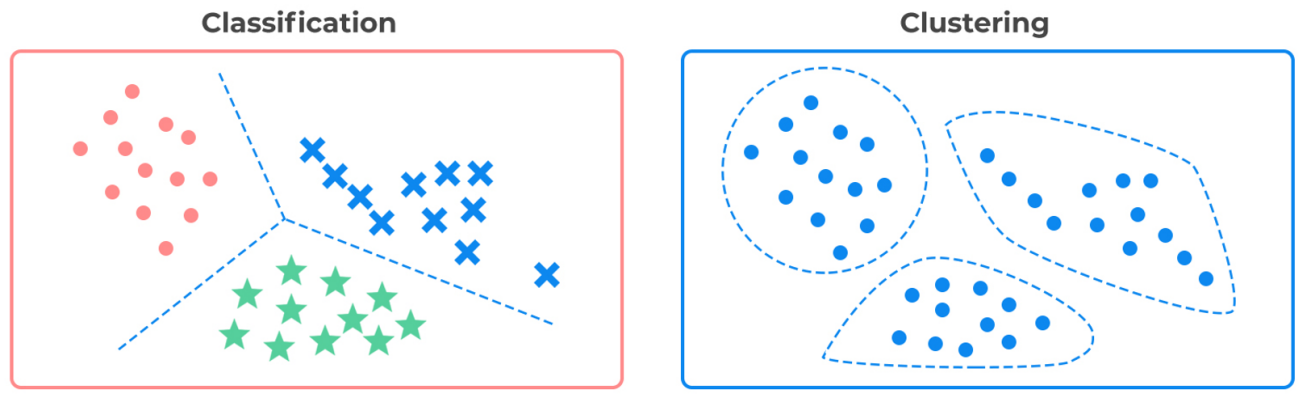
Klasifikasi:

- Menggunakan data yang memiliki label kelas
- Membuat bidang pemisah antara data pada label kelas yang satu dengan label kelas yang lain

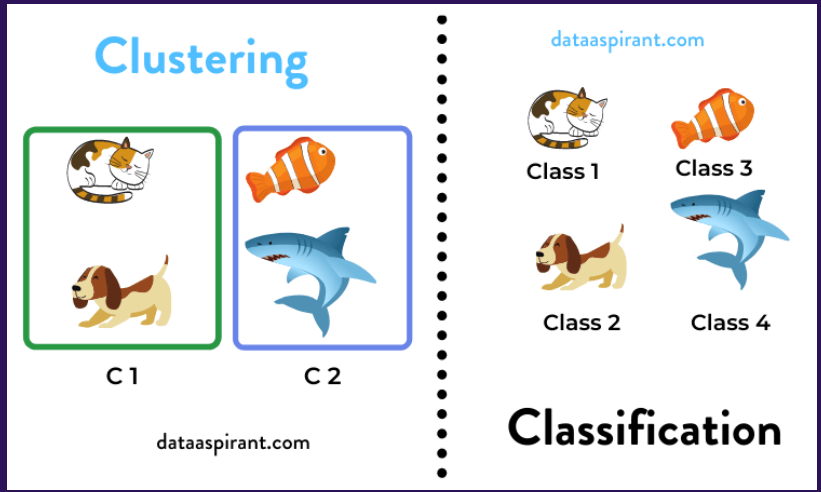
Clustering:

- Menggunakan data yang tidak memiliki label kelas
- Mengelompokkan data yang memiliki karakteristik yang mirip

Clustering vs Klasifikasi



Sumber: https://editor.analyticsvidhya.com/uploads/74251clustering.PNG



Sumber: https://i1.wp.com/dataaspirant.com/wp-content/uploads/2020/12/5-Clustering-Vs-Classification-Example.png

Manfaat *Clustering*

- Mengelompokkan data berdasarkan kemiripan karakteristik
- Mendapatkan informasi mengenai distribusi data
- Dapat digunakan dalam langkah praproses data untuk mengkarakterisasi data
- Mendeteksi data pencilan, yang memiliki karakteristik berbeda dari objek data yang lain

