Penambangan Data

Program Studi Magister Sistem Informasi
Fakultas Ilmu Komputer dan Rekayasa

Dr. Mardiani, S.Si., M.T.I

UNIVERSITAS MDP

Metode Clustering

KAMPUS INOVASI

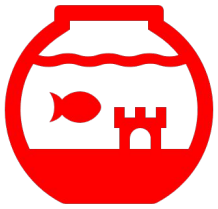
Pokok Bahasan

- Konsep Clustering
- Algoritma Clustering: K-Means, Hierarchical Clustering
- Implementasi K-Means dan Hierarchical Clustering
- Evaluasi hasil clustering

Clustering

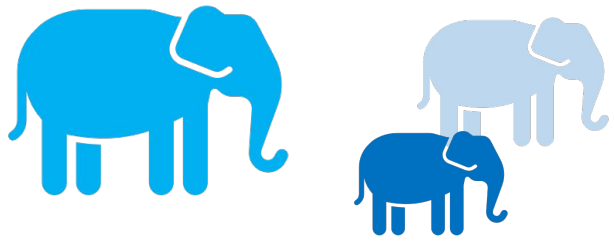
Pengklusteran merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan.

Beberapa algoritma pengelompokkan diantaranya adalah K-Means



Clustering Main Features

- Clustering – a data mining technique
- Usage:
  - Statistical Data Analysis
  - Machine Learning
  - Data Mining
  - Pattern Recognition
  - Image Analysis
  - Bioinformatics



Notion of a Cluster can be Ambiguous

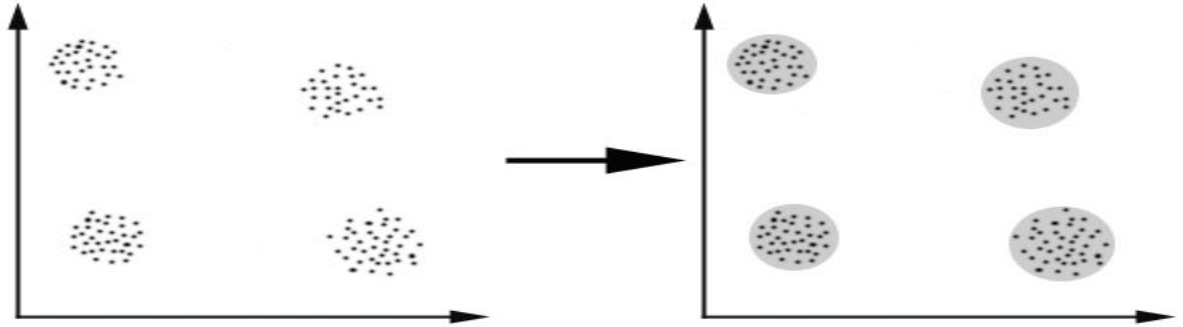
How many clusters?

