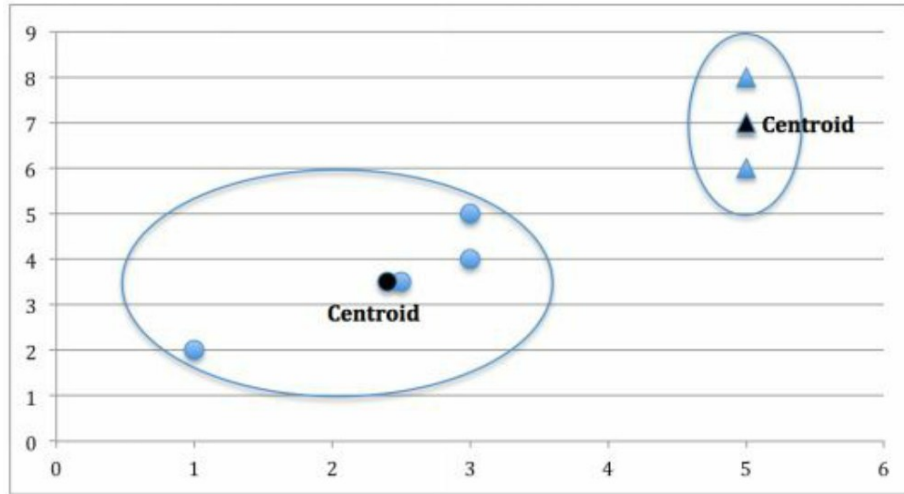


UNSUPERVISED LEARNING - *CLUSTERING*

Team Teaching

METODE UNSUPERVISED LEARNING



K-Means, Market Basket Analysis, PCA, dan sebagainya

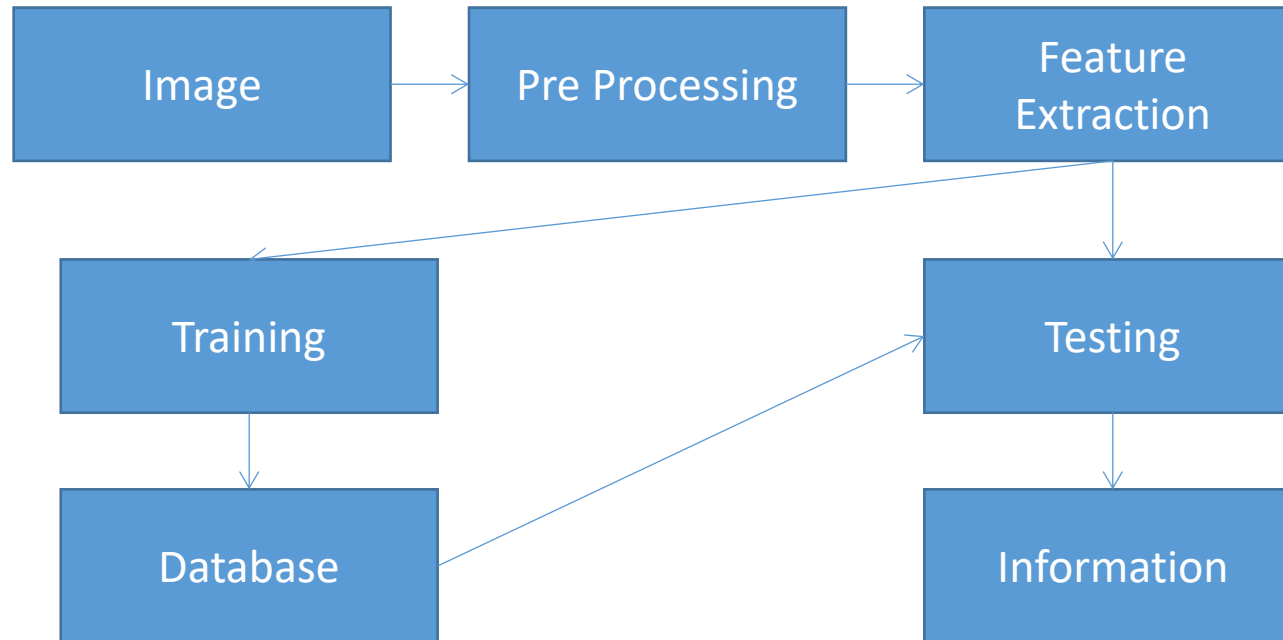
PENDAHULUAN

- Clustering merupakan salah satu teknik data mining yang digunakan untuk mendapatkan kelompok-kelompok dari objek-objek yang mempunyai karakteristik yang umum di data yang cukup besar.
- Clustering mengacu pada pengelompokan catatan, pengamatan, atau kasus ke dalam kelas objek serupa.

PENDAHULUAN

- Clustering membagi data ke dalam grup-grup, sedemikian sehingga masing-masing grup mempunyai persamaan karakteristik yang esensial.
- Clustering memegang peranan penting dalam aplikasi data mining, misalnya eksplorasi data ilmu pengetahuan, pengaksesan informasi dan text mining, aplikasi basis data spasial, dan analisis web.
- Clustering diterapkan dalam *search engine* di Internet. *Search engine* akan mencari dokumen yang cocok dengan kata kunci yang dimasukkan.

