



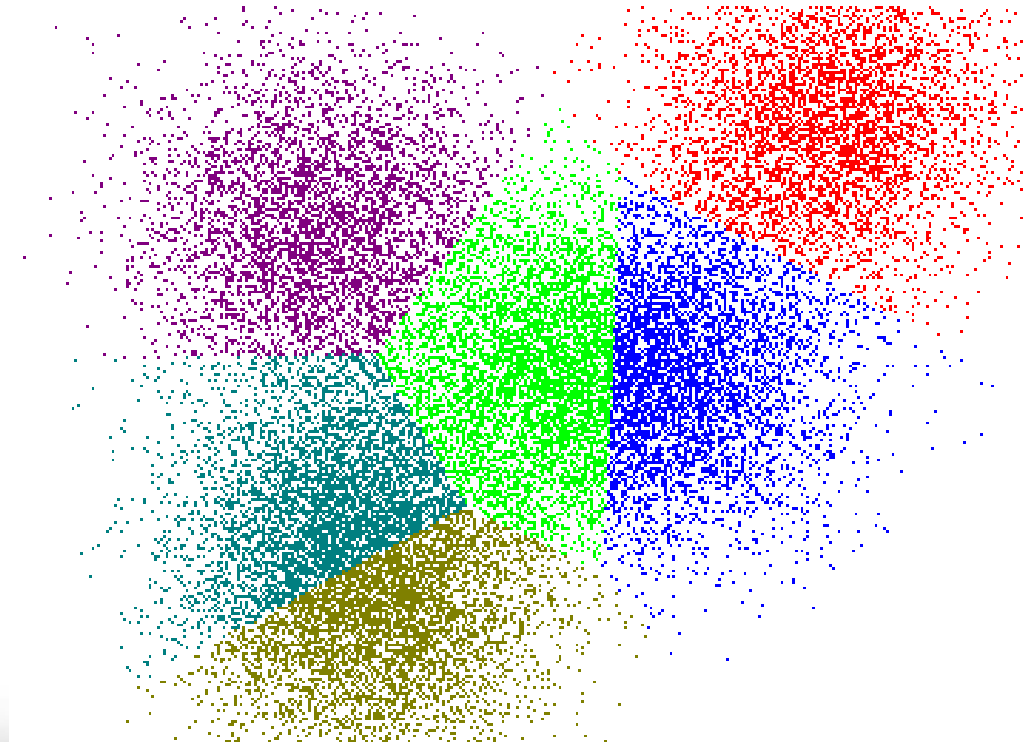
www.esaunggul.ac.id

**CLUSTERING
PERTEMUAN - 5
NOVIANDI**

PRODI MIK | FAKULTAS ILMU-ILMU KESEHATAN

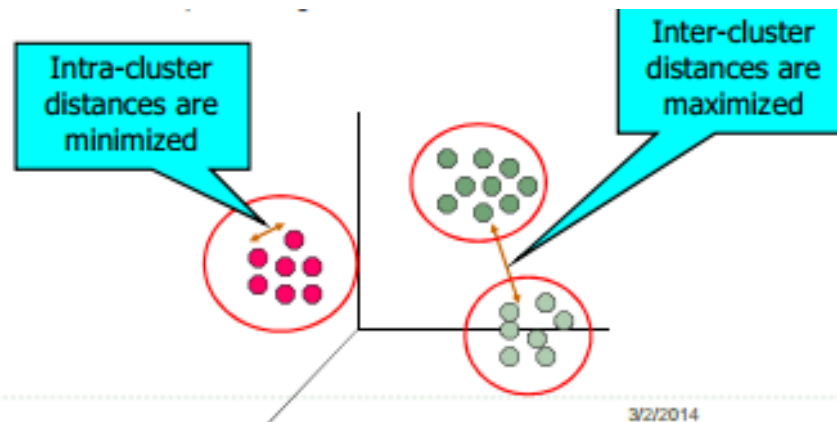
KEMAMPUAN AKHIR YANG DIHARAPKAN

Dapat menjelaskan konsep dasar cluster dan penerapannya pada data.