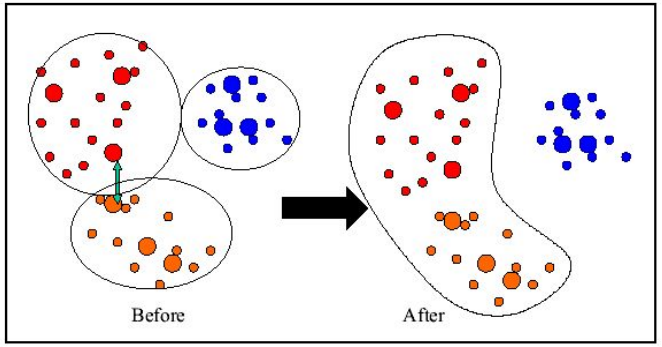
Six Clusters

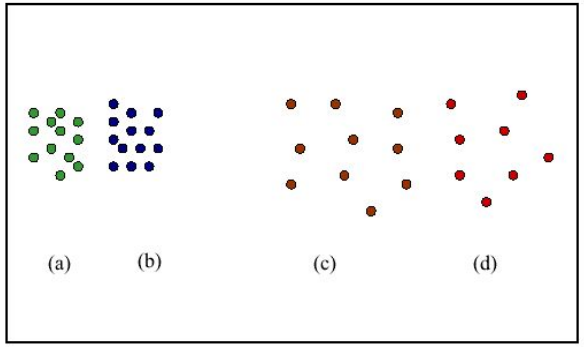
Two Clusters

Four Clusters

Distance based method

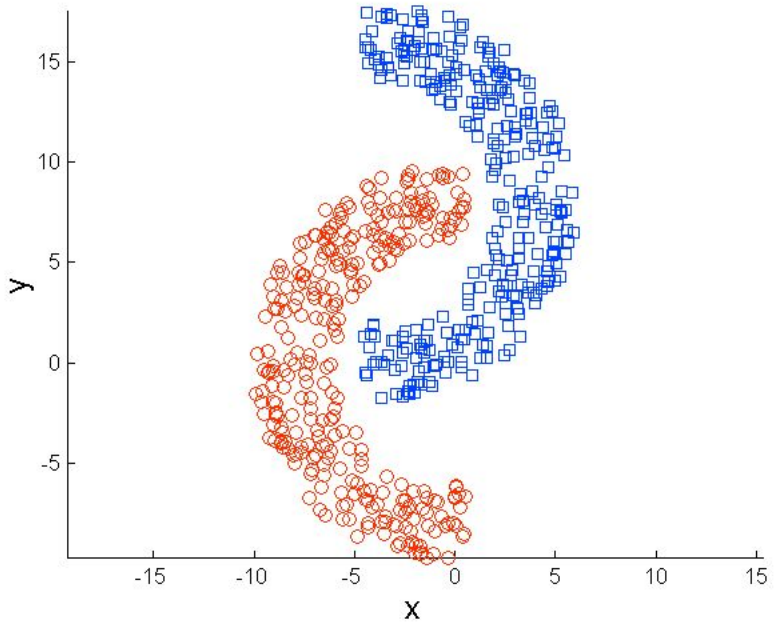




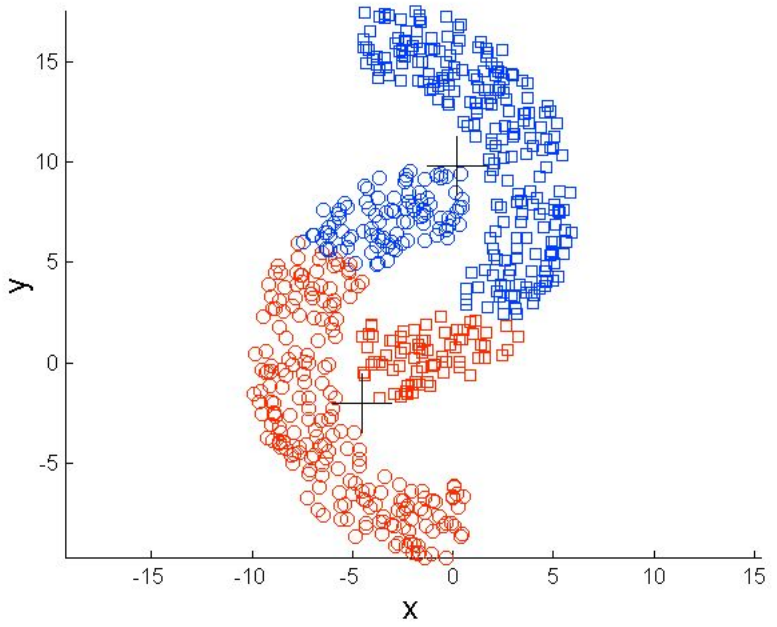


Dalam kasus ini, dapat dengan mudah diidentifikasi 4 cluster tempat data dapat dibagi; kriteria kesamaan adalah jarak: dua atau lebih objek termasuk dalam klaster yang sama jika keduanya “dekat” menurut jarak tertentu. Ini disebut pengelompokan berbasis jarak.

