Unsupervised Learning-K-means Clustering

Tujuan Pembelajaran

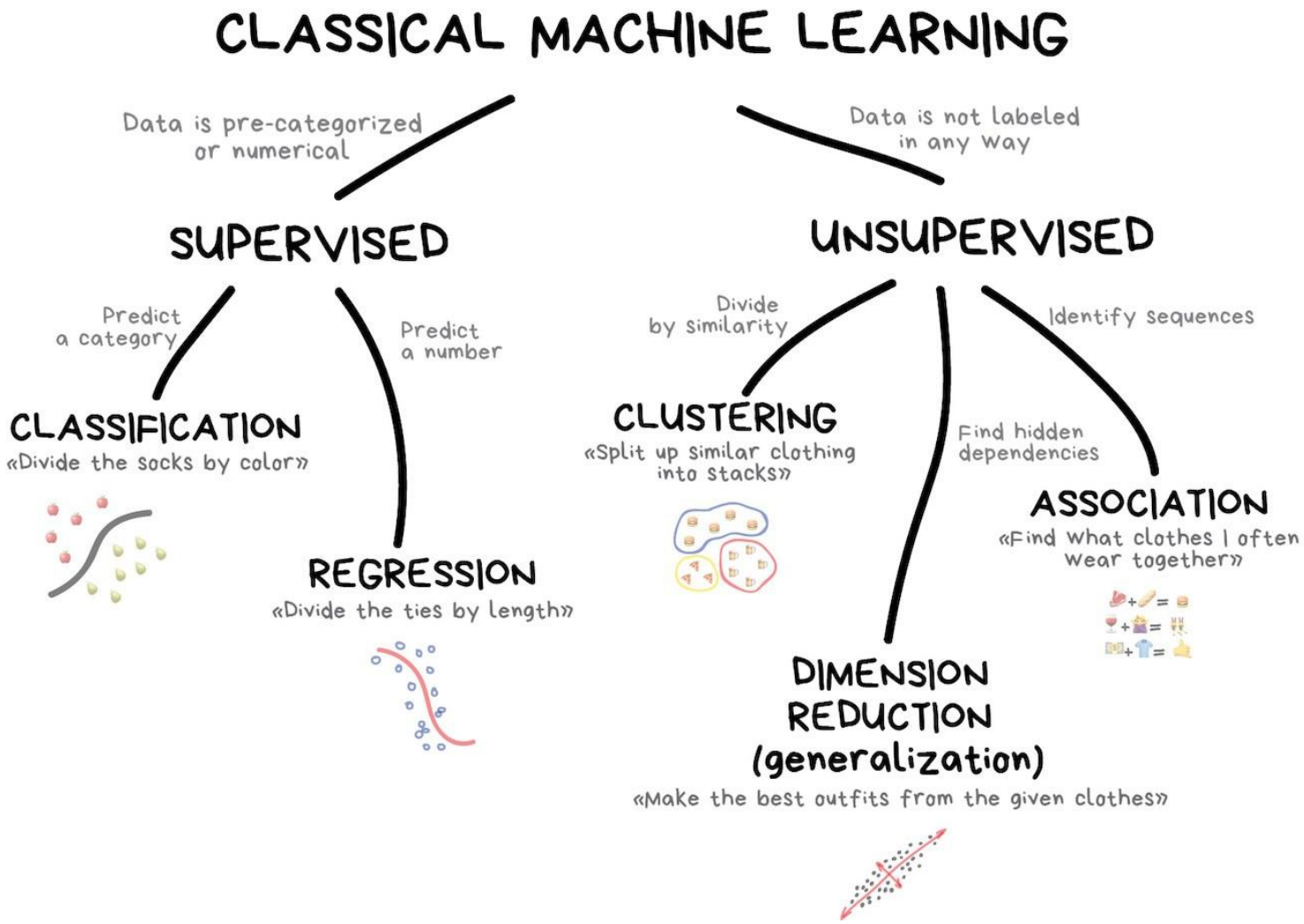
- Mahasiswa dapat memahami konsep pembelajaran tidak tersupervisi (unsupervised learning), khususnya menggunakan metode K-Means clustering
- Mahasiswa dapat menjelaskan beberapa penerapan algoritma klastering dalam menyelesaikan berbagai masalah.

Outline

- Pengantar Clustering
- Penerapan Algoritma Clustering
- Macam-macam Algoritma Clustering
- Langkah Algoritma K-Means Clustering
- Contoh Perhitungan Algoritma K-Means
- Optimasi Nilai k pada K-Means

Supervised vs Unsupervised Learning

- Supervised
  - Classification
  - Regression
- Unsupervised
  - Clustering
  - Association
  - Dimension Reduction



Pengantar

- Diberikan data profil pelanggan (customer), bagaimana memilih data pelanggan yang potensial untuk ditawarkan produk tertentu?

	Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Defaulted	Address	DebtIncomeRatio
0	1	41	2	6	19	0.124	1.073	0.0	NBA001	6.3
1	2	47	1	26	100	4.582	8.218	0.0	NBA021	12.8
2	3	33	2	10	57	6.111	5.802	1.0	NBA013	20.9
3	4	29	2	4	19	0.681	0.516	0.0	NBA009	6.3
4	5	47	1	31	253	9.308	8.908	0.0	NBA008	7.2
...	...	...	...	...	...	...	...	...	...	...
845	846	27	1	5	26	0.548	1.220	NaN	NBA007	6.8
846	847	28	2	7	34	0.359	2.021	0.0	NBA002	7.0
847	848	25	4	0	18	2.802	3.210	1.0	NBA001	33.4
848	849	32	1	12	28	0.116	0.696	0.0	NBA012	2.9
849	850	52	1	16	64	1.866	3.638	0.0	NBA025	8.6

850 rows × 10 columns

Kita diminta untuk mengelompokkan data customer di samping, berdasarkan kesamaan profil pelanggan