CLUSTERING

- ❖ *Clustering* adalah salah satu teknik *unsupervised learning* dimana tidak ada fase *learning*. *Cluster* berguna untuk mengelompokkan objek-objek data yang memiliki kemiripan ke dalam satu grup dan yang berbeda dikelompokkan ke dalam grup lainnya
- ❖ Semakin besar tingkat kemiripan/*similarity* (atau homogenitas) di dalam satu grup dan semakin besar tingkat perbedaan diantara grup, maka semakin baik (atau lebih berbeda) *clustering* tersebut.



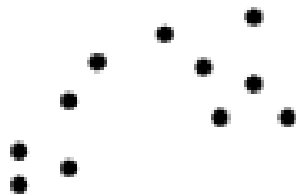
APLIKASI CLUSTERING

- Mengelompokkan dokumen-dokumen yang relatif, mengelompokkan gen dan protein yang memiliki fungsi yang sama.
- Mengurangi ukuran data yang besar

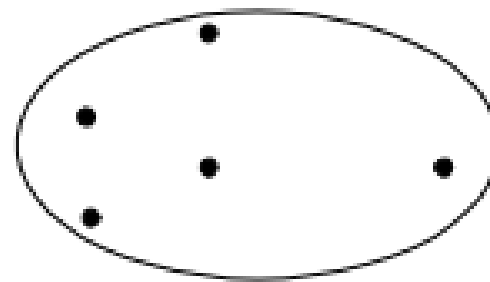
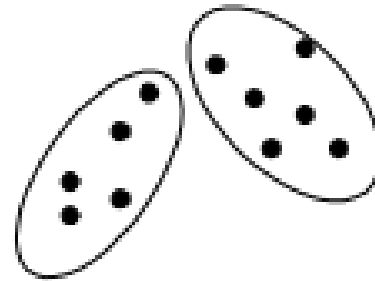
TIPE-TIPE *CLUSTERING*

- ❑ *Partitional clustering* adalah himpunan obyek data ke dalam sub-himpunan (cluster) yang tidak overlap, sehingga setiap obyek data berada dalam tepat satu cluster.
- ❑ *Hierarchical clustering* adalah cluster yang memiliki subcluster. Himpunan cluster besarang yang diatur dalam tree.

PARTITIONAL CLUSTERING

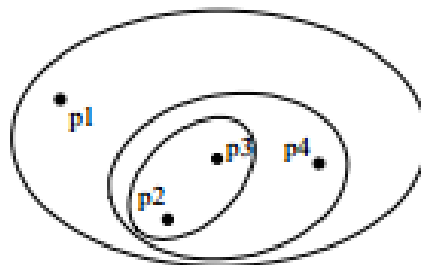


Original Points

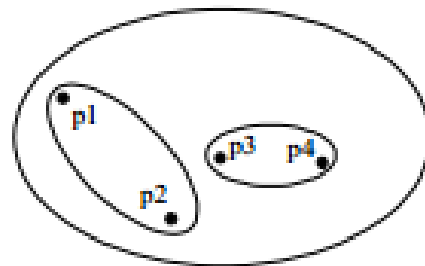


A Partitional Clustering

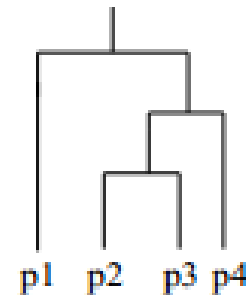
HIERARCHICAL CLUSTERING



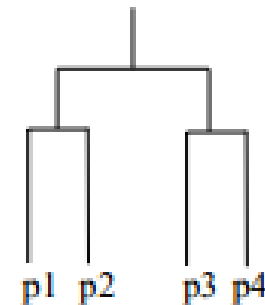
Traditional Hierarchical Clustering



Non-traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Dendrogram

SIMILARITY AND DISSIMILARITY BETWEEN OBJECTS

- Jarak biasanya digunakan untuk mengukur kemiripan dan ketidakmiripan diantara dua objek
- Rumus untuk mengukur jarak diantara dua objek:

$$d(i, j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)}$$

where $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ are two p -dimensional data objects, and q is a positive integer