Limitations of K-means: Non-globular Shapes

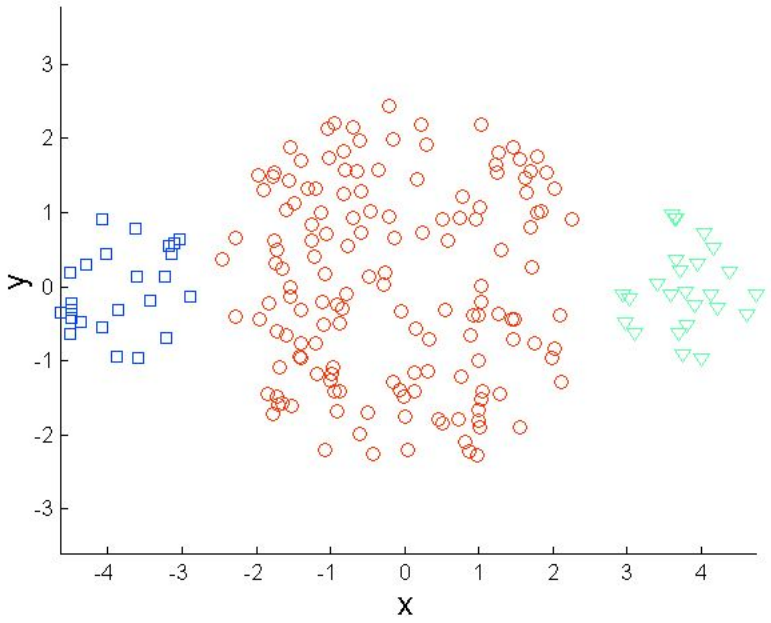


Original Points

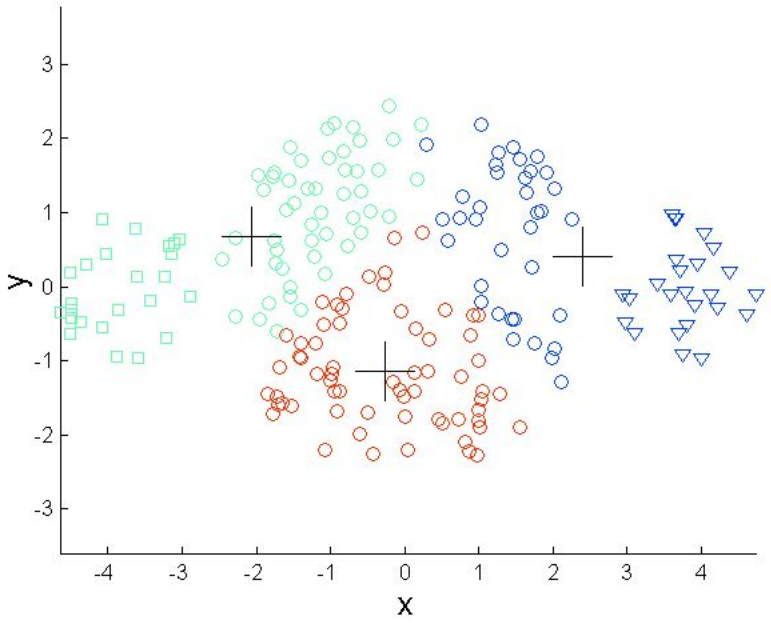


K-means (2 Clusters)

Limitations of K-means: Differing Sizes



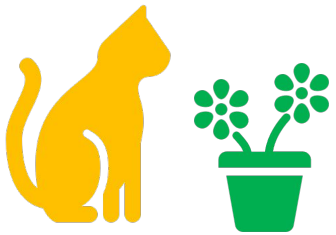
Original Points



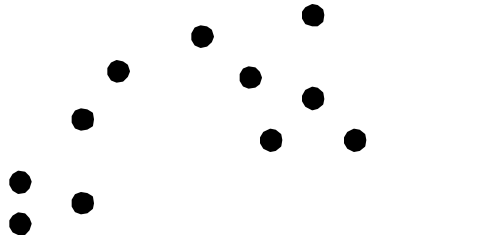
K-means (3 Clusters)

Types of Clustering

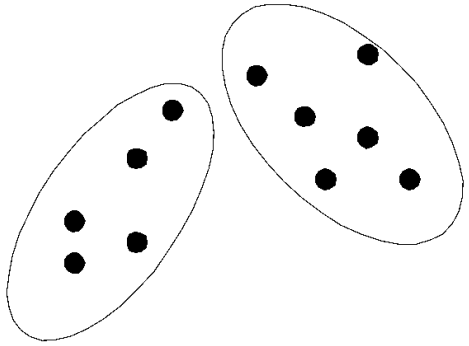
- Hierarchical
  - Menemukan cluster baru menggunakan cluster yang ditemukan sebelumnya
- Partitional
  - Menemukan semua cluster sekaligus