TRAINING & TESTING



TUJUAN CLUSTERING

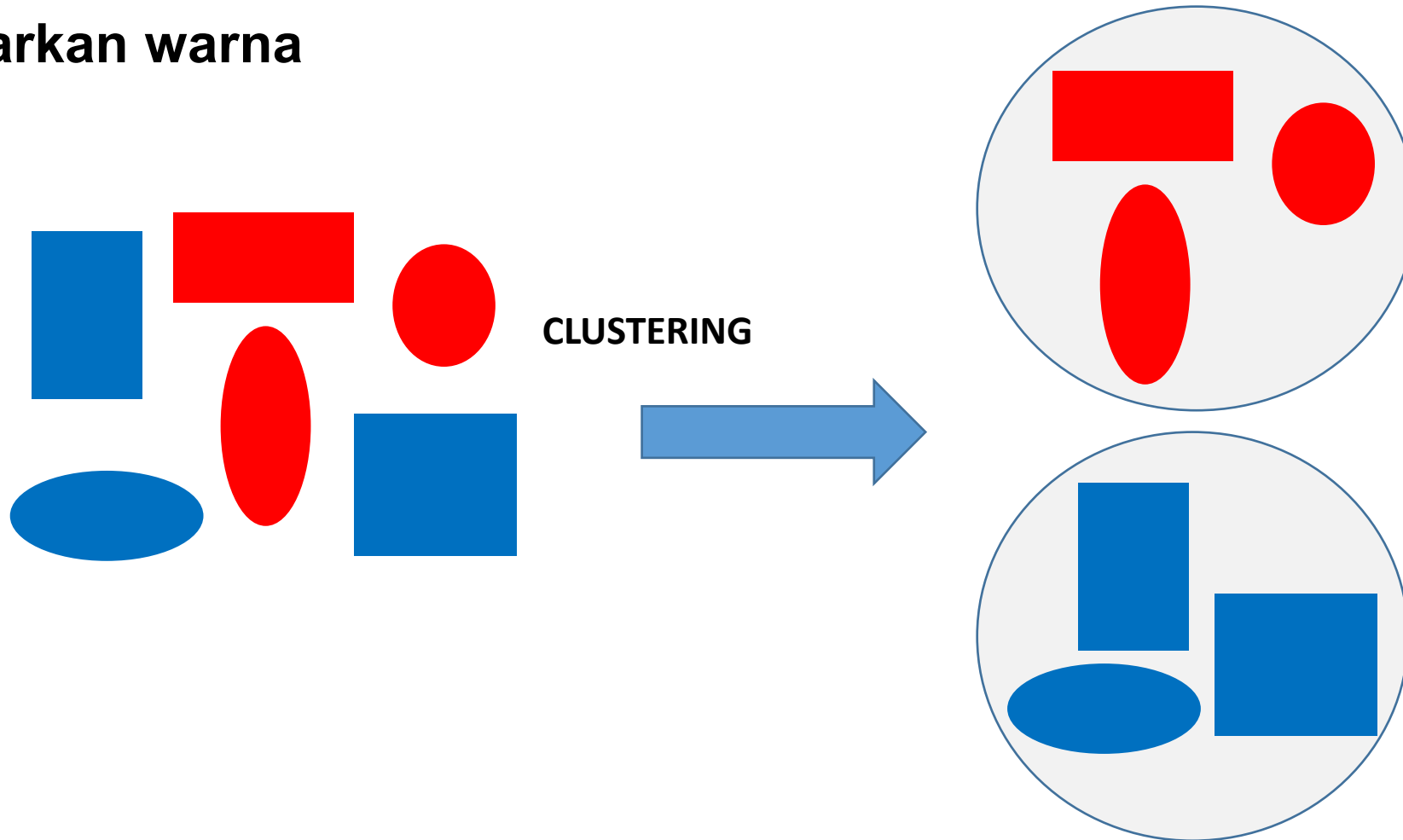
- Pengelompokan sejumlah data atau objek ke dalam cluster atau grup sehingga dalam setiap cluster akan berisi data yang semirip mungkin.
- Clustering melakukan pengelompokkan data yang didasarkan pada kesamaan antar objek, oleh karena itu klasterisasi digolongkan sebagai metode unsupervised learning.

TUGAS CLUSTERING

- Membagi seluruh kumpulan data menjadi subkelompok atau kelompok yang relatif homogen, di mana kesamaan catatan di dalam kelompok dimaksimalkan, dan kesamaan dengan catatan di luar kelompok ini diminimalkan.