Kebutuhan *Clustering*

- Kemampuan untuk melakukan *clustering* pada data berdimensi besar
- Kemampuan untuk menghadapi berbagai jenis atribut data
- Penemuan *cluster* dengan bentuk tertentu
- Kemampuan untuk menghadapi data berderau
- *Clustering* tambahan, jika ada tambahan data input
- Hasil *clustering* dapat diinterpretasikan dengan mudah

Metode *Clustering*

- Metode partisi → membagi data menjadi *k* kelompok
  - → *k-means*, *k-medoid*
- Metode hirarki → membuat dekomposisi hirarkis dari data yang diberikan
  - → *Agglomerative* (*bottom-up*), *divisive* (*top-down*)
- Metode *density-based* → membangun *cluster* selama kepadatan/jumlah objek dalam *cluster* melewati nilai batas tertentu
  - → DBSCAN, OPTICS
- Metode *grid-based* → mengkuantisasi ruang objek menjadi sejumlah sel yang membentuk struktur grid
  - → STING, CLIQUE

K-Means

- Teknik berbasis *centroid* (titik tengah *cluster*)
- *Centroid* digunakan untuk merepresentasikan sebuah *cluster*
- Penentuan titik *centroid* baru berdasarkan nilai rata-rata jarak antara objek di dalam *cluster* dengan posisi terakhir titik *centroid*
- Meminimalkan varian di dalam *cluster*
- Mengukur kualitas *cluster*:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{c_i})^2$$

K-Means – Algoritma

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

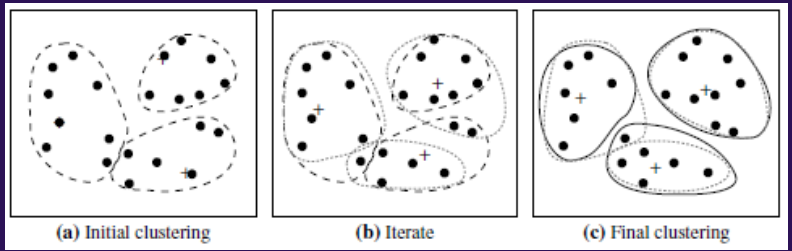
Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

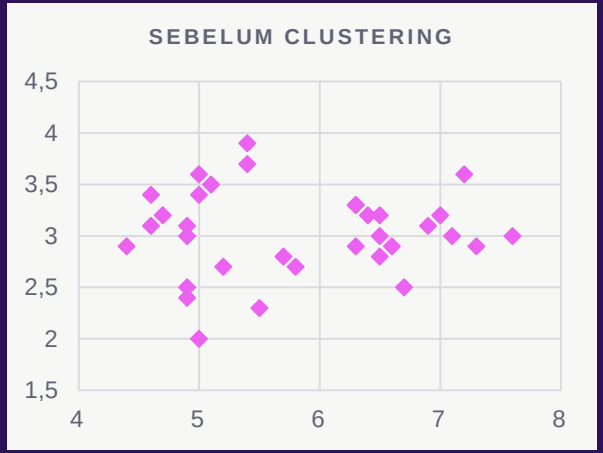
Method:

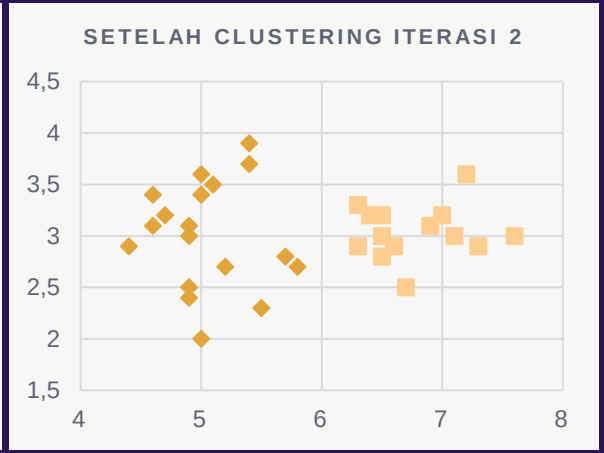
(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;