Customer Segmentation

Clustering

Pengantar

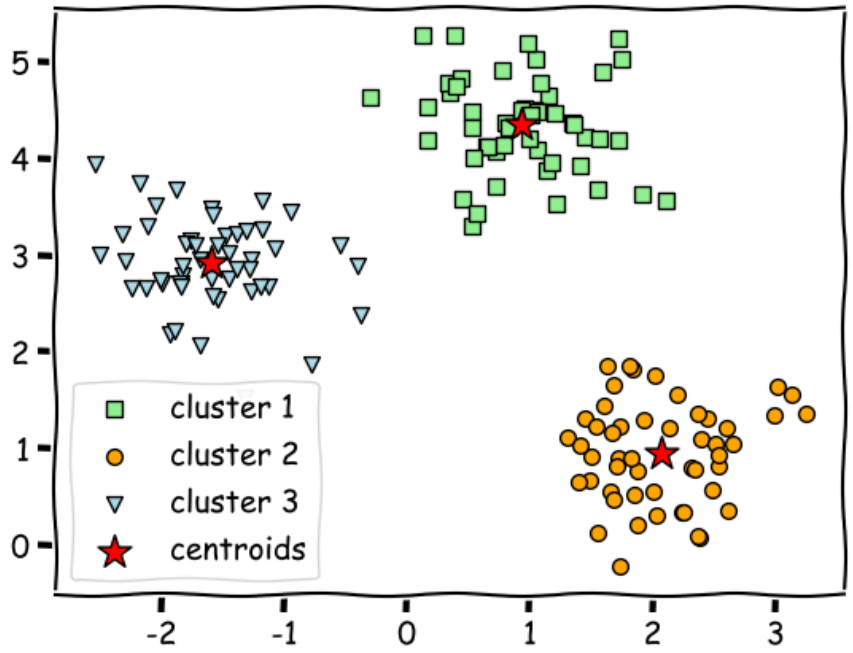
- Contoh hasil clustering / segmentasi pelanggan

	Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Defaulted	DebtIncomeRatio	Cluster
0	1	41	2	6	19	0.124	1.073	0.0	6.3	2
1	2	47	1	26	100	4.582	8.218	0.0	12.8	0
2	3	33	2	10	57	6.111	5.802	1.0	20.9	2
3	4	29	2	4	19	0.681	0.516	0.0	6.3	2
4	5	47	1	31	253	9.308	8.908	0.0	7.2	1
...	...	...	...	...	...	...	...	...	...	...
845	846	27	1	5	26	0.548	1.220	NaN	6.8	2
846	847	28	2	7	34	0.359	2.021	0.0	7.0	2
847	848	25	4	0	18	2.802	3.210	1.0	33.4	2
848	849	32	1	12	28	0.116	0.696	0.0	2.9	2
849	850	52	1	16	64	1.866	3.638	0.0	8.6	0

Setiap pelanggan berhasil dikelompokkan

Apa itu Clustering?

- **Cluster** adalah sekumpulan data / object yang memiliki **kesamaan (similarity)** diantara setiap anggota klaster, atau **ketidaksamaan (dissimilarity)** dengan data pada klaster yang lain



Contoh Penerapan Clustering

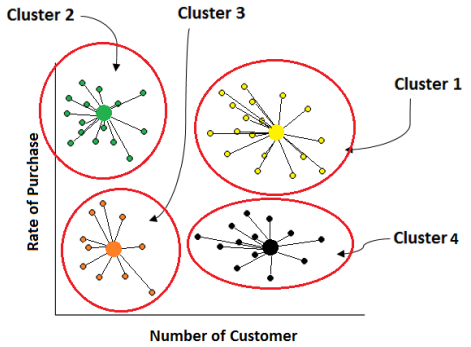
- Retail / Marketing
  - Analisis pola transaksi yang dilakukan pelanggan
  - Rekomendasi buku, film atau produk baru untuk pelanggan baru
- Perbankan
  - Deteksi fraud dalam transaksi perbankan
  - Pengelompokan nasabah (program loyalitas nasabah)
- Asuransi
  - Deteksi fraud dalam klaim asuransi
  - Analisis resiko asuransi bagi pelanggan
- Berita dan Penerbitan
  - Kategorisasi berita secara otomatis
  - Rekomendasi artikel / berita baru

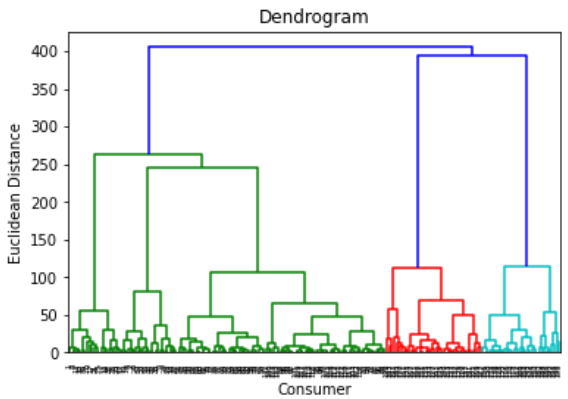
Penggunaan Algoritma Clustering

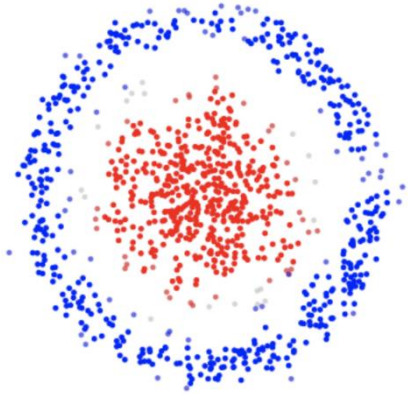
- Exploratory Data Analysis (EDA)
- Generate Rangkuman (summary generation)
- Deteksi pencilan (outlier detection)
- Mencari duplikat (finding duplicates)
- Tahap pra-pemrosesan data (Data pre-processing)
- Kompresi data / image
- Optimasi algoritma k-NN
- dll

Algoritma Clustering

- Partitioning-based clustering
  - K-Means,
  - K-Medoid,
  - K-Medians,
  - Fuzzy C-Means, dll
- Hierarchical Clustering
  - Agglomerative
  - Divisive, dll
- Density-based Clustering
  - DBSCAN, dll