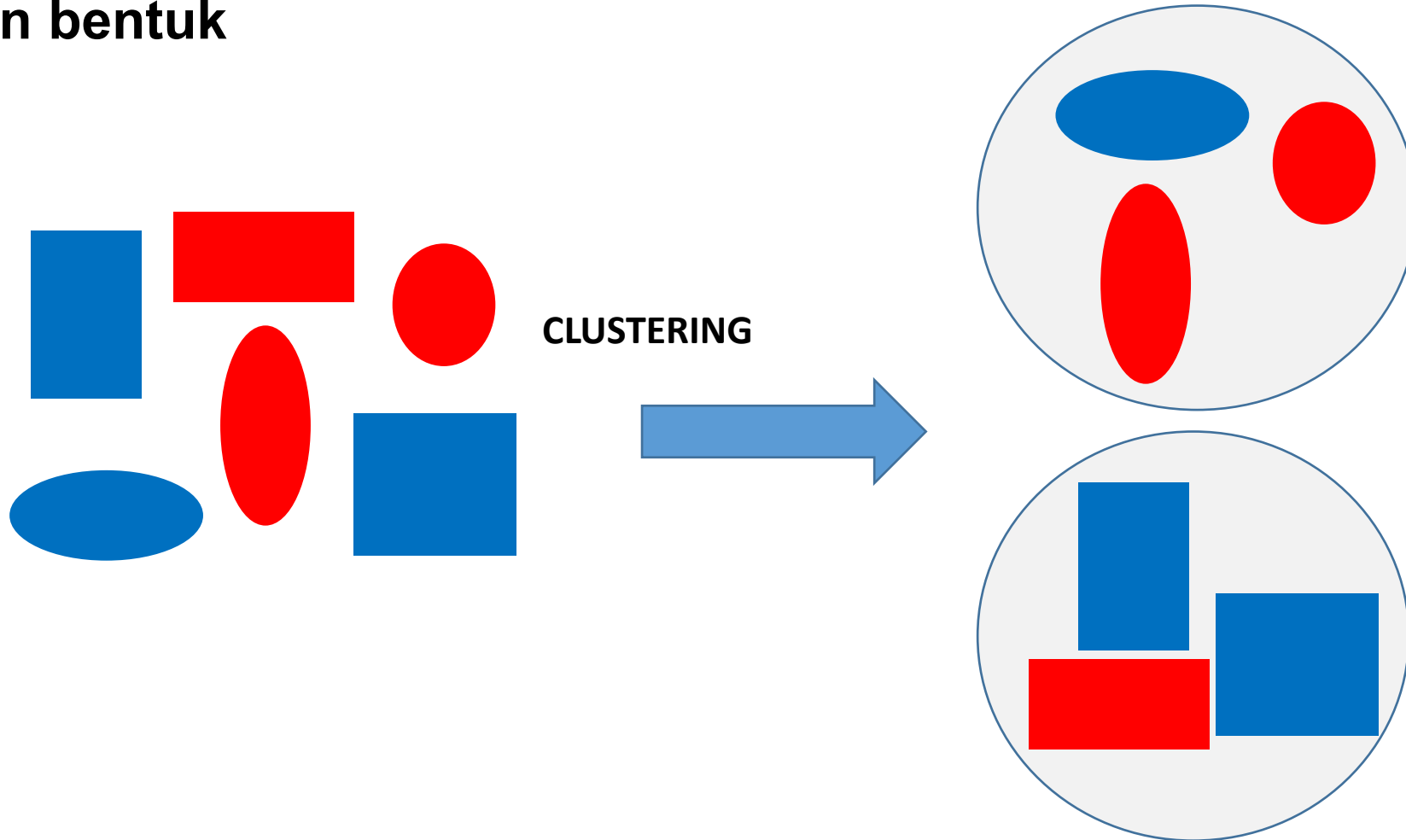
CONTOH CLUSTERING

Berdasarkan warna



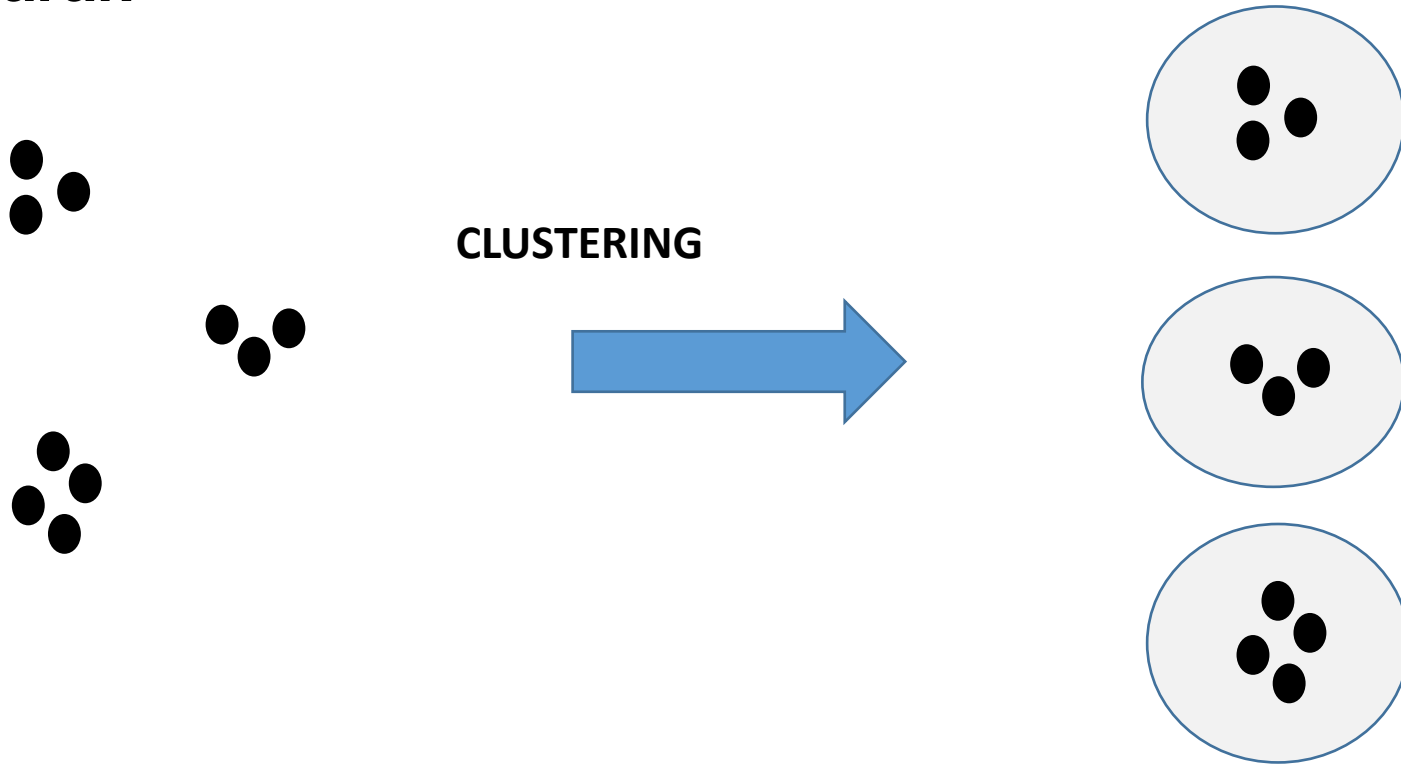
CONTOH CLUSTERING

Berdasarkan bentuk



CONTOH CLUSTERING

Berdasarkan jarak



KLASIFIKASI Vs CLUSTERING

	CLASSIFICATION	CLUSTERING
DATA	Supervised	Unsupervised
LABEL	Ya	Tidak
ANALISIS HASIL	Error ratio	Variance

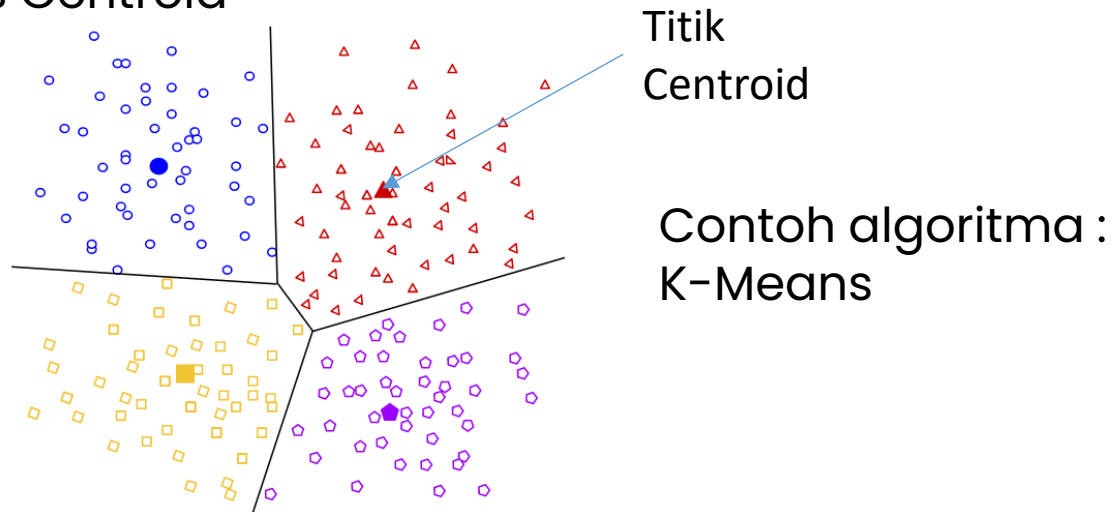
JENIS CLUSTERING

1. Clustering Non-Hierarkis

Merupakan jenis clustering yang menggabungkan atau memisahkan cluster dengan tidak berbentuk struktur seperti pohon. Teknik ini mengelompokkan data untuk memaksimalkan atau meminimalkan beberapa kriteria evaluasi antara lain :

a. Clustering Berbasis Centroid

Artinya setiap cluster diasosiasikan dengan centroid. Tujuan utama dari algoritma ini adalah untuk meminimalkan jumlah jarak antara titik data dan cluster yang sesuai.