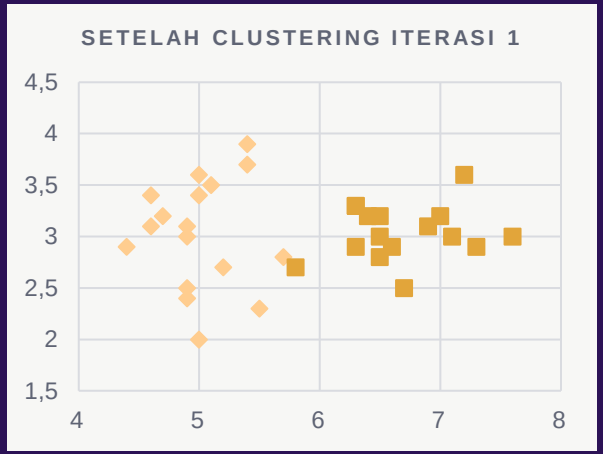
(a) Initial clustering **(b)** Iterate **(c)** Final clustering

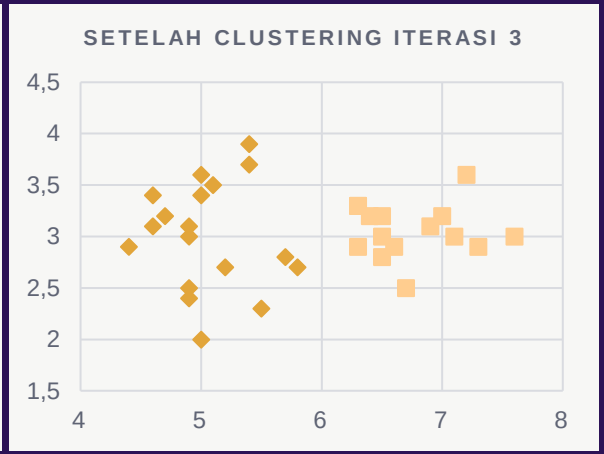
K-Means – Contoh

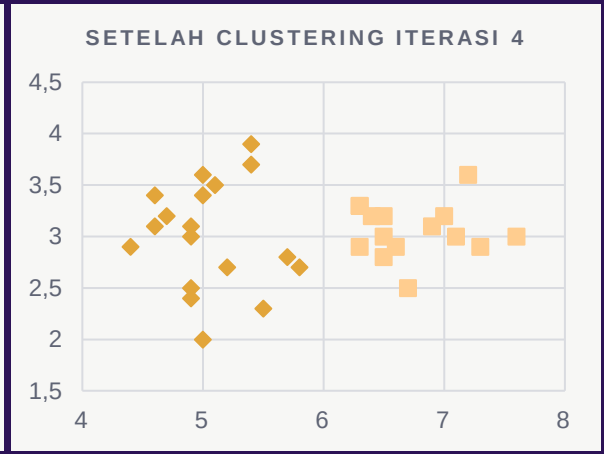




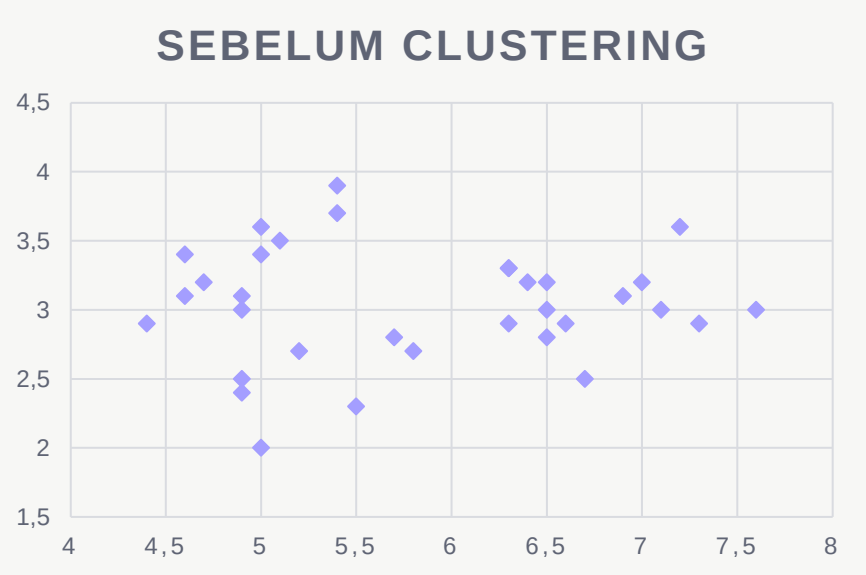
Data: klik di sini
Atribut yang digunakan:
sepal_length & sepal_width
Analisis *cluster*: klik di sini
Jumlah *cluster* = 2
Proses selesai di iterasi ke-4