Jika nilai $q=1$, maka jarak tersebut diukur dengan *Manhattan distance*

SIMILARITY AND DISSIMILARITY BETWEEN OBJECTS

Jika $q=2$, maka jarak tersebut diukur dengan Euclidean distance

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

- $d(i,j) \geq 0$
- $d(i,j) = 0$
- $d(i,j) = d(j,i)$

ALGORITMA CLUSTERING

- *K-Means*
- *K-Medoids*
- *Hierarchical Clustering*

K-MEANS CLUSTERING

- ❑ Pendekatan *partitional clustering*
- ❑ Setiap cluster diasosiasikan dengan sentroid
- ❑ Setiap titik di tandai ke *cluster* dengan sentroid terdekat
- ❑ K menandakan jumlah cluster yang akan terbentuk
- ❑ Algoritma *Clustering*:
 1. Menentukan jumlah cluster
 2. Menentukan nilai centroid biasanya dilakukan secara random atau biasanya menggunakan rumus rata-rata
 3. Menghitung jarak antara titik centroid dengan titik tiap objek. Biasanya menggunakan jarak Euclidean distance.
 4. Mengelompokkan objek berdasarkan jarak terdekat
 5. Kembali ke tahap ke 2 dan lakukan perulangan hingga nilai centroid yang dihasilkan tetap dan anggota cluster tidak berpindah ke cluster lain.

CONTOH SOAL

DATA	X	Y
M1	2	5.0
M2	2	5.5
M3	5	3.5
M4	6.5	2.2
M5	7	3.3
M6	3.5	4.8
M7	4	4.5

$C1=(3,4)$ dan $C2=(6,4)$

CONTOH SOAL

Iterasi 1

a. Menghitung Euclidean distance dari semua data ke tiap titik pusat pertama

$$D_{11} = \sqrt{(M_{1x} - C_{1x})^2 + (M_{1y} - C_{1y})^2} = \sqrt{(2 - 3)^2 + (5 - 4)^2} = \sqrt{2} = 1.41$$

$$D_{12} = \sqrt{(M_{2x} - C_{1x})^2 + (M_{2y} - C_{1y})^2} = \sqrt{(2 - 3)^2 + (5.5 - 4)^2} = \sqrt{3.25} = 1.80$$

$$D_{13} = \sqrt{(M_{3x} - C_{1x})^2 + (M_{3y} - C_{1y})^2} = \sqrt{(5 - 3)^2 + (3.5 - 4)^2} = \sqrt{4.25} = 2.06$$

$$D_{14} = \sqrt{(M_{4x} - C_{1x})^2 + (M_{4y} - C_{1y})^2} = \sqrt{(6.5 - 3)^2 + (2.2 - 4)^2} = \sqrt{2} = 3.94$$

$$D_{15} = \sqrt{(M_{5x} - C_{1x})^2 + (M_{5y} - C_{1y})^2} = \sqrt{(7 - 3)^2 + (3.3 - 4)^2} = \sqrt{2} = 4.06$$

$$D_{16} = \sqrt{(M_{6x} - C_{1x})^2 + (M_{6y} - C_{1y})^2} = \sqrt{(3.5 - 3)^2 + (4.8 - 4)^2} = \sqrt{2} = 0.94$$

$$D_{17} = \sqrt{(M_{7x} - C_{1x})^2 + (M_{7y} - C_{1y})^2} = \sqrt{(4 - 3)^2 + (4.5 - 4)^2} = \sqrt{2} = 1.12$$

Dengan cara yang sama hitung jarak tiap titik ke titik pusat ke dan kita akan mendapatkan $D_{21} = 4.12$, $D_{22} = 4.27$, $D_{23} = 1.18$, $D_{24} = 1.86$, $D_{25} = 1.22$, $D_{26} = 2.62$, $D_{27} = 2.06$