Partitioning-based Clustering

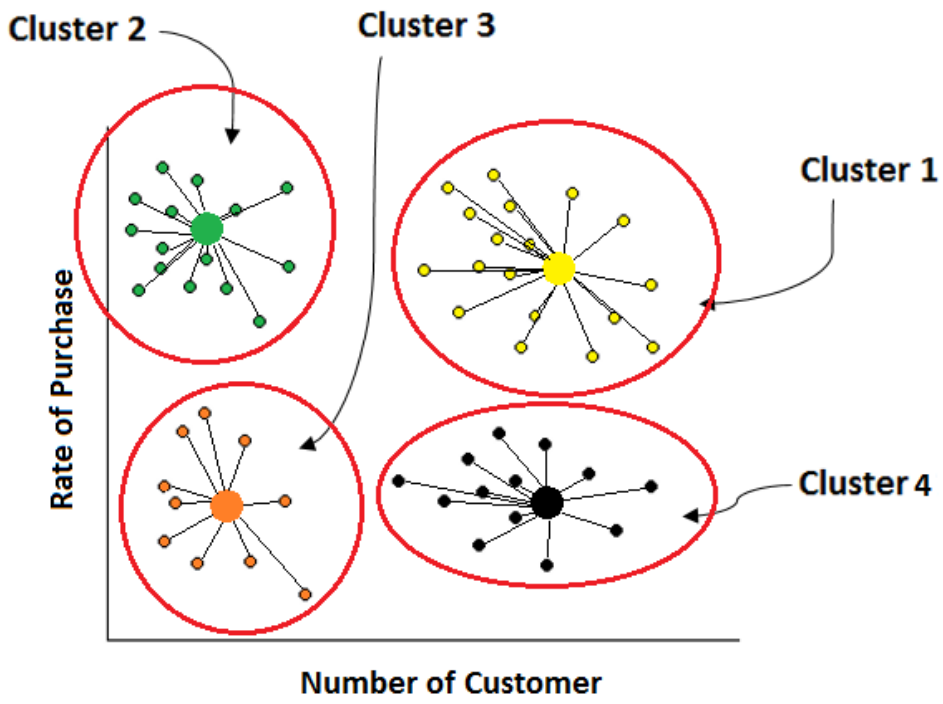
Algoritma K-means

K-Means Clustering

- Algoritma K-Means adalah salah satu algoritma clustering yang bersifat iteratif yang mencoba untuk mempartisi dataset menjadi subkelompok non-overlapping berbeda yang ditentukan oleh K (cluster) yang mana setiap titik data hanya dimiliki oleh satu kelompok.
- K-Means mencoba membuat titik data intra-cluster semirip mungkin dengan titik data yang lain pada satu cluster.
- K-Means menetapkan poin data ke cluster sedemikian rupa sehingga jumlah jarak kuadrat antara titik data dan pusat massa cluster (centroid) adalah minimal.
- Semakin sedikit variasi dalam sebuah cluster, semakin homogen (serupa) titik data dalam cluster yang sama.

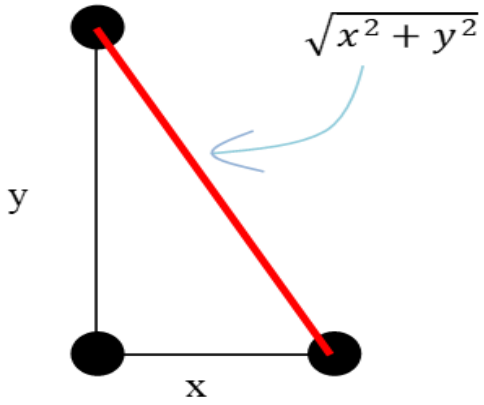
K-Means Clustering: Similarity / Dissimilarity

- Intra-cluster:
  - Memaksimalkan similarity (kesamaan) di dalam klaster
  - Meminimalkan dissimilarity di dalam klaster
- Inter-cluster:
  - Meminimalkan similarity antar-klaster
  - Memaksimalkan dissimilarity antar-klaster

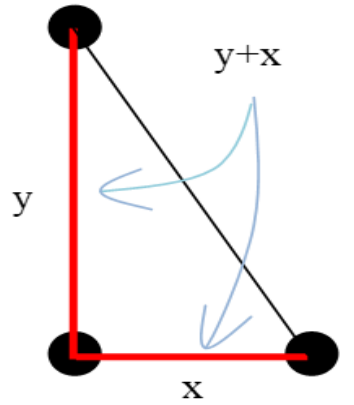


K-Means Clustering: Metode Perhitungan Similarity

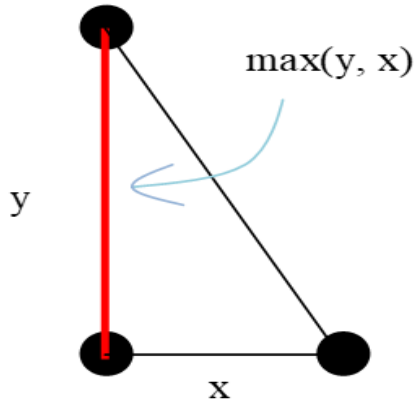
- Jarak Euclidean
- Jarak City-Block
- Jarak Kotak Catur (Chebychef)
- Jarak Minkowski
- Jarak Canberra
- Jarak Bray-Curtis (Sorensen)
- Divergensi Kullback Leibler
- Divergensi Jensen Shannon
- dll



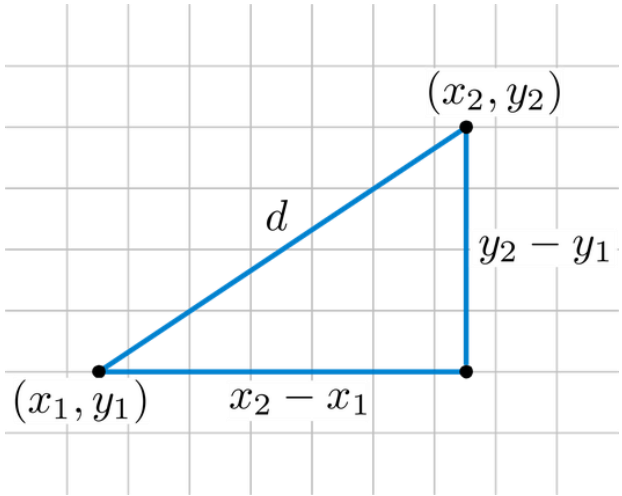
(a) jarak *Euclidean*



(b) Jarak *city-block*



(c) Jarak *Chebychef*



$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Langkah / Algoritma K-Means Clustering

1. Tentukan jumlah klaster (nilai k)
2. Inisialisasi nilai centroid awal setiap klaster secara acak
3. Hitung jarak setiap titik data dengan setiap centroid
4. Masukkan setiap titik data ke dalam klaster berdasarkan jarak terdekat dengan pusat klaster
5. Untuk setiap klaster, tentukan nilai centroid baru berdasarkan rerata (means) dari setiap data di dalam klaster
6. Ulangi langkah 3-5 sedemikian hingga tidak ada perubahan anggota klaster.