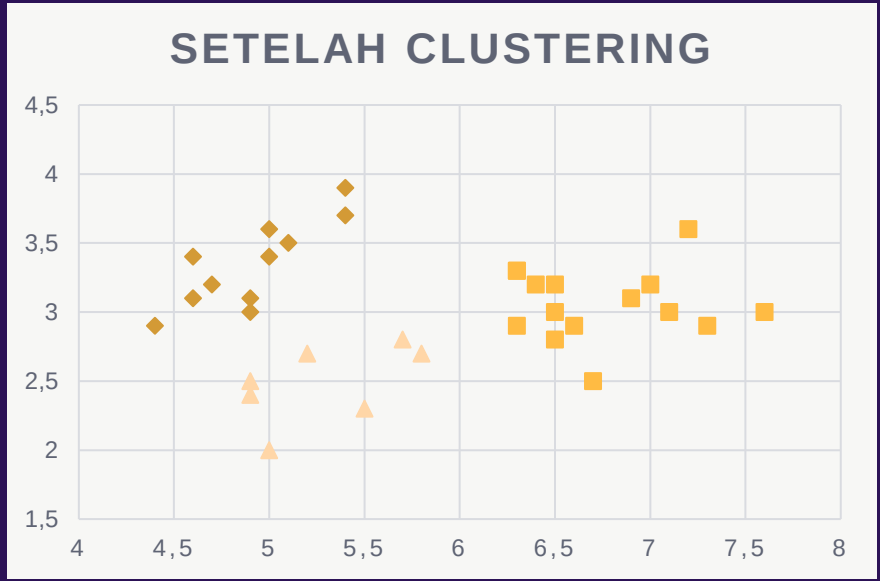




K-Means – Contoh

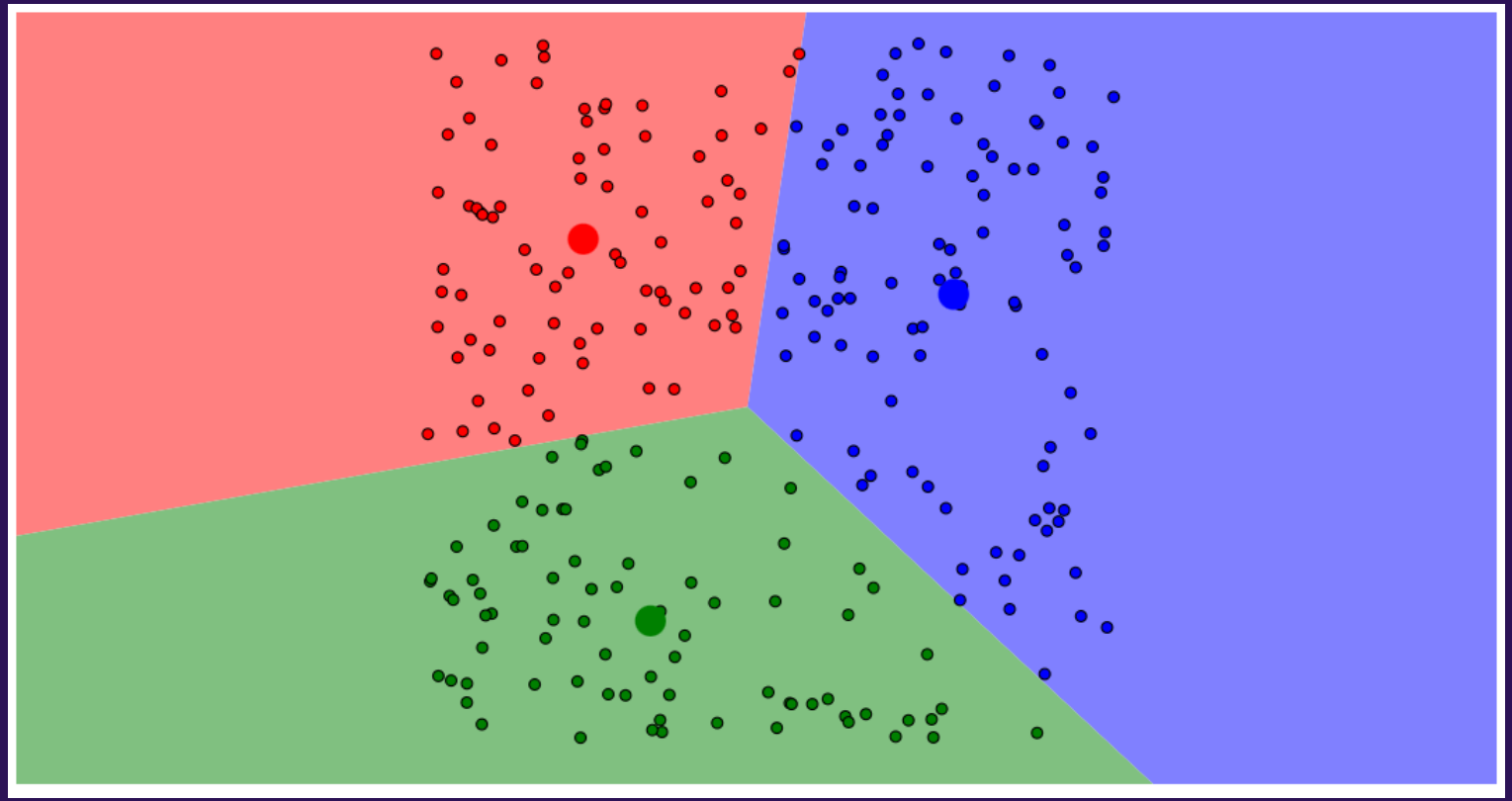


Data: klik di sini
Atribut yang digunakan: sepal_length & sepal_width
Analisis *cluster*: klik di sini
Jumlah *cluster* = 3
Proses selesai di iterasi ke-6



K-Means – Contoh

Contoh visual: https://www.naftaliharris.com/blog/visualizing-k-means-clustering/



K-Means – Beberapa Catatan

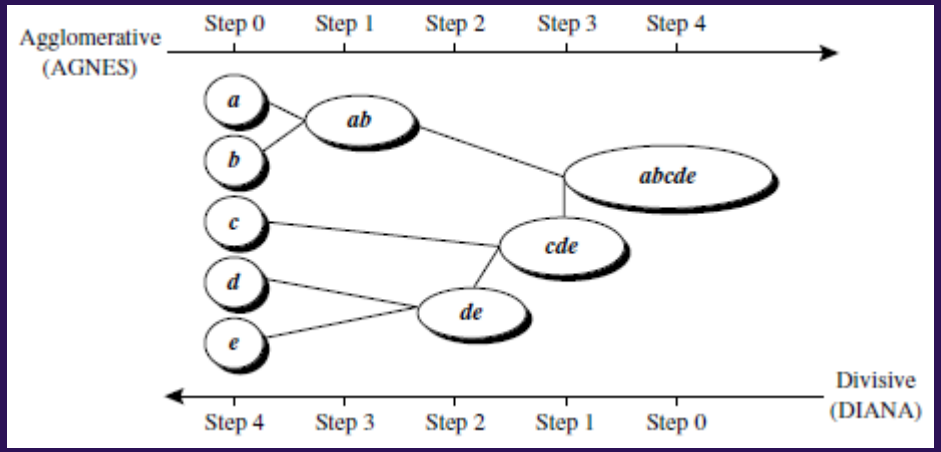
- *K-means* tidak menjamin konvergen pada *global optimum*, & terkadang prosesnya selesai di optimum lokal
- Hasil *clustering* bergantung pada pemilihan *centroid* di awal
- Hanya dapat diterapkan jika rata-rata objek dapat didefinisikan
- Sensitif terhadap derau & titik data pencilan, karena nilai derau & pencilan dapat mempengaruhi nilai rata-rata
- Varian lainnya:
  - *K-modes* → menggunakan nilai modus dari data