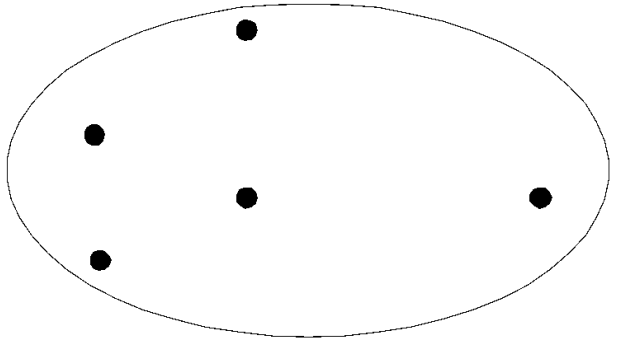


Partitional Clustering



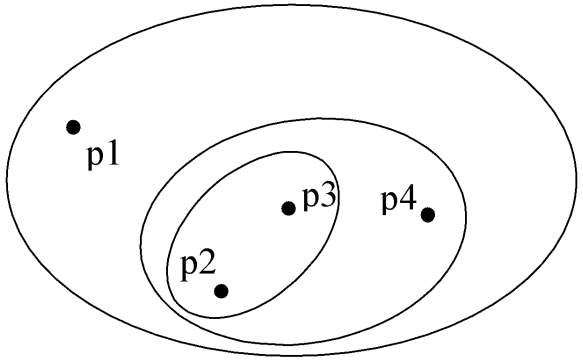
Original Points



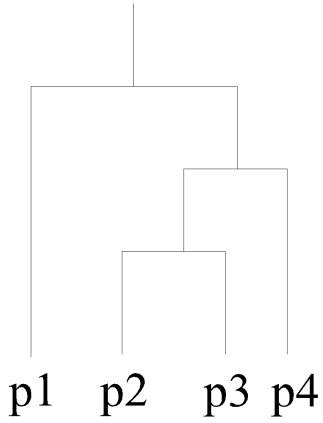


A Partitional Clustering

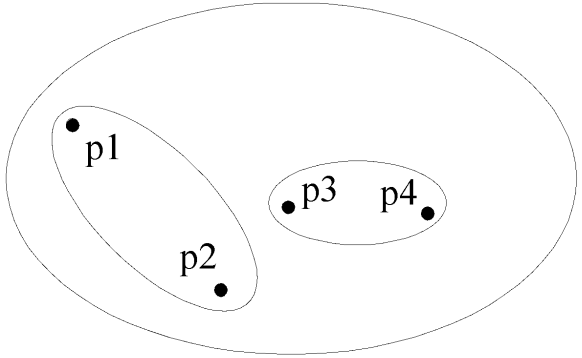
Hierarchical Clustering



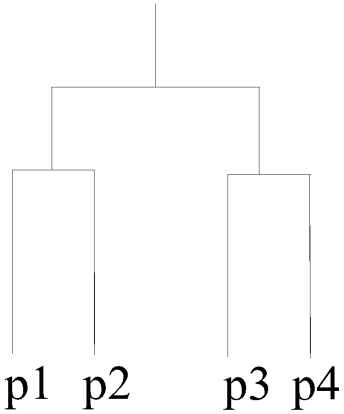
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Algoritma Pengelompokan K-Means

Langkah-langkah algoritma K-Means:

1. Tentukan berapa kelompok yang akan dibuat sebanyak k kelompok.
2. Secara sembarang pilih k buah catatan yang ada sebagai pusat-pusat keompok awal.
3. Setiap catatan akan ditentukan pusat kelompok terdekatnya.
4. Perbarui pusat-pusat kelompok.
5. Pusat kelompok yang terdekat pada setiap catatan akan ditentukan, dan seterusnya sampai nilai rasio tidak membesar lagi.

Rumus Jarak Euclidean dua titik:

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$$

Between Cluster Variation (BCV):

BCV=$d(m_1,m_2)+d(m_1,m_3)+d(m_2,m_3)$

Dalam hal ini, $d(m_{i,j})$ menyatakan jarak m_i ke m_j

Within Cluster Variation (WCV):

WCV=Σ(jarak pusat tiap cluster yang paling minimum)^2

$$\text{rasio}=\frac{\text{BCV}}{\text{WCV}}$$

Untuk mengikuti Mata Kuliah ITI 372 Penelitian Operasional II, seorang mahasiswa harus lulus dua mata kuliah prasyarat, yaitu ITI 371 Penelitian Operasional I, AMA 213 Matriks dan Ruang Vektor serta AMA 214 Kalkulus. Diketahui data 6 orang mahasiswa sebagai berikut (A=4,B=3,C=2 dan D=1):

Mahasiswa	I	II	III	IV	V	VI
ITI 371	D=1	A=4	C=2	B=3	B=3	A=4
AMA 213	D=1	C=2	C=2	B=3	C=2	A=4
AMA 214	C=2	A=4	C=2	B=3	B=3	A=4

Kelompokkan mahasiswa-mahasiswa tersebut menjadi 2 kelompok, dengan ketentuan awal kelompok pertama beranggotakan mahasiswa I,II dan IV serta kelompok kedua beranggotakan mahasiswa III,V dan VI. Pusat setiap kelompok masing-masing adalah mahasiswa I dan mahasiswa VI, gunakan algoritma K-Means sampai iterasi kedua untuk penentuan cluster tersebut!

Iterasi pertama :
Kelompok 1=I,II dan IV $m_1(1,1,2)$
Kelompok 2=III,V dan VI dengan $m_2(4,4,4)$
Rumus jarak dua titik:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Misal jarak I ke kel 2:
I(1,1,2) dan $m_2(4,4,4)$ ada 3 titik:

$$d(I,m_2) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

$$= \sqrt{(1-4)^2 + (1-4)^2 + (2-4)^2} = \sqrt{9+9+4} = \sqrt{22} = 4{,}69$$

dst

Catatan	Jarak kel 1	Jarak kel 2	Jarak terdekat
I	0	4,6904158	C1
II	3,741657	2	C2
III	1,414214	3,4641016	C1
IV	3	1,7320508	C2
V	2,44949	2,4494897	C2
VI	4,690416	0	C2

Between Cluster Variation (BCV):

$BCV = d(m_1,m_2) + d(m_1,m_3) + d(m_2,m_3)$

$BCV = 4.6904$

Within Cluster Variation (WCV):

$WCV = \Sigma(\text{jarak pusat tiap cluster yang paling minimum})^2$

$WCV = 0^2 + 2^2 + 1{,}411^2 + 1{,}732^2 + 2{,}449^2 + 0^2$

$WCV = 0+4+2+3+6+0 = 15$

$$\text{rasio} = \frac{BCV}{WCV}$$

Ratio = 4.6904/15 = 0.3126

Iterasi kedua:

Cari rata-rata untuk nilai tengah setiap kelompok:

Kelompok 1=I dan III m_1(1.5 , 1.5, 2)

Kelompok 2=II,IV,V dan VI dengan m_2(3.5 , 2.75 , 3.5)