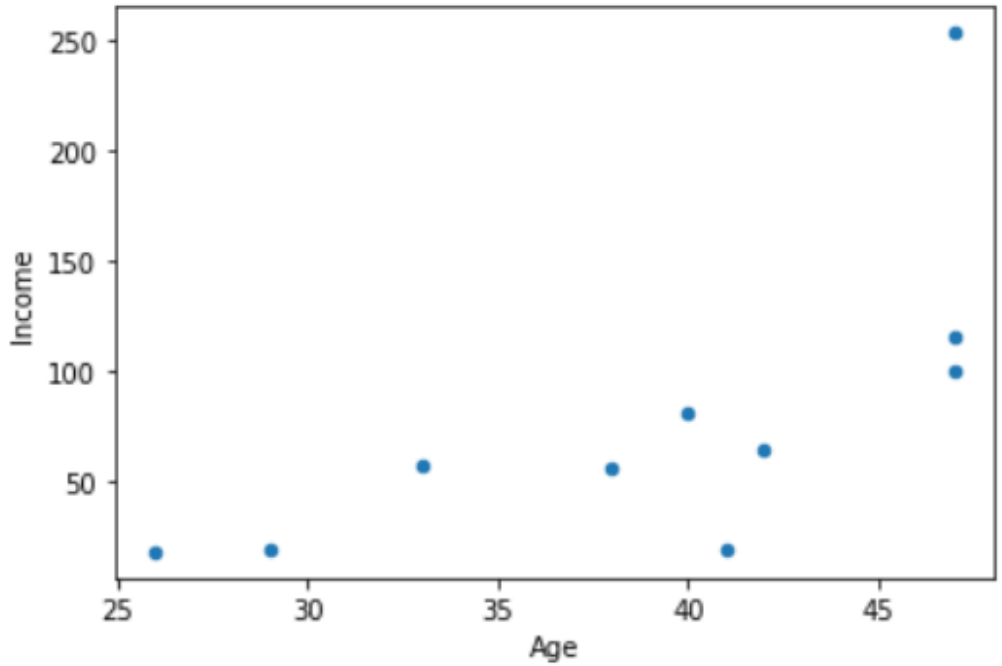
Ilustrasi Cara Kerja Algoritma K-Means

Contoh Kasus: Klasterisasi Pelanggan

Data Pelanggan

CustID	Age	Income
1	41	19
2	47	100
3	33	57
4	29	19
5	47	253
6	40	81
7	38	56
8	42	64
9	26	18
10	47	115

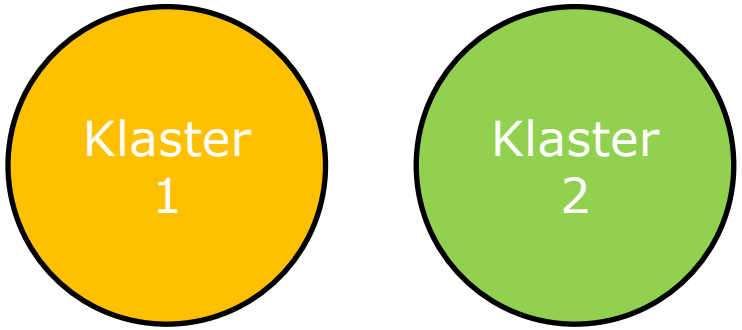
Diketahui data pelanggan sebagaimana tabel di samping, kita diminta mengelompokkan data pelanggan menjadi 2 (dua) kelompok.



Contoh Kasus: Klasterisasi Pelanggan

1. Tentukan jumlah klaster. Dalam contoh kasus ini kita gunakan nilai k=2

CustID	Age	Income
1	41	19
2	47	100
3	33	57
4	29	19
5	47	253
6	40	81
7	38	56
8	42	64
9	26	18
10	47	115



Contoh Kasus: Klasterisasi Pelanggan

2. Inisialisasi nilai centroid awal setiap klaster secara acak

CustID	Age	Income
1	41	19
2	47	100
3	33	57
4	29	19
5	47	253
6	40	81
7	38	56
8	42	64
9	26	18
10	47	115

Cara penentuan centroid awal:

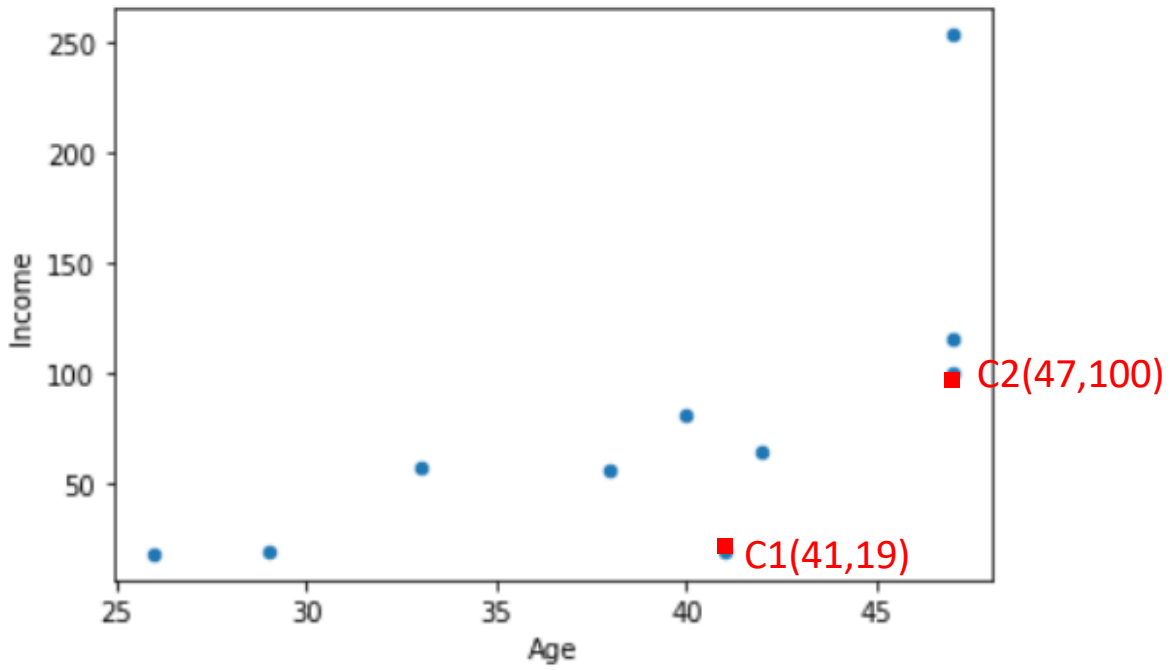
1. Memilih salah satu data untuk atribut "Age" dan "Income" secara acak
2. Membangkitkan bilangan acak sesuai rentang nilai "Age" dan "Income"