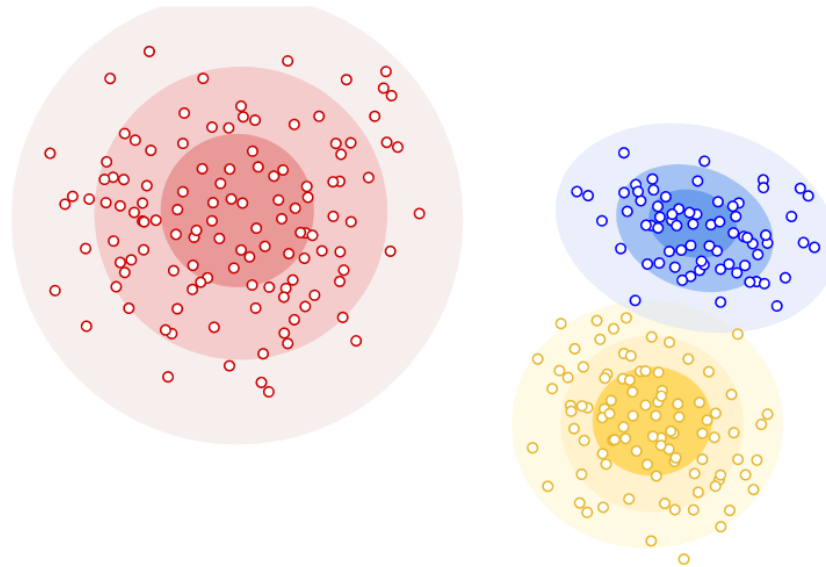


JENIS CLUSTERING

1. Clustering Non-Hierarkis

b. Clustering Berbasis Distribusi

Model pengelompokan ini didasarkan pada gagasan tentang seberapa besar kemungkinan semua titik data dalam kluster memiliki distribusi yang sama



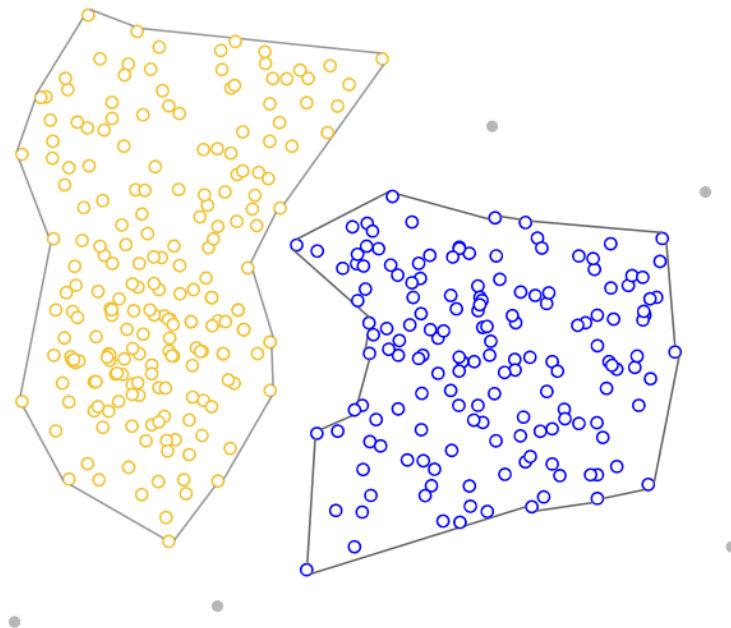
Contoh algoritma :
Gaussian Mixture Model

JENIS CLUSTERING

1. Clustering Non-Hierarkis

c. Clustering Berbasis Density

Model-model ini mencari ruang data untuk area dengan kepadatan titik data yang bervariasi dalam ruang data. Mereka mengisolasi wilayah padat yang berbeda dan menetapkan titik data dalam wilayah ini ke cluster yang



Contoh algoritma :
DBSCAN

JENIS CLUSTERING

2. Clustering Hierarkis

Merupakan pembuatan kluster dalam urutan yang telah ditentukan. Kluster diurutkan dari atas ke bawah. Dalam jenis pengelompokan ini, kelompok yang serupa dikelompokkan bersama dan disusun secara hierarkis. Dalam pengelompokan ini, kami menghubungkan pasangan cluster semua objek data yang ada dalam hierarki.

$$d_{\text{Euclidean}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

$$d_{\text{cityblock}}(\mathbf{x}, \mathbf{y}) = \sum_i |x_i - y_i|$$

$$d_{\text{Minkowski}}(\mathbf{x}, \mathbf{y}) = \sum_i |x_i - y_i|^q$$

KARAKTERISTIK CLUSTERING

- Partitioning Clustering
- Hierarchical Clustering
- Overlapping Clustering
- Hybrid