CONTOH SOAL

Iterasi 1

b. Dari perhitungan Euclidean distance, kita dapat membandingkan

DATA	C1	C2
M1	1.41	4.12
M2	1.80	4.27
M3	2.06	1.18
M4	3.94	1.86
M5	4.06	1.22
M6	0.94	2.62
M7	1.12	2.06

$\{M_1, M_2, M_6, M_7\}$ anggota C_1 and $\{M_3, M_4, M_5\}$ anggota C_2

CONTOH SOAL

Iterasi 1

c. Hitung titik pusat baru

$$C_1 = \left(\frac{2 + 2 + 3 + 4}{4}, \frac{5 + 5.5 + 4.8 + 4.5}{4} \right) = (2.75, 4.9)$$

$$C_2 = \left(\frac{5 + 6.5 + 7}{3}, \frac{3.5 + 2.2 + 3.3}{3} \right) = (6.17, 3)$$

Lakukan iterasi ke 2 seperti iterasi ke 1, sampai anggota kelompok C1 dan C2 tidak berubah lagi seperti sebelumnya,

Kesimpulan $\{M_1, M_2, M_6, M_7\}$ anggota C_1 dan $\{M_3, M_4, M_5\}$ anggota C_2

HIERARCHICAL CLUSTERING

Strategi pengelompokkannya umumnya ada dua jenis, yaitu:

- Agglomerative (Bottom-Up)
- Devisive (Top-Down)

Algoritma Agglomerative Hierarchical Clustering :

1. Hitung Matrik Jarak antar data.
2. Ulangi langkah 3 dan 4 hingga hanya satu kelompok yang tersisa.
3. Gabungkan dua kelompok terdekat berdasarkan metode pengelompokan (*Single Linkage*, *Complete Linkage*, *Average Linkage*)
4. Perbarui Matrik Jarak antar data untuk merepresentasikan kedekatan diantara kelompok baru dan kelompok yang masih tersisa.
5. Selesai

Metode Pengelompokan Hierarki Aglomeratif

Beberapa metode pengelompokan secara hierarki Aglomeratif:

- ❑ Single Linkage (Jarak Terdekat)

$$d_{uv} = \min\{d_{uv}\}, d_{uv} \in D$$

- ❑ Complete Linkage (Jarak Terjauh)

$$d_{uv} = \max\{d_{uv}\}, d_{uv} \in D$$

- ❑ Average Linkage (Jarak rata-rata)

$$d_{uv} = \text{average}\{d_{uv}\}, d_{uv} \in D$$

CONTOH STUDI KASUS

Data	Fitur x	Fitur y
1	1	1
2	4	1
3	1	2
4	3	4
5	5	4

Kelompokkan dataset tersebut dengan menggunakan metode AHC (Single Linkage) menggunakan jarak Manhattan!

CONTOH STUDI KASUS

- Menghitung Jarak Pada Semua Pasangan dua data :

Data	Fitur x	Fitur y
1	1	1
2	4	1
3	1	2
4	3	4
5	5	4

$$D_{\text{man}}(\text{Data}_1, \text{Data}_1) = \sum_{j=1}^2 |x_j - y_j| = |1-1| + |1-1| = 0$$