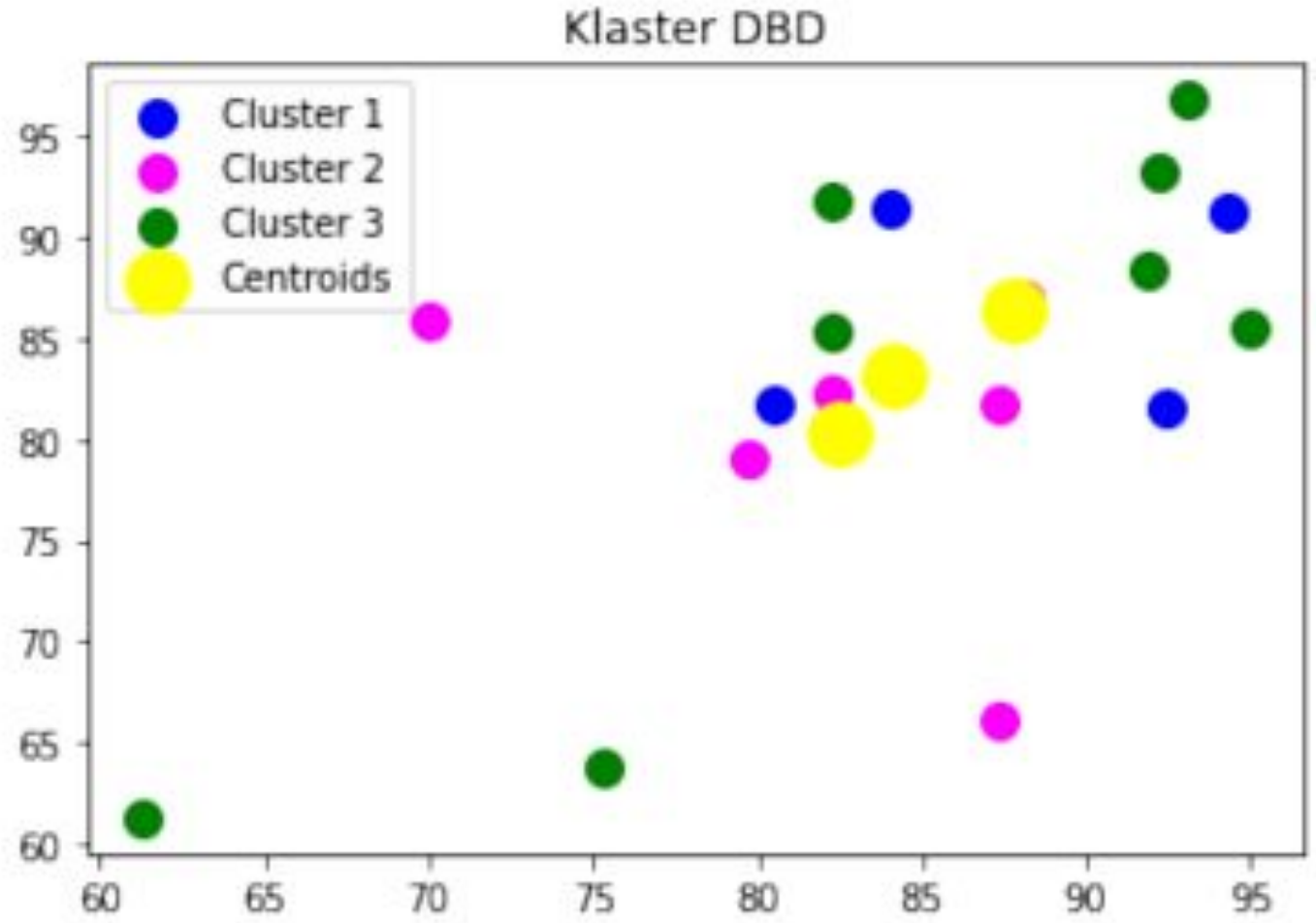
Catatan	Jarak kel 1	Jarak kel 2	Jarak terdekat
I	0,707107	3,400368	C1
II	3,24037	1,030776	C2
III	0,707107	2,25	C1
IV	2,345208	0,75	C2
V	1,870829	1,030776	C2
VI	4,062019	1,436141	C2

BCV = 2.795

WCV = 0.5+1.0625+0.5+0.5625+1.0625+2.0625=5.75

Ratio = 2.795/5.75 = 0.486

Contoh penjelasan dari hasil Cluster



Sehingga dapat disimpulkan hasil klasterisasi wilayah DBD dengan K=3 di kabupaten Bangkalan adalah

Klaster	Kecamatan
Klaster 1	Galis, Bangkalan, Geger, Kokop
Klaster 2	Labang, Kwanyar Modung, Tragah, Arosbaya, sepulu
Klaster 3	Kamal, Blega, Konang,Tanah Merah, Socah, Burneh, Tanjung bumi, Klampis

Kesimpulan

Klasterisasi atau pengelompokan dari Kasus Demam Berdarah (DBD) yang ada dikabupaten Bangkalan sampai tahun 2019 bertujuan untuk mengelompokkan wilayah kecamatan terhadap kasus DBD. clustering adalah suatu Teknik untuk kelompok data yang homogen sedemikian rupa sehingga titik data di setiap cluster semirip mungkin menurut ukuran kesamaan seperti jarak berbasis euclidean atau jarak berbasis korelasi. Keputusan tentang ukuran kemiripan dapat ditentukan melalui jumlah centroid masing-masing kelompok dan jaraknya.