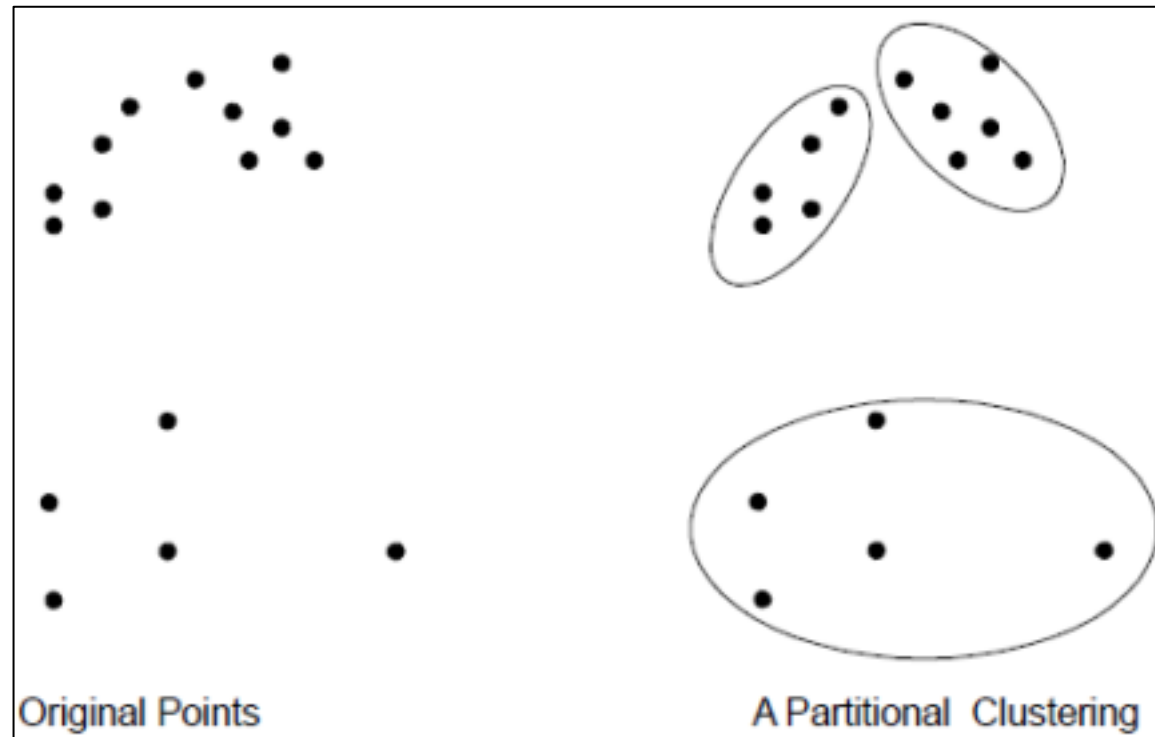
PARTITIONAL CLUSTERING

- Pendekatan partisi yaitu data dikelompokkan ke dalam sejumlah cluster tanpa adanya struktur hirarki antara satu dengan yang lainnya.
- Pada metode ini, setiap cluster memiliki titik pusat cluster (centroid) dan secara umum metode ini memiliki fungsi tujuan yaitu meminimumkan jarak (dissimilarity) dari seluruh data ke pusat cluster masing-masing.

PARTITIONING CLUSTERING

- Disebut juga exclusive clustering.
- Setiap data harus masuk dalam cluster tertentu.
- Memungkinkan bagi setiap data yang masuk ke dalam cluster tertentu pada suatu tahapan, untuk kemudian pindah ke cluster lain pada tahapan berikutnya.
- Contoh: k-means.

PARTITIONAL CLUSTERING



HIERARCHICAL CLUSTERING

- Pendekatan hirarki adalah suatu metode pengelompokan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kesamaan paling dekat.
- Kemudian proses diteruskan ke objek lain yang memiliki kedekatan kedua.
- Demikian seterusnya sehingga cluster akan membentuk semacam pohon dimana ada hierarki (tingkatan) yang jelas antar objek, dari yang paling mirip sampai yang paling tidak mirip.

HIERARCHICAL CLUSTERING

- Setiap data harus masuk dalam cluster tertentu.
- Data yang sudah masuk ke dalam cluster tertentu pada suatu tahapan, tidak dapat berpindah ke cluster lain.
- Contoh: Single Linkage, Centroid Linkage, Complete Linkage, Average Centroid.
- Direction of Hierarchy:
 - Divisive
 - 1 cluster divided/dibagi into k cluster
 - Agglomerative
 - n cluster merge to k cluster ($n > k$)

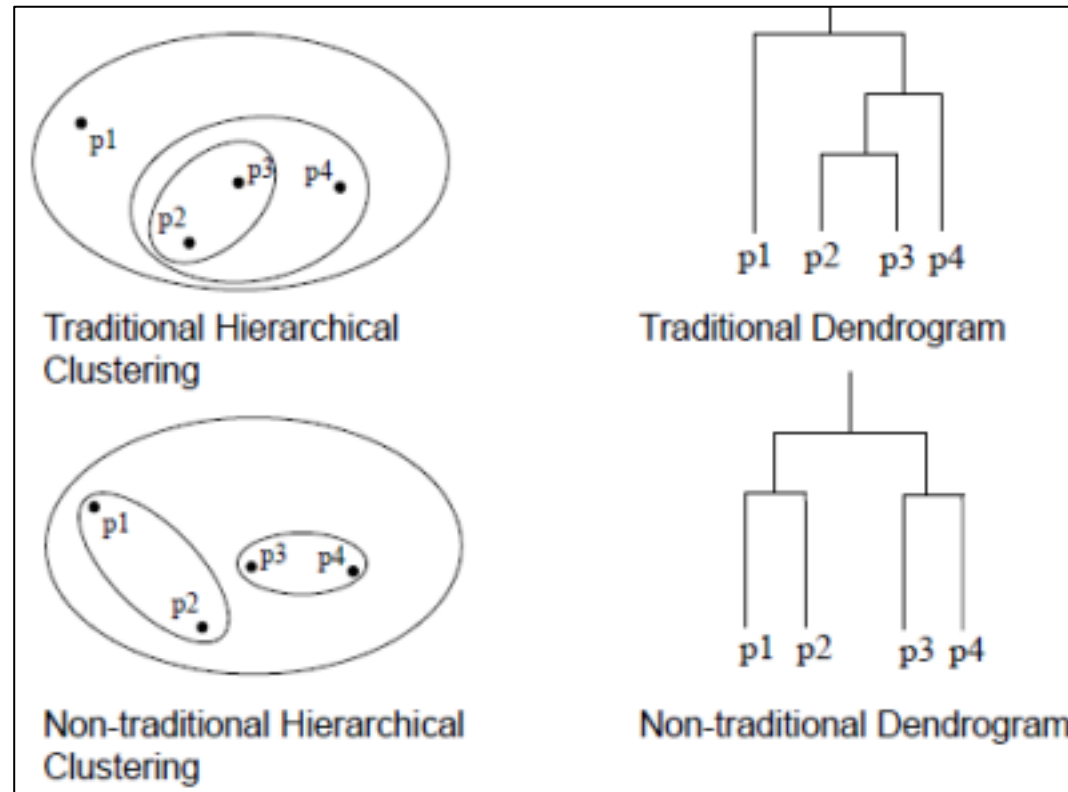
HIERARCHICAL CLUSTERING

- Divisive (Pembagian)
 - Mulai dari 1 cluster besar
 - Pada tiap tahap, pisahkan (split) sebuah cluster hingga tiap cluster memiliki 1 titik (atau sampai terdapat k cluster)
 - Algoritma hierarchical clustering tradisional menggunakan matriks jarak atau similaritas

HIERARCHICAL CLUSTERING

- Agglomerative (Penggabungan)
 - Mulai dari titik-titik sebagai cluster individu
 - Pada tiap tahap, gabungkan (merge) pasangan cluster terdekat sampai terbentuk 1 (atau k) cluster

HIERARCHICAL CLUSTERING



OVERLAPPING & HYBRID CLUSTERING

- Overlapping Clustering:
 - Setiap data mungkin masuk ke dalam beberapa cluster.
 - Data memiliki nilai keanggotaan pada beberapa cluster.
 - Contoh: Fuzzy C-Means
- Hybrid:
 - Mengawinkan karakteristik dari Partitioning, Hierarchical, dan Overlapping.