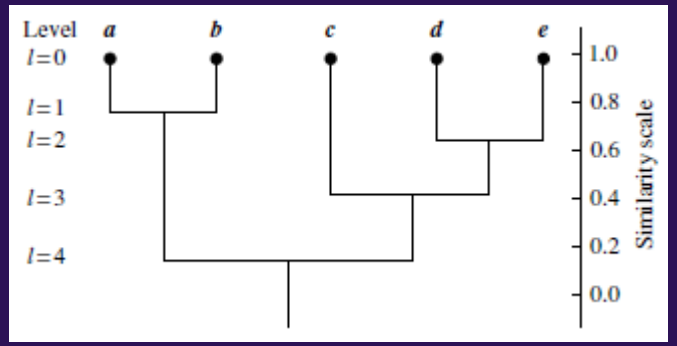
Metode *Clustering* Hirarki

- Mengelompokkan data menjadi hirarki/"pohon" *cluster*
- Dapat berupa:
  - *Agglomerative*: strategi *bottom-up*
    - ➔ Dimulai dari membuat setiap objek menjadi *cluster* sendiri, kemudian secara bertahap menggabungkan beberapa *cluster* menjadi satu *cluster* yang berukuran lebih besar.
    - ➔ Proses selesai ketika semua objek tergabung menjadi sebuah *cluster*
  - *Divisive*: strategi *top-down*
    - ➔ Dimulai dari menempatkan semua objek dalam sebuah *cluster*, kemudian secara bertahap memecah *cluster* besar menjadi *cluster* yang lebih kecil
    - ➔ Proses selesai ketika sudah terbentuk *cluster* yang paling kecil, yaitu setiap objek menjadi *cluster* sendiri

Metode *Clustering* Hirarki

- AGNES: AGglomerative NESting
- DIANA: DIvisive ANAlysis
- Dendrogram: struktur pohon yang umum digunakan untuk menggambarkan proses *clustering* hirarki





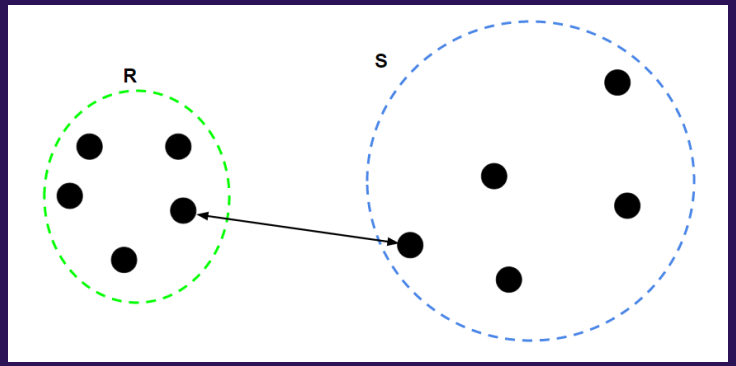
Ukuran jarak antar *cluster*:

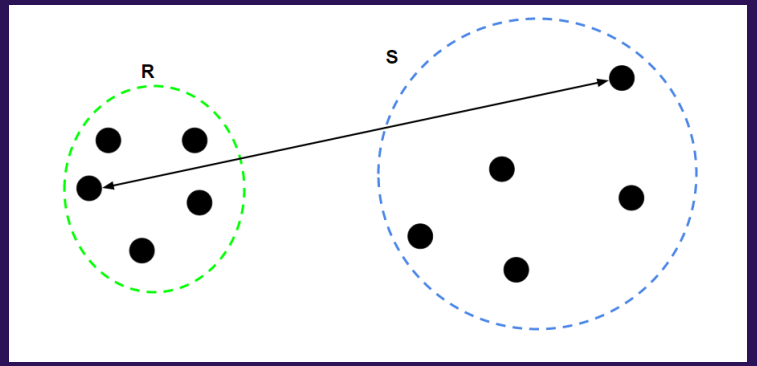
- Jarak minimum: $dist_{min}(C_i, C_j) = \min_{\boldsymbol{p} \in C_i, \boldsymbol{p}' \in C_j} \{|\boldsymbol{p} - \boldsymbol{p}'|\}$

  → Algoritma yang menggunakan jarak minimum = algoritma *single-linkage*

- Jarak maksimum: $dist_{max}(C_i, C_j) = \max_{\boldsymbol{p} \in C_i, \boldsymbol{p}' \in C_j} \{|\boldsymbol{p} - \boldsymbol{p}'|\}$