$$D_{\text{man}}(\text{Data}_1, \text{Data}_2) = |1-4| + |1-1| = 3$$

$$D_{\text{man}}(\text{Data}_1, \text{Data}_3) = |1-1| + |1-2| = 1$$

$$D_{\text{man}}(\text{Data}_1, \text{Data}_4) = |1-3| + |1-4| = 2 + 3 = 5$$

$$D_{\text{man}}(\text{Data}_3, \text{Data}_4) = |1-3| + |2-4| = 2 + 2 = 4$$

$$D_{\text{man}}(\text{Data}_1, \text{Data}_5) = |1-5| + |1-4| = 4 + 3 = 7$$

$$D_{\text{man}}(\text{Data}_3, \text{Data}_5) = |1-5| + |2-4| = 4 + 2 = 6$$

$$D_{\text{man}}(\text{Data}_2, \text{Data}_3) = |4-1| + |1-2| = 3 + 1 = 4$$

$$D_{\text{man}}(\text{Data}_4, \text{Data}_5) = |3-5| + |4-4| = 2 + 0 = 2$$

$$D_{\text{man}}(\text{Data}_2, \text{Data}_4) = |4-3| + |1-4| = 1 + 3 = 4$$

$$D_{\text{man}}(\text{Data}_2, \text{Data}_5) = |4-5| + |1-4| = 1 + 3 = 4$$

CONTOH STUDI KASUS

Hasil Matrik Jarak :

Dman	1	2	3	4	5
1	0	3	1	5	7
2	3	0	4	4	4
3	1	4	0	4	6
4	5	4	4	0	2
5	7	4	6	2	0

- Menggunakan Metode Single Linkage :

Dengan memperlakukan data sebagai kelompok, selanjutnya kita pilih jarak dua kelompok yang terkecil.

D _{man}	1	2	3	4	5
1	0	3	1	5	7
2	3	0	4	4	4
3	1	4	0	4	6
4	5	4	4	0	2
5	7	4	6	2	0

$$\min(D_{man}) = \min(d_{13}) = 1$$

terpilih kelompok 1 dan 3, sehingga kedua kelompok ini digabungkan. (Melanjutkan pengelompokan).

- Menghitung jarak antar kelompok (1 dan 3) dengan kelompok lain yang tersisa, yaitu 2, 4 dan 5.

$$d_{(13)2} = \min\{d_{12}, d_{32}\} = \min\{3, 4\} = 3$$

$$d_{(13)4} = \min\{d_{14}, d_{34}\} = \min\{5, 4\} = 4$$

$$d_{(13)5} = \min\{d_{15}, d_{35}\} = \min\{7, 6\} = 6$$

- Menghitung jarak antar kelompok (1 dan 3) dengan kelompok lain yang tersisa, yaitu 2, 4 dan 5.

$$d_{(13)2} = \min\{d_{12}, d_{32}\} = \min\{3, 4\} = 3 \quad \text{●}$$

$$d_{(13)4} = \min\{d_{14}, d_{34}\} = \min\{5, 4\} = 4 \quad \text{●}$$

$$d_{(13)5} = \min\{d_{15}, d_{35}\} = \min\{7, 6\} = 6 \quad \text{●}$$

- Dengan menghapus baris-baris dan kolom-kolom matrik jarak yang bersesuaian dengan kelompok 1 dan 3, serta menambahkan baris dan kolom untuk kelompok (13)

Dman	1	2	3	4	5
1	0	3	3	5	7
2	3	0	4	4	4
3	3	4	0	4	6
4	5	4	4	0	2
5	7	4	6	2	0

→


Dman	(13)	2	4	5
(13)	0	3	4	6
2	3	0	4	4
4	4	4	0	2
5	6	4	2	0

Selanjutnya dipilih jarak dua kelompok yang terkecil.

$$\min(D_{man}) = \min(d_{45}) = 2$$

- Dengan menghapus baris-baris dan kolom-kolom matrik jarak yang bersesuaian dengan kelompok 1 dan 3, serta menambahkan baris dan kolom untuk kelompok (13).

Dman	1	2	3	4	5
1	0	3	1	5	7
2	3	0	4	4	4
3	1	4	0	4	6
4	5	4	4	0	2
5	7	4	6	2	0



Dman	(13)	2	4	5
(13)	0	3	4	6
2	3	0	4	4
4	4	4	0	2
5	6	4	2	0

- Selanjutnya dipilih jarak dua kelompok yang terkecil.

$$\min(D_{man}) = \min(d_{45}) = 2$$

- Menghitung jarak antar kelompok (4 dan 5) dengan kelompok lain yang tersisa, yaitu (13) dan 2.