Algoritma Clustering

- K-means
- Meanshift Algorithm
- Agglomerative Hierarchical Clustering

K-MEANS

- Salah satu metode data clustering non hirarki yang berusaha mempartisi data yang ada ke dalam satu atau lebih cluster/kelompok.
- Metode ini mempartisi data ke dalam cluster/kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain.

K-MEANS

- Merupakan algoritma clustering yang sangat populer.
- Pada awalnya ditentukan berapa cluster yang akan dibentuk.
- Sembarang objek dalam cluster dapat dipilih untuk dijadikan sebagai titik tengah (centroid point) cluster.
- Kemudian lakukan langkah berikut hingga tidak ada objek yang dapat berpindah:
 1. Tentukan koordinat titik tengah setiap cluster,
 2. Tentukan jarak setiap objek terhadap koordinat titik tengah,
 3. Kelompokkan objek-objek tersebut berdasarkan jarak minimumnya.

K-MEANS

- Langkah-langkah metode k-means
 - Pilih secara acak k buah data sebagai pusat cluster.
 - Jarak antara data dan pusat cluster dihitung menggunakan euclidian distance.

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- Ket:
 - $d(x, y)$: jarak antara data pada titik x dan y
 - x : titik data pertama (pusat klaster)
 - y : titik data kedua (data dari N)
 - n : jumlah atribut data

K-MEANS

1. Tentukan jumlah cluster yang ingin dibentuk.
2. Tentukan centroid awal untuk masing-masing cluster secara random.
3. Hitung jarak (euclidian distance) setiap data ke masing-masing centroid cluster.
4. Setiap data memilih centroid yang terdekat.
5. Tentukan centroid baru dari masing-masing cluster, dengan cara menghitung nilai rata-rata dari data-data dalam cluster.
6. Jika nilai centroid baru tidak sama dengan centroid lama (pada cluster yang sama), maka ulang ke langkah 3.

K-MEANS

- Link : <https://bit.ly/ClusteringML>
- Cara Kerja K-Means



KELEMAHAN K-MEANS

1. Sangat bergantung pada pemilihan nilai awal centroid.
2. Tidak jelas berapa banyak cluster yang terbaik.
3. Hanya bekerja pada atribut numerik (data kuantitatif).

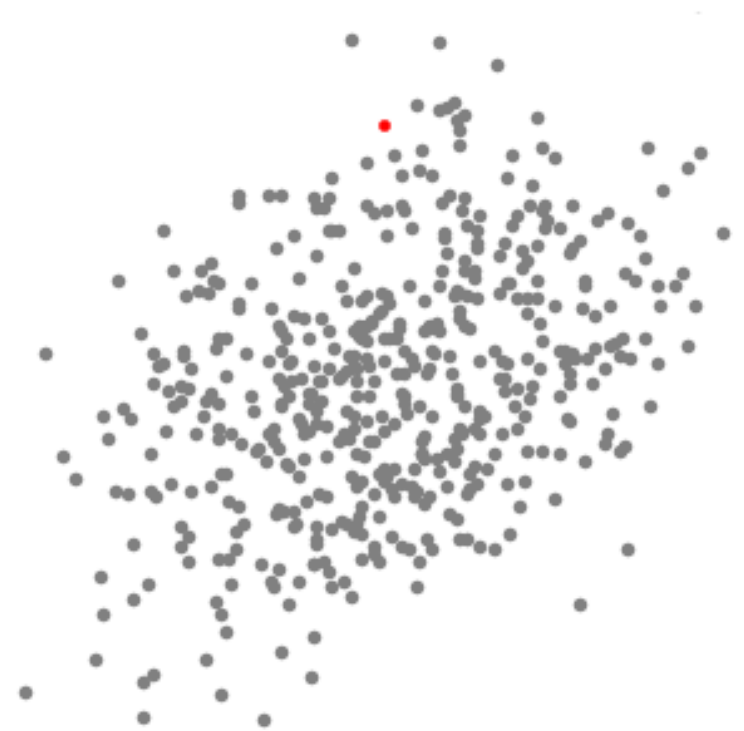
MEANSHIFT

- Algoritma berbasis centroid yang bertujuan untuk menemukan titik pusat dari setiap kelompok/kelas, yang bekerja dengan memperbarui kandidat centroid menjadi rata-rata titik di dalam jendela geser.
- Kandidat jendela ini kemudian disaring dalam tahap pasca-pemrosesan untuk menghilangkan duplikat yang hampir mendekati, sehingga membentuk titik pusat dan grup yang sesuai.

MEANSHIFT

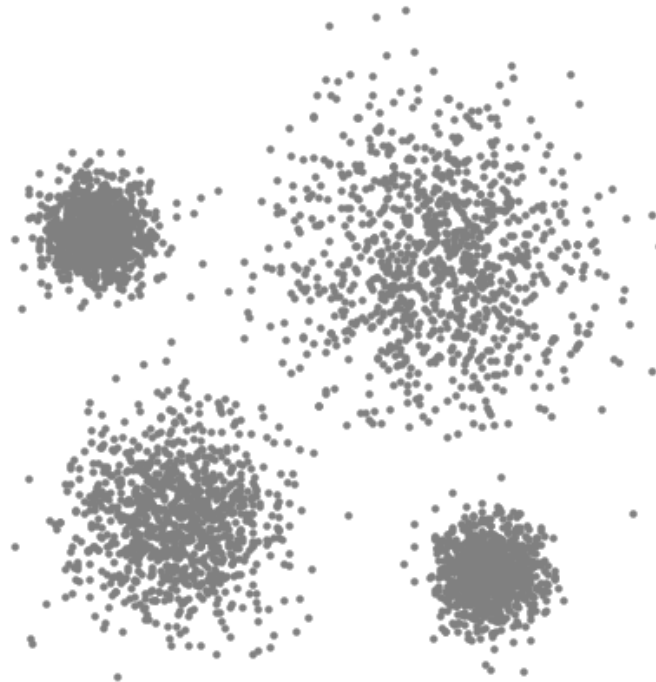
Langkah-Langkah Algoritma Meanshift:

- Memulai dengan menetapkan titik yang dipilih secara acak.
- Titik ini kemudian menjadi pusat area berbentuk lingkaran beradius sebesar r .
- Menggeser centroid dari area kerapatan rendah ke area dengan kerapatan yang lebih tinggi.
- Menggeser area lingkaran hingga area tidak mendapatkan kerapatan yang lebih tinggi.
- Mengulangi langkah 1-3 hingga titik data dikelompokkan sesuai dengan lingkaran dimana dia berada.