Dalam contoh ini kita memilih centroid awal dengan cara 1, kita tentukan C1 = (41,19) dan C2 = (47,100)

Contoh Kasus: Klasterisasi Pelanggan

2. Inisialisasi nilai centroid awal setiap klaster secara acak

CustID	Age	Income
1	41	19
2	47	100
3	33	57
4	29	19
5	47	253
6	40	81
7	38	56
8	42	64
9	26	18
10	47	115



Contoh Kasus: Klasterisasi Pelanggan

3. Hitung jarak setiap titik data dengan setiap centroid. Contoh: Euclidean Distance

CustID	Age	Income	Jarak ke C1(41,19)	Jarak ke C2(47,100)
1	41	19	$\sqrt{(41-41)^2+(19-19)^2}=0$	$\sqrt{(41-47)^2+(19-100)^2}=81{,}22$
2	47	100	$\sqrt{(47-41)^2+(100-19)^2}=81{,}22$	$\sqrt{(47-47)^2+(100-100)^2}=0$
3	33	57	$\sqrt{(33-41)^2+(57-19)^2}=38{,}83$	$\sqrt{(33-47)^2+(57-100)^2}=45{,}22$
4	29	19	$\sqrt{(29-41)^2+(19-19)^2}=12{,}0$	$\sqrt{(29-47)^2+(19-100)^2}=82{,}98$
5	47	253	$\sqrt{(47-41)^2+(253-19)^2}=234{,}08$	$\sqrt{(47-47)^2+(253-100)^2}=153{,}0$
6	40	81	$\sqrt{(40-41)^2+(81-19)^2}=62{,}01$	$\sqrt{(40-47)^2+(81-100)^2}=20{,}25$
7	38	56	$\sqrt{(38-41)^2+(56-19)^2}=37{,}12$	$\sqrt{(38-47)^2+(56-100)^2}=44{,}91$
8	42	64	$\sqrt{(42-41)^2+(64-19)^2}=45{,}01$	$\sqrt{(42-47)^2+(64-100)^2}=36{,}35$
9	26	18	$\sqrt{(26-41)^2+(18-19)^2}=15{,}03$	$\sqrt{(26-47)^2+(18-100)^2}=84{,}65$
10	47	115	$\sqrt{(47-41)^2+(115-19)^2}=96{,}19$	$\sqrt{(47-47)^2+(115-100)^2}=15{,}0$