$$d_{(45)(13)} = \min\{d_{41}, d_{43}, d_{51}, d_{53}\} = \min\{5, 4, 7, 6\} = 4$$

$$d_{(45)2} = \min\{d_{42}, d_{52}\} = \min\{4, 4\} = 4$$

- Menghapus baris dan kolom matrik yang bersesuaian dengan kelompok 4 dan 5, serta menambahkan baris dan kolom untuk kelompok (45)

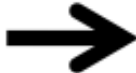
- Menghitung jarak antar kelompok (4 dan 5) dengan kelompok lain yang tersisa, yaitu (13) dan 2.

$$d_{(45)(13)} = \min\{d_{41}, d_{43}, d_{51}, d_{53}\} = \min\{5, 4, 7, 6\} = 4 \quad \bullet$$

$$d_{(45)2} = \min\{d_{42}, d_{52}\} = \min\{4, 4\} = 4 \quad \bullet$$

- Menghapus baris dan kolom matrik yang bersesuaian dengan kelompok 4 dan 5, serta menambahkan baris dan kolom untuk kelompok (45)

Dman	(13)	2	4	5
(13)	0	3	4	6
2	3	0	4	4
4	4	4	0	2
5	6	4	2	0



Dman	(45)	(13)	2
(45)	0	4	4
(13)	4	0	3
2	4	3	0

- Selanjutnya dipilih jarak dua kelompok yang terkecil.

$$\min(D_{man}) = \min(d_{(13)2}) = 3$$

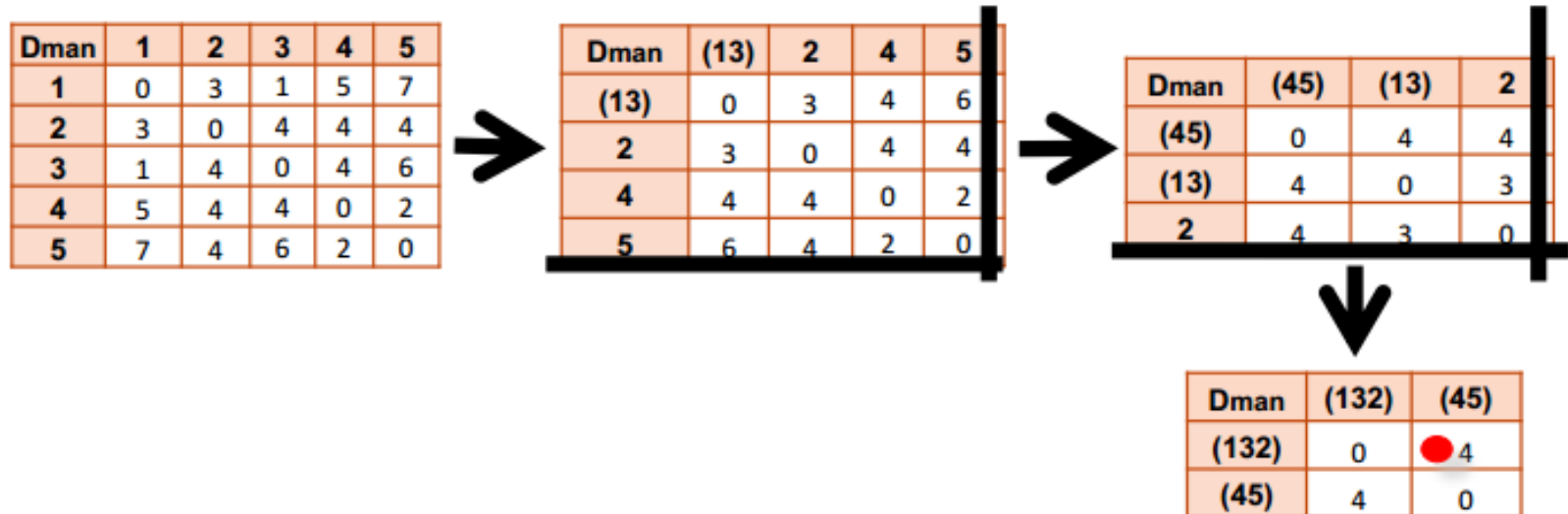
terpilih kelompok (13) dan 2, sehingga kedua kelompok ini digabungkan. (Melanjutkan pengelompokan).