Dalam proses klasterisasi ini, seleksi fitur dilakukan dengan Teknik Feature Score dan Matriks korelasi Fitur menggunakan heatmap. Hasil seleksi fitur kedua Teknik tersebut menghasilkan fitur yang sedikit berbeda. Sedangkan dari dinas Kesehatan juga memberikan faktor yang mempengaruhi jumlah kasus DBD secara umum. Maka pada tugas ini digunakan fitur berdasarkan Dinas Kesehatan serta mempertimbangkan hasil dari Teknik seleksi fitur. Terdapat 10 Fitur yang digunakan meliputi ABJ 2017,ABJ2018,ABJ2019, DBD2017,DBD2018,DBD2019, luas, tinggi wilayah, hujan bulanan 2019, jumlah penduduk

Evaluasi Hasil Clustering

- **Inertia (Within-Cluster Sum of Squares, WCSS)**:

Inertia mengukur total jarak kuadrat antara data dalam kluster dan pusat klusternya. Semakin rendah inertia, semakin baik hasil clustering.

- **Metode Elbow**

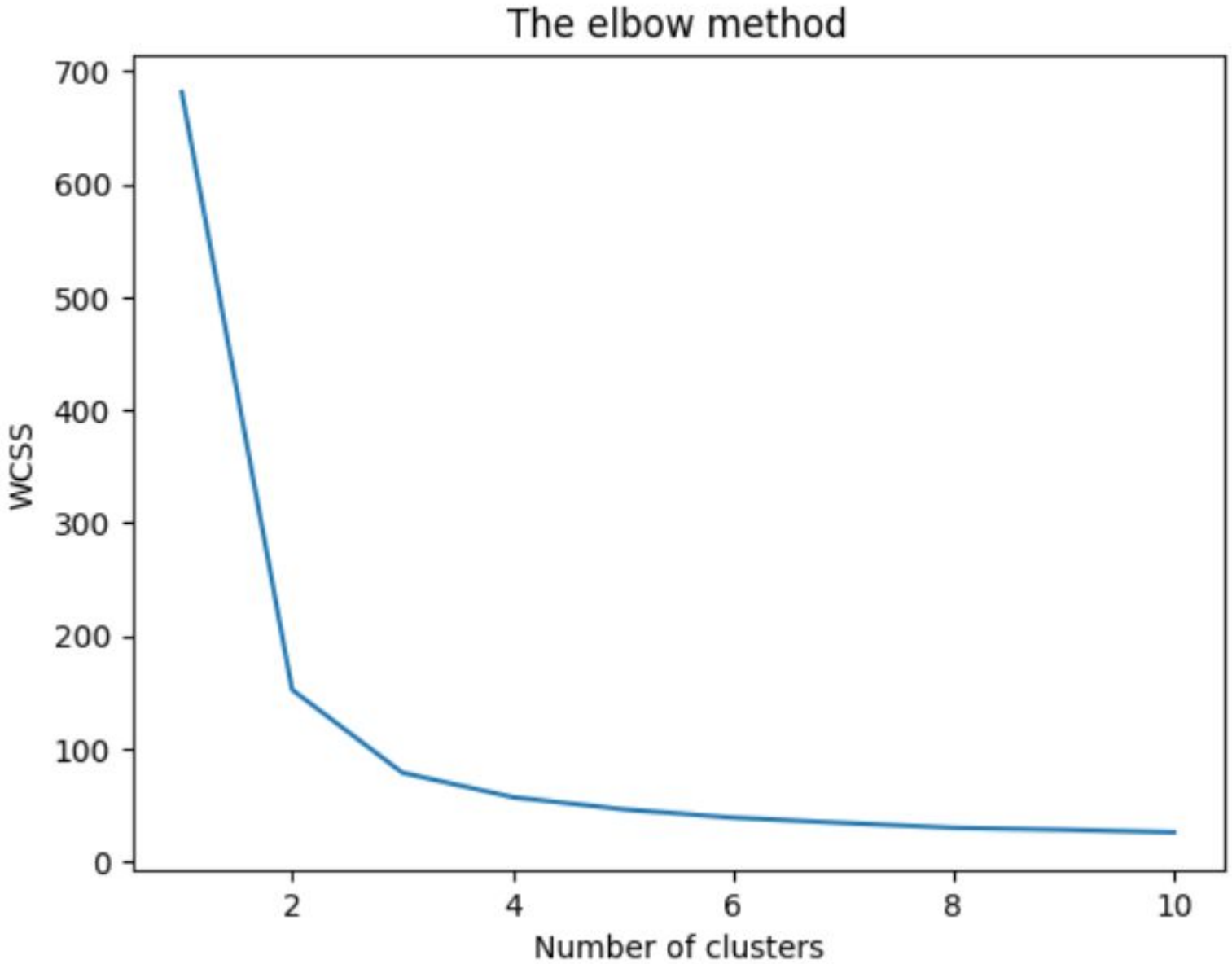
Dalam memilih jumlah kluster K, grafik inertia vs jumlah kluster K akan membentuk sudut atau "elbow" pada nilai K yang optimal.