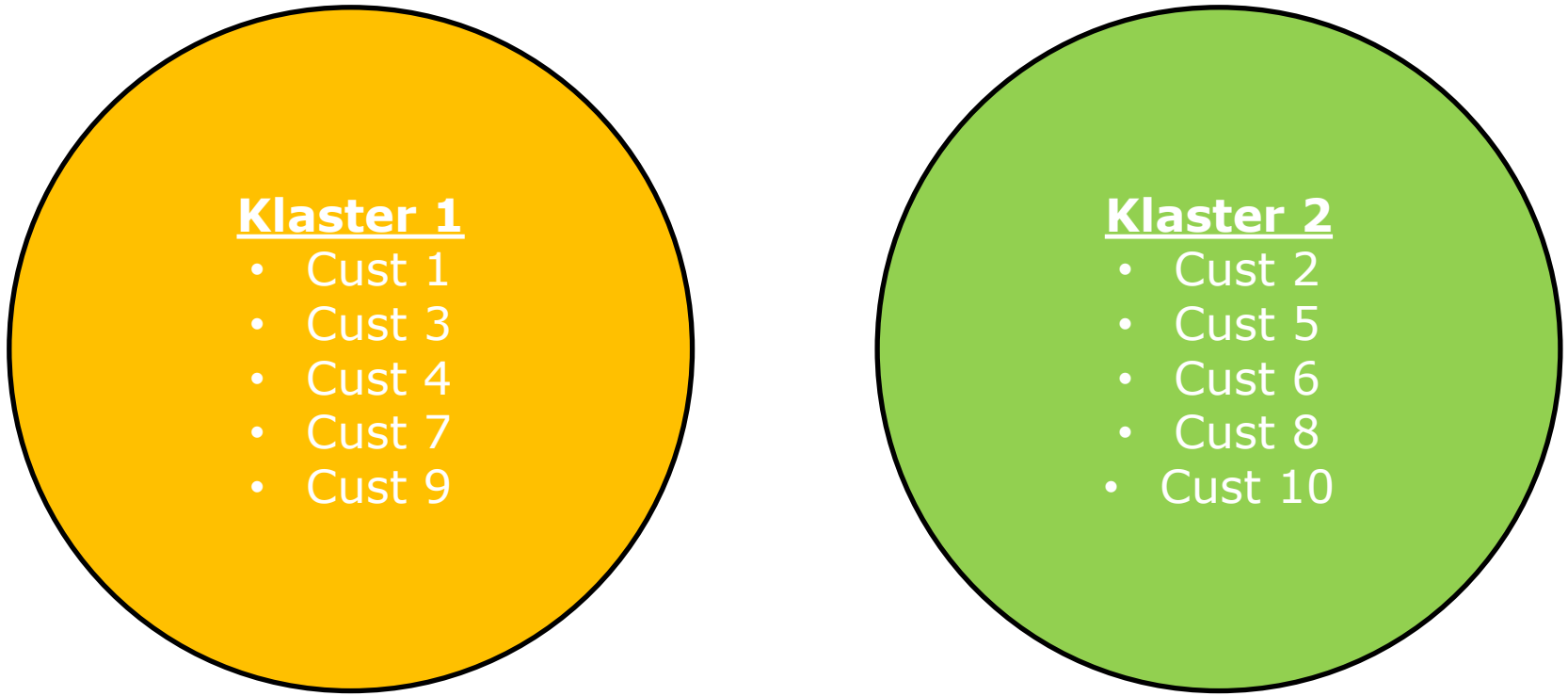
Contoh Kasus: Klasterisasi Pelanggan

4. Masukkan setiap titik data ke dalam klaster berdasarkan jarak terdekat dengan centroid

CustID	Age	Income	Jarak ke C1(41,19)	Jarak ke C2(47,100)	Klaster
1	41	19	$\sqrt{(41-41)^2+(19-19)^2}=0$	$\sqrt{(41-47)^2+(19-100)^2}=81{,}22$	1
2	47	100	$\sqrt{(47-41)^2+(100-19)^2}=81{,}22$	$\sqrt{(47-47)^2+(100-100)^2}=0$	2
3	33	57	$\sqrt{(33-41)^2+(57-19)^2}=38{,}83$	$\sqrt{(33-47)^2+(57-100)^2}=45{,}22$	1
4	29	19	$\sqrt{(29-41)^2+(19-19)^2}=12{,}0$	$\sqrt{(29-47)^2+(19-100)^2}=82{,}98$	1
5	47	253	$\sqrt{(47-41)^2+(253-19)^2}=234{,}08$	$\sqrt{(47-47)^2+(253-100)^2}=153{,}0$	2
6	40	81	$\sqrt{(40-41)^2+(81-19)^2}=62{,}01$	$\sqrt{(40-47)^2+(81-100)^2}=20{,}25$	2
7	38	56	$\sqrt{(38-41)^2+(56-19)^2}=37{,}12$	$\sqrt{(38-47)^2+(56-100)^2}=44{,}91$	1
8	42	64	$\sqrt{(42-41)^2+(64-19)^2}=45{,}01$	$\sqrt{(42-47)^2+(64-100)^2}=36{,}35$	2
9	26	18	$\sqrt{(26-41)^2+(18-19)^2}=15{,}03$	$\sqrt{(26-47)^2+(18-100)^2}=84{,}65$	1
10	47	115	$\sqrt{(47-41)^2+(115-19)^2}=96{,}19$	$\sqrt{(47-47)^2+(115-100)^2}=15{,}0$	2

Contoh Kasus: Klasterisasi Pelanggan

4. Masukkan setiap titik data ke dalam klaster berdasarkan jarak terdekat dengan centroid



Contoh Kasus: Klasterisasi Pelanggan

5. Untuk setiap klaster, hitung nilai centroid baru berdasarkan rerata (means) dari setiap data di dalam klaster

CustID	Age	Income	Jarak ke C1(41,19)	Jarak ke C2(47,100)	Klaster
1	41	19	$\sqrt{(41-41)^2+(19-19)^2}=0$	$\sqrt{(41-47)^2+(19-100)^2}=81{,}22$	1
2	47	100	$\sqrt{(47-41)^2+(100-19)^2}=81{,}22$	$\sqrt{(47-47)^2+(100-100)^2}=0$	2
3	33	57	$\sqrt{(33-41)^2+(57-19)^2}=38{,}83$	$\sqrt{(33-47)^2+(57-100)^2}=45{,}22$	1
4	29	19	$\sqrt{(29-41)^2+(19-19)^2}=12{,}0$	$\sqrt{(29-47)^2+(19-100)^2}=82{,}98$	1
5	47	253	$\sqrt{(47-41)^2+(253-19)^2}=234{,}08$	$\sqrt{(47-47)^2+(253-100)^2}=153{,}0$	2
6	40	81	$\sqrt{(40-41)^2+(81-19)^2}=62{,}01$	$\sqrt{(40-47)^2+(81-100)^2}=20{,}25$	2
7	38	56	$\sqrt{(38-41)^2+(56-19)^2}=37{,}12$	$\sqrt{(38-47)^2+(56-100)^2}=44{,}91$	1
8	42	64	$\sqrt{(42-41)^2+(64-19)^2}=45{,}01$	$\sqrt{(42-47)^2+(64-100)^2}=36{,}35$	2
9	26	18	$\sqrt{(26-41)^2+(18-19)^2}=15{,}03$	$\sqrt{(26-47)^2+(18-100)^2}=84{,}65$	1
10	47	115	$\sqrt{(47-41)^2+(115-19)^2}=96{,}19$	$\sqrt{(47-47)^2+(115-100)^2}=15{,}0$	2

Centroid Baru **C1** = (mean(41;33;29;38;26), mean(19;57;19;56;18)) = **(33,4; 33,8)**