  → Algoritma yang menggunakan jarak maksimum = algoritma *complete-linkage*





Metode *Clustering* Hirarki – Contoh

Algoritma *single-linkage*

- Hitung jarak antar objek/*cluster*.
- Selama masih terdapat > 1 *cluster*:
  - Periksa tabel jarak, pilih pasangan objek yang jaraknya paling dekat, kemudian gabungkan objek-objek tersebut menjadi 1 *cluster*
  - Buat tabel jarak baru, perbarui dengan nilai jarak antar objek terdekat

	A	B	C	D	E
A	0				
B	9	0			
C	3	7	0		
D	6	5	9	0	
E	11	10	2	8	0

→

	A	B	C, E	D
A	0			
B	9	0		
C, E	3	7	0	
D	6	5	8	0

→

	A, C, E	B	D
A, C, E	0		
B	7	0	
D	6	5	0

→

	A, C, E	B, D
A, C, E	0	
B, D	6	0

Sumber: https://online.stat.psu.edu/stat555/node/86/

Metode *Clustering* Hirarki – Contoh

Algoritma *complete-linkage*

- Hitung jarak antar objek/*cluster*.
- Selama masih terdapat > 1 *cluster*:
  - Periksa tabel jarak, pilih pasangan objek yang jaraknya paling dekat, kemudian gabungkan objek-objek tersebut menjadi 1 *cluster*
  - Buat tabel jarak baru, perbarui dengan nilai jarak antar objek terjauh

	A	B	C	D	E
A	0				
B	9	0			
C	3	7	0		
D	6	5	9	0	
E	11	10	2	8	0

→

	A	B	C, E	D
A	0			
B	9	0		
C, E	11	10	0	
D	6	5	9	0

→

	A	B, D	C, E
A	0		
B, D	9	0	
C, E	11	10	0

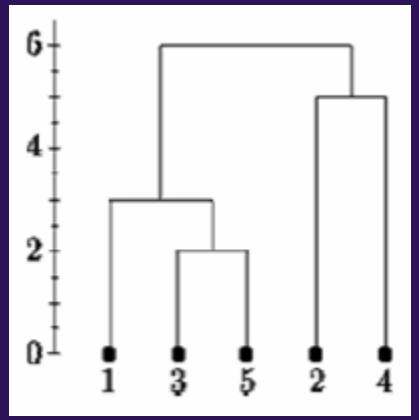
↘

	A, B, D	C, E
A, B, D	0	
C, E	11	0

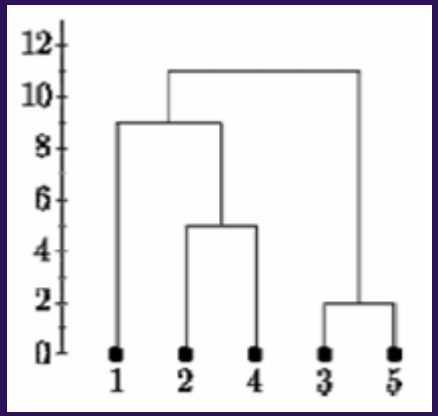
Sumber: https://online.stat.psu.edu/stat555/node/86/

Metode *Clustering* Hirarki – Contoh

Dendrogram:



Single-linkage



Complete-linkage

Permasalahan: “Berapa jumlah *cluster* yang terbentuk dari *clustering* hirarki?”

➔ Tidak ada cara yang objektif untuk menyatakan jumlah *cluster* yang terbentuk

➔ Penentuan jumlah *cluster* dilakukan secara subjektif

Soal Latihan

Diberikan sebuah *worksheet*, berisi 15 buah objek yang memiliki 2 atribut. Akan diterapkan *clustering k-means* dengan jumlah *cluster* 2 buah pada objek-objek tersebut.

Tugas :

1. Tentukan titik awal *centroid* 1 & 2, sebelum melakukan iterasi pertama *k-means*
2. Lakukan proses algoritma *k-means* sebanyak 5 iterasi di *file worksheet* tersebut
3. Gambarkan *scatter plot* dari hasil *k-means* iterasi ke-5

Worksheet dapat diakses di sini

Recap

- Definisi *clustering*
- *Clustering* vs Klasifikasi
- Varian metode *clustering*
- *K-Means*
- *Clustering* hirarki

Next: apakah ada materi yang ingin dibahas kembali?

つづく