- WCSS (Within-Cluster Sum of Squares) adalah metrik yang mengukur total jumlah kuadrat jarak antara setiap titik data dalam suatu kluster dengan centroid (pusat) klusternya. Ini memberikan indikasi seberapa "kompak" kluster tersebut. Semakin kecil nilai WCSS, semakin baik hasil clustering karena berarti data dalam kluster lebih dekat dengan pusat kluster.

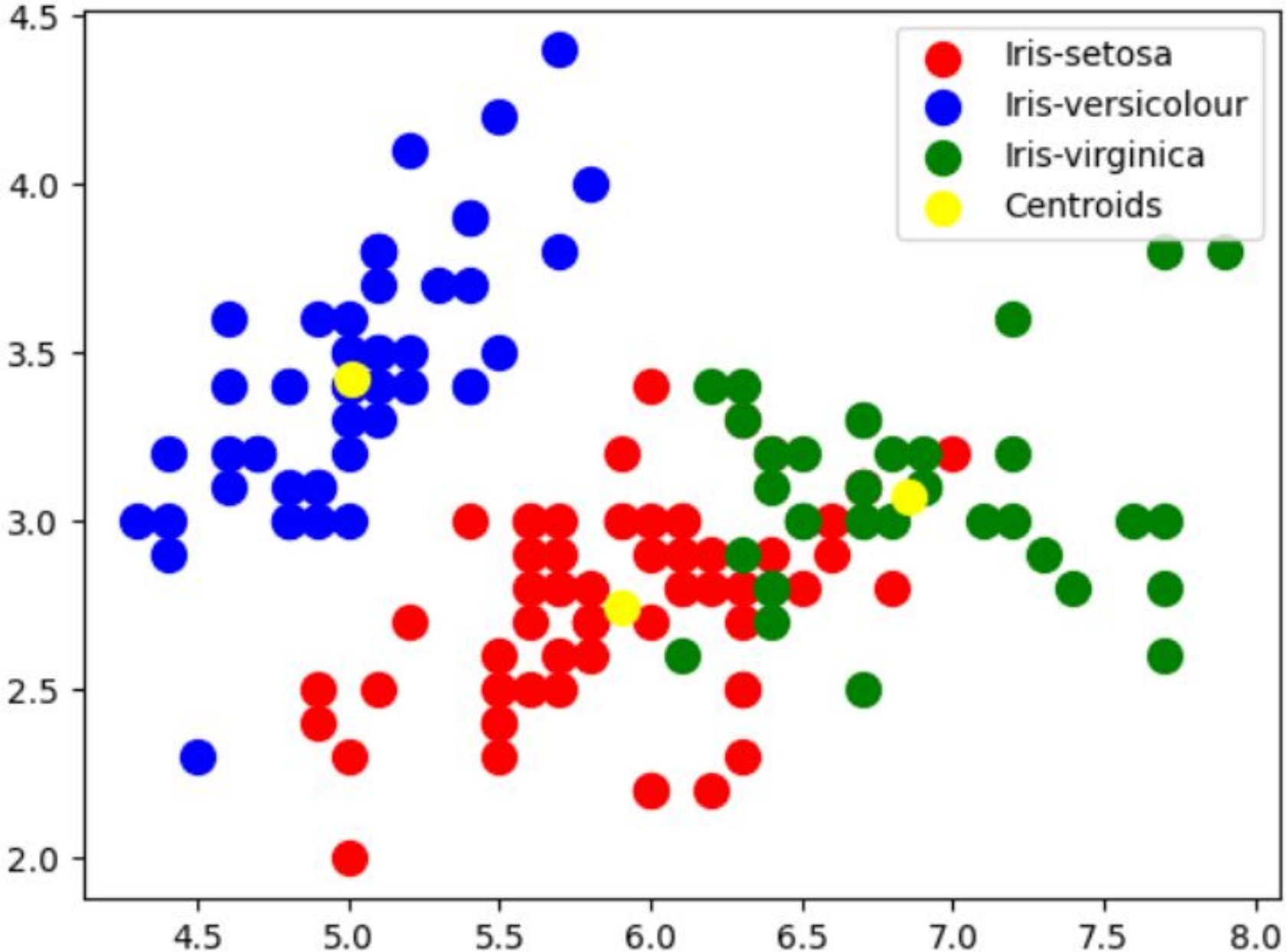
$$\text{WCSS} = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2$$

di mana:

- K: jumlah kluster.
- C_i: himpunan data dalam kluster i.
- x: titik data dalam kluster C_i.
- μ_i: centroid atau rata-rata titik-titik dalam kluster C_i.
- $\|x - \mu_i\|^2$: jarak Euclidean antara titik x dan centroid μ_i, yang kemudian dikuadratkan.