Contoh Kasus: Klasterisasi Pelanggan

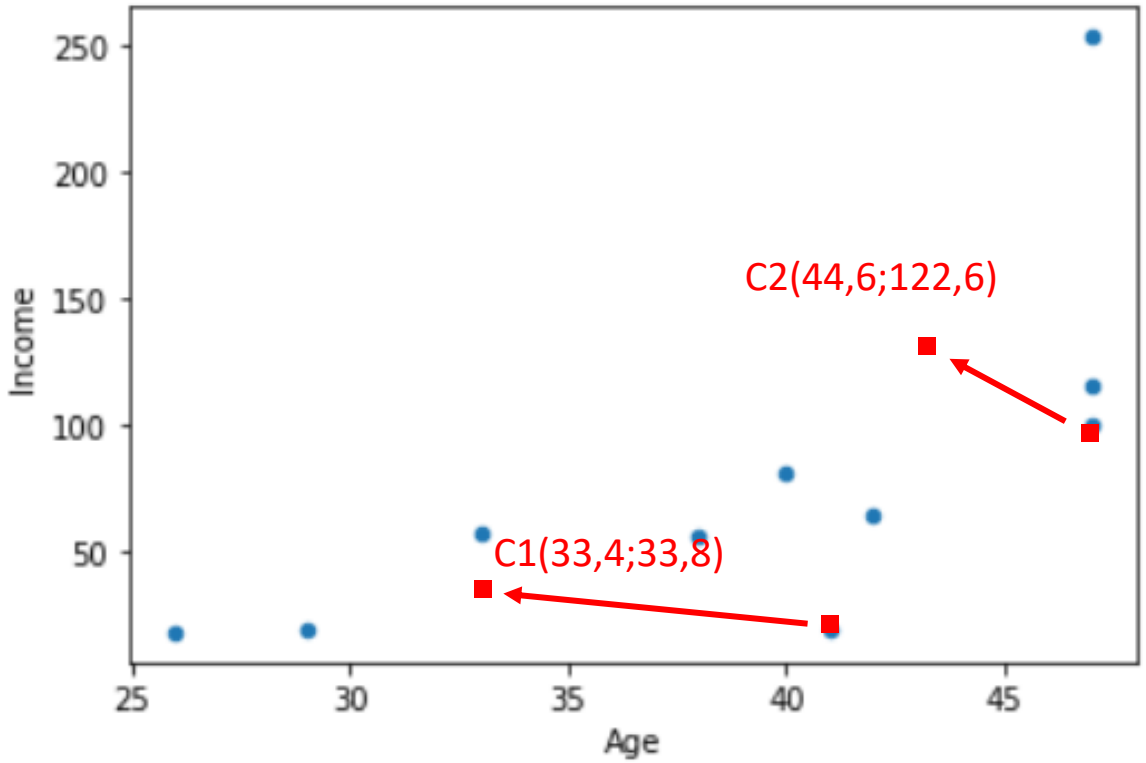
5. Untuk setiap klaster, hitung nilai centroid baru berdasarkan rerata (means) dari setiap data di dalam klaster

CustID	Age	Income	Jarak ke C1(41,19)	Jarak ke C2(47,100)	Klaster
1	41	19	$\sqrt{(41-41)^2+(19-19)^2}=0$	$\sqrt{(41-47)^2+(19-100)^2}=81{,}22$	1
2	47	100	$\sqrt{(47-41)^2+(100-19)^2}=81{,}22$	$\sqrt{(47-47)^2+(100-100)^2}=0$	2
3	33	57	$\sqrt{(33-41)^2+(57-19)^2}=38{,}83$	$\sqrt{(33-47)^2+(57-100)^2}=45{,}22$	1
4	29	19	$\sqrt{(29-41)^2+(19-19)^2}=12{,}0$	$\sqrt{(29-47)^2+(19-100)^2}=82{,}98$	1
5	47	253	$\sqrt{(47-41)^2+(253-19)^2}=234{,}08$	$\sqrt{(47-47)^2+(253-100)^2}=153{,}0$	2
6	40	81	$\sqrt{(40-41)^2+(81-19)^2}=62{,}01$	$\sqrt{(40-47)^2+(81-100)^2}=20{,}25$	2
7	38	56	$\sqrt{(38-41)^2+(56-19)^2}=37{,}12$	$\sqrt{(38-47)^2+(56-100)^2}=44{,}91$	1
8	42	64	$\sqrt{(42-41)^2+(64-19)^2}=45{,}01$	$\sqrt{(42-47)^2+(64-100)^2}=36{,}35$	2
9	26	18	$\sqrt{(26-41)^2+(18-19)^2}=15{,}03$	$\sqrt{(26-47)^2+(18-100)^2}=84{,}65$	1
10	47	115	$\sqrt{(47-41)^2+(115-19)^2}=96{,}19$	$\sqrt{(47-47)^2+(115-100)^2}=15{,}0$	2

Centroid Baru **C2** = (mean(47;47;40;42;47), mean(100;253;81;64;115)) = **(44,6; 122,6)**

Contoh Kasus: Klasterisasi Pelanggan

Pergeseran Centroid setiap klaster. C1 =(33,4; 33,8) dan C2 = (44,6; 122,6)



Contoh Kasus: Klasterisasi Pelanggan

6. Ulangi langkah 3-5 menggunakan centroid baru

CustID	Age	Income	Jarak ke C1(33,4; 33,8)	Jarak ke C2(44,6; 122,6)	Klaster
1	41	19	16,64	103,66	1
2	47	100	67,58	22,73	2
3	33	57	23,20	66,62	1
4	29	19	15,44	104,77	1
5	47	253	219,62	130,42	2
6	40	81	47,66	41,85	2
7	38	56	22,67	66,93	1
8	42	64	31,40	58,66	1
9	26	18	17,45	106,24	1
10	47	115	82,33	7,97	2

Contoh Kasus: Klasterisasi Pelanggan

6. Ulangi langkah 3-5 menggunakan centroid baru

CustID	Age	Income	Jarak ke C1(33,4; 33,8)	Jarak ke C2(44,6; 122,6)	Klaster
1	41	19	16,64	103,66	1
2	47	100	67,58	22,73	2
3	33	57	23,20	66,62	1
4	29	19	15,44	104,77	1
5	47	253	219,62	130,42	2
6	40	81	47,66	41,85	2
7	38	56	22,67	66,93	1
8	42	64	31,40	58,66	1
9	26	18	17,45	106,24	1
10	47	115	82,33	7,97	2

Apakah hasil klasterisasinya sama dengan tahap sebelumnya?