MEANSHIFT

- Proses perpindahan centroid hingga membentuk cluster/kelompok
- Link : <https://bit.ly/ClusteringML>



AGGLOMERATIVE HIERARCHICAL CLUSTERING

- Suatu metode hierarchical clustering yang bersifat bottom-up yaitu menggabungkan n buah klaster menjadi satu klaster tunggal.
- LINK : <https://bit.ly/ClusteringML>

AGGLOMERATIVE HIERARCHICAL CLUSTERING

Langkah-langkah

- Metode ini dimulai dengan meletakkan setiap objek data sebagai sebuah klaster tersendiri (atomic cluster)
- Selanjutnya menggabungkan klaster-klaster tersebut menjadi klaster yang lebih besar dan lebih besar lagi sampai akhirnya semua objek data menjadi satu dalam sebuah klaster tunggal.
- Kunci dari metode AHC adalah perhitungan proximity antara 2 cluster.

AGGLOMERATIVE HIERARCHICAL CLUSTERING

Contoh:

- Untuk memudahkan contoh perhitungan, di lakukan filter term yang akan diproses = a y (a dan y).
- Tiap dokumen yang ada di inialisasi sebagai cluster, jika ada 4 dokumen maka jumlah cluster juga ada 4.

D1 = a j h y i a i a y t

D2 = s d r y a e i r b g

D3 = f y t h t r a y t y

D4 = y a r y i o j g r a

AGGLOMERATIVE HIERARCHICAL CLUSTERING

- Jawab:

cluster	a	y
(D1)	3	2
(D2)	1	1
(D3)	1	3
(D4)	2	2

$$d(1,2) = \sqrt{(3-1)^2 + (2-1)^2} = 2.24$$

$$d(1,3) = \sqrt{(3-1)^2 + (2-3)^2} = 2.24$$

$$d(1,4) = \sqrt{(3-2)^2 + (2-2)^2} = 1.00$$

.....

$$d(3,4) = \sqrt{(1-2)^2 + (3-2)^2} = 1.42$$

AGGLOMERATIVE HIERARCHICAL CLUSTERING

- Jawab:

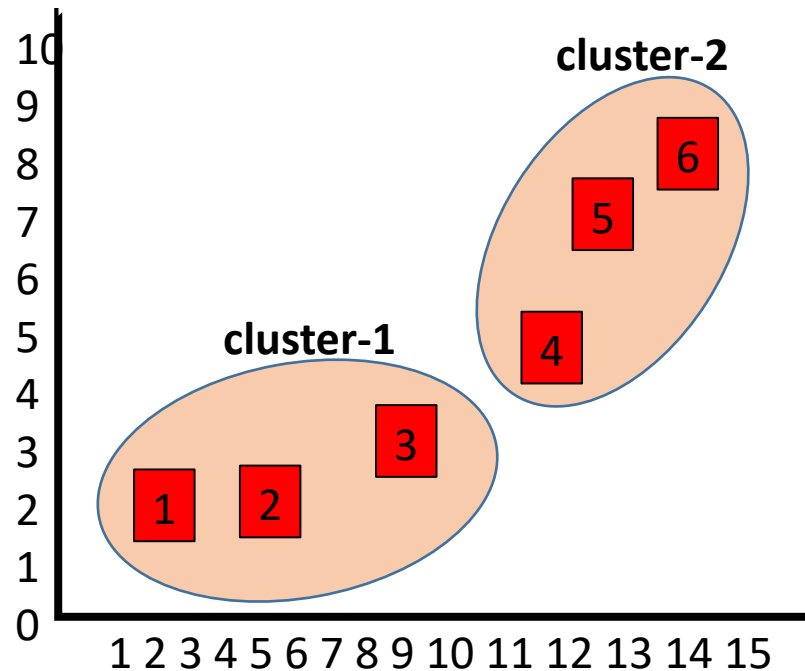
	1	2	3	4
1		2.24	2.24	1
2	2.24		2	1.42
3	2.24	2		1.42
4	1	1.42	1.42	

cluster	a	y
(D1,D4)	2.5	2
(D2)	1	1
(D3)	1	3

PERHITUNGAN SIMILARITY

- Single Linkage
jarak minimum antar cluster
- Centroid Linkage
jarak centroid antar cluster
- Complete Linkage
jarak maksimum antar cluster
- Average Linkage
jarak rata-rata antar cluster

PENGUKURAN JARAK



- Jarak dari cluster-1 ke cluster-2:
- Single Linkage
jarak objek-3 dan objek-4 (min).
- Centroid Linkage
jarak centroid cluster-1 & centroid cluster-2.
- Complete Linkage
jarak objek-1 dan objek-6 (maks).
- Average Linkage

$$\text{jarak} = \frac{A}{m * n}$$

- A = jarak objek-1 & objek-4 + jarak objek-1 & objek-5 + jarak objek-1 & objek-6 + jarak objek-2 & objek-4 + jarak objek-2 & objek-5 + jarak objek-2 & objek-6 + jarak objek-3 & objek-4 + dst..
- m = jumlah objek cluster-1
- n = jumlah objek cluster-2

VALIDASI HASIL CLUSTERING

- **Kebutuhan validasi hasil clustering**
 - **Evaluasi**
 - Mengevaluasi kualitas (seberapa baik) hasil clustering
 - **Stabilitas**
 - Mengetahui sensitivitas hasil clustering terhadap parameter algoritma, misalnya jumlah cluster
 - **Tendensi**
 - Menilai kesesuaian hasil clustering, misalnya apakah data memiliki karakteristik pengelompokan tertentu

EVALUASI KUALITAS HASIL CLUSTERING

- Internal
 - Menggunakan informasi internal, seberapa **compact** dan **separated** cluster-cluster yang dihasilkan
 - ***Silhouette coefficient***
- Eksternal
 - Membandingkan hasil clustering dengan label kelas eksternal yang tersedia (ground truth)
 - **Entropy, purity, F-measure**
- Relatif
 - Membandingkan dua hasil clustering, misalnya yang didapat dari satu algoritma yang sama dengan nilai parameter (***k***, centroid awal) yang berbeda
 - **SSE atau entropy**
 - Digunakan untuk menentukan hasil clustering optimal, misalnya dengan memvariasikan nilai parameter jumlah cluster ***k***, posisi awal centroid

SILHOUETTE COEFFICIENT

- *Silhouette* dapat memberikan representasi grafis yang ringkas tentang seberapa baik **setiap objek** telah dikelompokkan.
- Nilai *silhouette* adalah ukuran seberapa mirip suatu objek dengan clusternya sendiri (kohesi) dibandingkan dengan cluster lain.
- Nilai *silhouette* berkisar antara -1 hingga $+1$. Nilai 1 menyatakan hasil terbaik dimana suatu data sangat compact dengan data lain pada cluster yang sama dan jauh dari cluster-cluster lain.
- Silhouette dapat dihitung dengan rumus jarak apa pun, seperti jarak Euclidean atau jarak Manhattan.
- Jika sebagian besar objek memiliki nilai *silhouette* tinggi, maka hasil clustering sudah baik.
- Jika banyak objek memiliki nilai *silhouette* rendah atau negatif, maka hasil clustering masih kurang baik, mungkin memiliki terlalu banyak atau terlalu sedikit jumlah cluster.