The elbow method
WCSS
Number of clusters
700
600
500
400
300
200
100
0
2
4
6
8
10

Clustering dengan Python


Iris-setosa
Iris-versicolour
Iris-virginica
Centroids
4.5
4.0
3.5
3.0
2.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

Clustering dengan Rapidminer

<new process*> – Altair AI Studio Educational 2024.1.0 @ DESKTOP-92VTQN7
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Repository
Import Data
Deals
Deals-Testset
Golf
Golf-Testset
Iris
Labor-Negotiations
Market-Data
Polynomial
Operators
mean
Segmentation (6)
k-Means
k-Means (H2O)
k-Means (Kernel)
k-Means (fast)
We found "MeaningCloud Text Analytics", "Prescriptive Analytics" and one more result in the Marketplace. Show me!
Process
Process
Process
Retrieve Iris
Clustering
inp
out
exa
clu
clu
res
res
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Clustering (k-Means)
add cluster attribute
add as label
remove unlabeled
k 5
max runs 10
Show advanced parameters
Change compatibility (10.5.000)
Help
k-Means
Concurrency
Tags: Unsupervised, Clustering, Segment Grouping, Similarity, Similarities, Euclide Distances, Centroids, K Means, K means, Kmeans
Synopsis

Result History

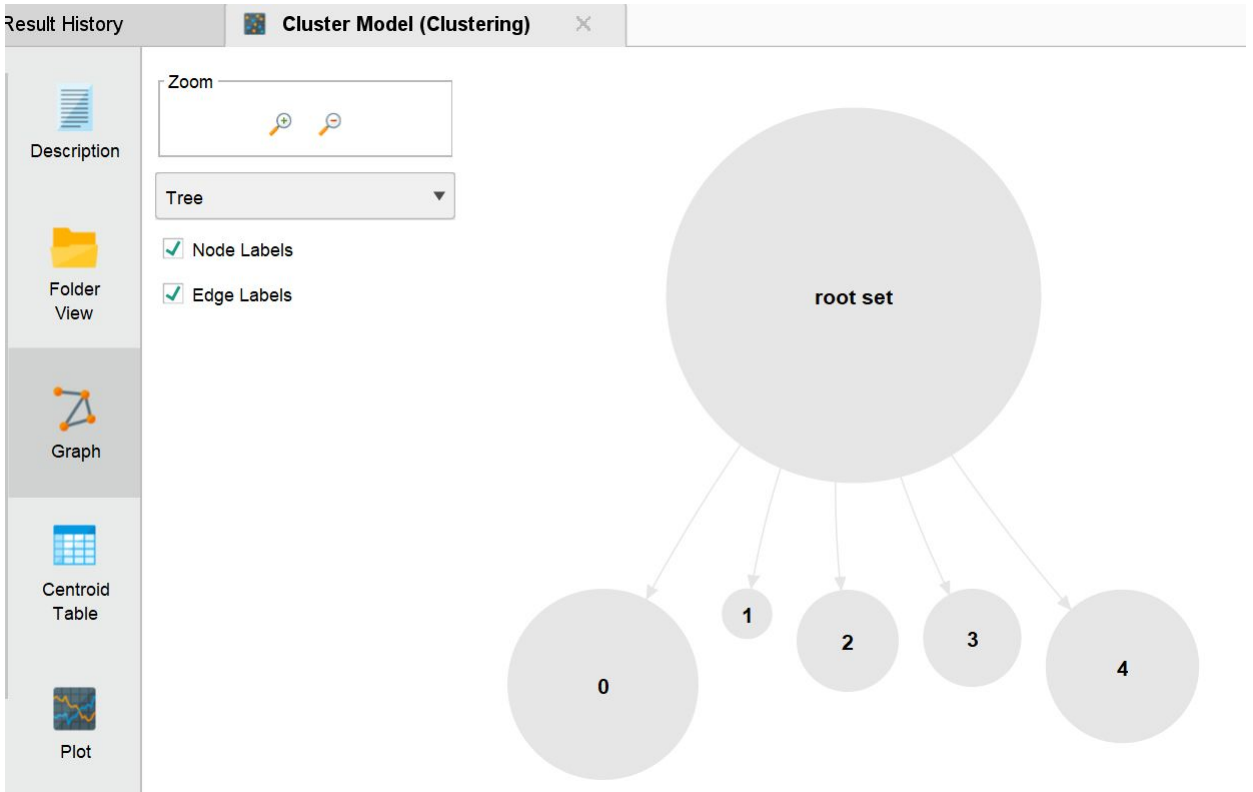
Cluster Model (Clustering)

Description

Folder View

Cluster Model

```
Cluster 0: 50 items
Cluster 1: 12 items
Cluster 2: 25 items
Cluster 3: 24 items
Cluster 4: 39 items
Total number of items: 150
```



Result History

Cluster Model (Clustering)

Description

Folder View

Graph

Centroid Table

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
a1	5.006	7.475	5.508	6.529	6.208
a2	3.418	3.125	2.600	3.058	2.854
a3	1.464	6.300	3.908	5.508	4.746
a4	0.244	2.050	1.204	2.162	1.564

UNIVERSITAS MDP

TERIMA KASIH

KAMPUS INOVASI

Dr. Mardiani, S.Si., M.T.I