- Jika sama, hentikan proses klasterisasi
- Jika belum sama, ulangi langkah 3-5

Contoh Kasus: Klasterisasi Pelanggan

Data			ITERASI 1	
CustID	**Age**	**Income**	C1(41,19)	C2(47,100)
1	41	19	0,00	81,22
2	47	100	81,22	0,00
3	33	57	38,83	45,22
4	29	19	12,00	82,98
5	47	253	234,08	153,00
6	40	81	62,01	20,25
7	38	56	37,12	44,91
8	42	64	45,01	36,35
9	26	18	15,03	84,65
10	47	115	96,19	15,00
	Centroid Baru		33,4	44,6
			33,8	122,6

ITERASI 2	
C1	C2
16,64	103,66
67,58	22,73
23,20	66,62
15,44	104,77
219,62	130,42
47,66	41,85
22,67	66,93
31,40	58,66
17,45	106,24
82,33	7,97
34,83	45,25
38,83	137,25

ITERASI 3	
C1	C2
20,77	118,33
62,36	37,29
18,26	81,18
20,67	119,36
214,51	115,76
42,48	56,49
17,46	81,57
26,17	73,32
22,63	120,79
77,13	22,32
35,57	47,00
44,86	156

ITERASI 4	
C1	C2
26,42	137,13
56,31	56,00
12,41	99,98
26,68	138,18
208,46	97,00
36,41	75,33
11,40	100,40
20,19	92,14
28,51	139,59
71,07	41,00
SELESAI	

Optimasi Nilai k pada K-Means

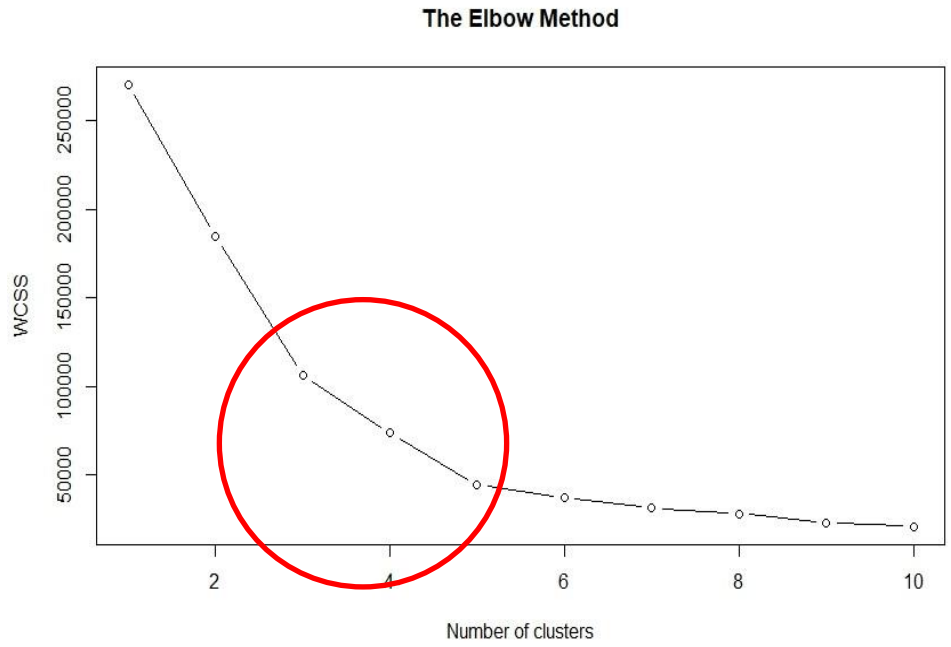
- Salah satu faktor krusial baik tidaknya metode K-Means adalah **jumlah klusternya (nilai K)**. Hasil pengelompokan akan menghasilkan analisa yang berbeda untuk jumlah klaster yang berbeda juga.
- **Semakin kecil nilai K** (misal 2), maka pembagian kluster menjadi cepat, namun mungkin ada informasi tersembunyi yang tidak terungkap.
- **Semakin besar nilai K** (misal K=10), maka terlalu banyak kluster. Mungkin akan terlalu sulit untuk membuat analisa atau memilih dukungan keputusan dari hasil cluster.

Optimasi Nilai k pada K-Means

- Penentuan nilai k terbaik dapat dilakukan berdasarkan ukuran kualitas hasil klasterisasi.
- Beberapa ukuran kualitas klaster:
  - Sum Square Error (SSE)
  - Davies-Bouldin Index (DBI)
  - Silhouette Coefficient
  - Rand Index
  - Mutual Information
  - Calinski-Harabasz Index (C-H Index)
  - Dunn Index

Penentuan Nilai k Terbaik dengan Metode Elbow

- Untuk mengetahui jumlah klaster yang paling baik adalah dengan cara melihat perbandingan kualitas klaster untuk setiap pilihan nilai k (missal k=2, 3, 4, 5, …).
- Nilai k yang dipilih adalah nilai k yang memiliki perubahan kualitas signifikan, seperti sebuah **siku (elbow)**.



Kesimpulan

- Clustering merupakan salah satu metode pembelajaran tidak tersupervisi
- Metode clustering dibagi menjadi 3 jenis: partitioning-based, hierarchical, dan density-based.
- Algoritma K-means adalah salah satu algoritma clustering yang bersifat iteratif yang mempartisi dataset menjadi subkelompok non-overlapping berbeda yang ditentukan oleh K (cluster)
- Algoritma K-Means:
  - Relatif efisien untuk data kecil hingga besar
  - Menghasilkan kelompok klaster
  - Memerlukan inisialisasi nilai k

Latihan

- Diketahui dataset
- Lakukan clustering dengn jumlah cluster = 2
- Dua cluster merepresentasikan :
  - Penyempitan pembuluh darah > 50%
  - Penyempitan kurang dari 50 %

	diameter narrowing	age	cholesterol	max HR	ST by exercise
1	0	63	233	150	2.3
2	1	67	286	108	1.5
3	1	67	229	129	2.6
4	0	37	250	187	3.5
5	0	41	204	172	1.4
6	0	56	236	178	0.8
7	1	62	268	160	3.6
8	0	57	354	163	0.6
9	1	63	254	147	1.4
10	1	53	203	155	3.1
11	0	57	192	148	0.4
12	0	56	294	153	1.3
13	1	56	256	142	0.6
14	0	44	263	173	0.0
15	0	52	199	162	0.5
16	0	57	168	174	1.6
17	1	48	229	168	1.0
18	0	54	239	160	1.2
19	0	48	275	139	0.2
20	0	49	266	171	0.6
21	0	64	211	144	1.8
22	0	58	283	162	1.0
23	1	58	284	160	1.8
24	1	58	224	173	3.2
25	1	60	206	132	2.4
26	0	50	219	158	1.6
27	0	58	340	172	0.0
28	0	66	226	114	2.6
29	0	43	247	171	1.5
30	1	40	167	114	2.0