SILHOUETTE COEFFICIENT

- Silhouette coefficient dari setiap data x dari hasil clustering dapat dihitung dengan :

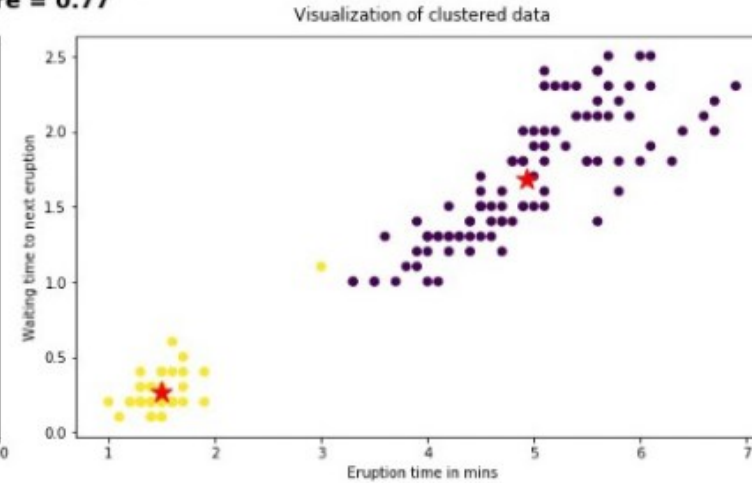
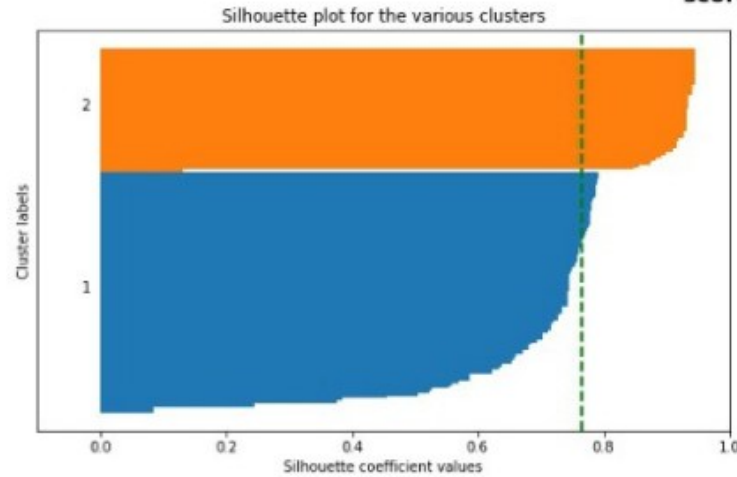
$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

- $a(x)$ adalah rata-rata jarak antara x dan semua data lain pada cluster yang sama
- $b(x)$ adalah minimum rata-rata jarak antara x dengan semua cluster yang lain
- Untuk setiap data $x \in C_x$ (data x anggota cluster C_x)

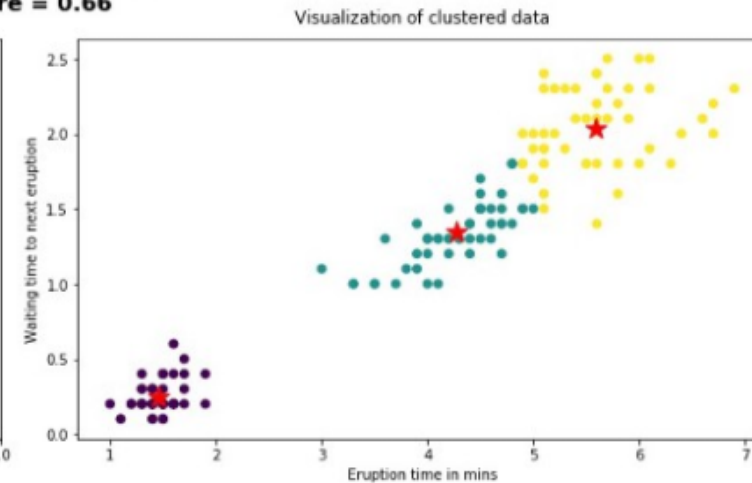
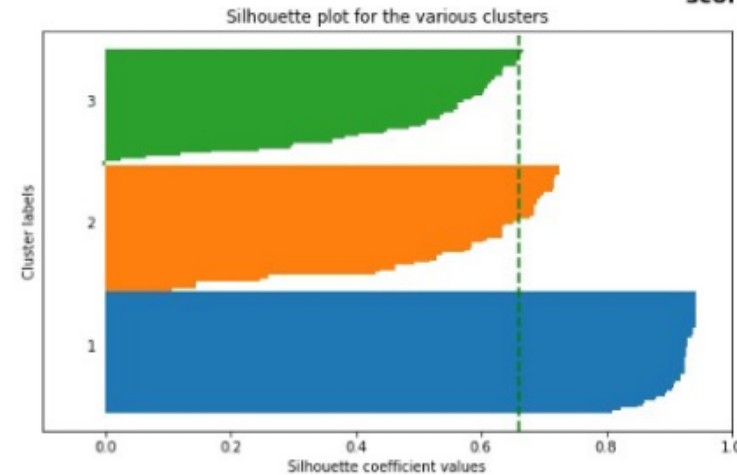
$$a(x) = \frac{1}{|C_x|-1} \sum_{j \in C_x, x \neq y} d(x, y) \text{ dan } b(x) = \min_{z \neq x} \frac{1}{|C_z|} \sum_{y \in C_z} d(x, y)$$

SILHOUETTE COEFFICIENT

Silhouette analysis using $k = 2$
score = 0.77



Silhouette analysis using $k = 3$
score = 0.66



TERIMA KASIH