- Menghitung jarak antar kelompok ((13) dan 2) dengan kelompok lain yang tersisa, yaitu (45).

- Menghitung jarak antar kelompok ((13) dan 2) dengan kelompok lain yang tersisa, yaitu (45).

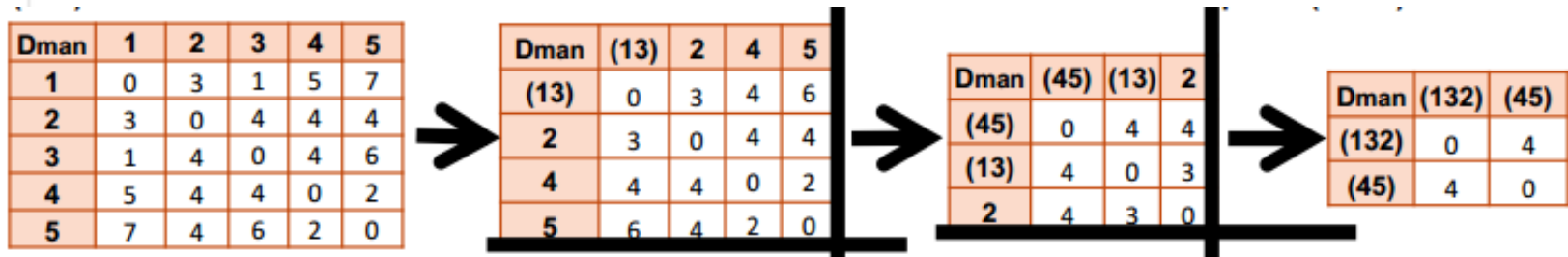
$$d_{(132)(45)} = \min\{d_{14}, d_{15}, d_{34}, d_{35}, d_{24}, d_{25}\} = \min\{5, 7, 4, 6, 4, 4\} = 4 \quad \bullet$$

- Menghapus baris dan kolom matrik yang bersesuaian dengan kelompok (13) dan 2, serta menambahkan baris dan kolom untuk kelompok (123)

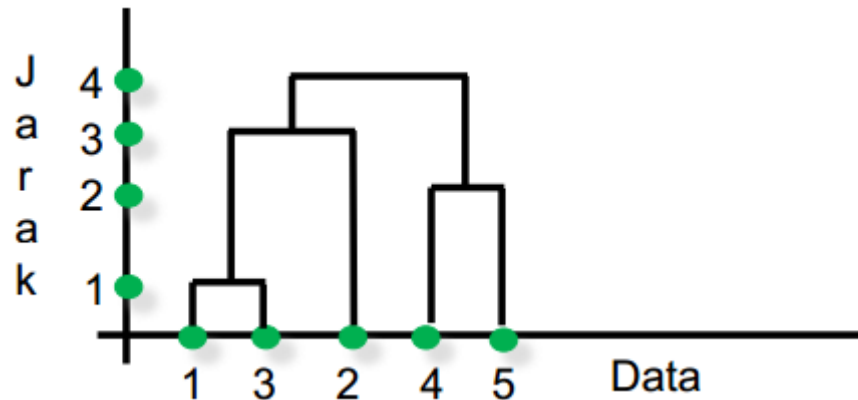


- Jadi kelompok (132) dan (45) digabung untuk menjadi kelompok tunggal dari lima data, yaitu kelompok (13245) dengan jarak terdekat 4.

- Menghapus baris dan kolom matrik yang bersesuaian dengan kelompok (13) dan 2, serta menambahkan baris dan kolom untuk kelompok (132).



- Jadi kelompok (132) dan (45) digabung untuk menjadi kelompok tunggal dari lima data, yaitu kelompok (13245) dengan jarak terdekat 4. Berikut Dendrogram Hasil Metode Single Linkage :



TERIMA KASIH 😊