2

ANALISIS CLUSTER NON-HIERARKI

(K-MEANS)

Muhammad Amanda [1)], Syifa Azzahra [2)], Indah Lestari[3)]
[1]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta
email: muhammadamanda263@gmail.com
[2]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta
email: Arhazza.afiys@gmail.com
[3]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta
email: il4620715@gmail.com

A. Pengertian Analisis *Cluster*

Analisis *Cluster* merupakan suatu teknik multivariat yang bertujuan untuk membentuk kelompok-kelompok atau mengklasifikasikan objek-objek ke dalam kelompok yang homogen berdasarkan karekteristik tertentu yang dimilikinya. Teknik ini mengklasifikasikan objek-objek sehingga setiap objek yang paling dekat kesamaannya dengan objek lainnya berada dalam satu kelompok yang sama.

Ciri-ciri *cluster* yang baik adalah *cluster* yang mempunyai :

1. Homogenitas atau kesamaan yang tinggi antar anggota dalam satu kelompok (*within-cluster*).
2. Heterogenitas atau perbedaan yang tinggi antar kelompok yang satu dengan kelompok yang lainnya (*between-cluster*).

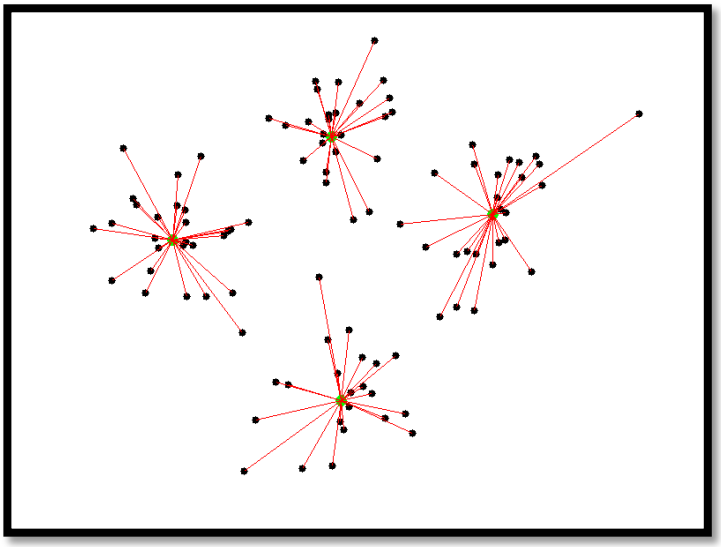
Secara umum, metode yang digunakan dalam analisis *cluster* yaitu metode hirarki dan non-hirarki. Pada metode hirarki memulai pengelompokkan dengan 2 atau lebih objek yang mempunyai kesamaan paling dekat, terdapat tingkatan yang jelas antar objek, dari yang paling mirip hingga yang paling tidak mirip, biasanya dibantu dengan dendogram. Sedangkan mettode non-hirarki dimulai dengan menentukan dahulu jumlah kelompok yang diinginkan. Tidak seperti metode hirarki yang meliputi proses *treelike construction,* metode non-hirarki langsung menempatkan objek-objek ke dalam kelompok secara sekaligus.

B. Metode Non-hirarki K-Means

Metode non-hirarki *K-means* biasanya digunakan untuk ukuran dara yang besar. Kelemahan dalam metode terletak pada pemilihan pusat cluster yang sembarang dan banyaknya *cluster* yang sudah ditentukan terlebih dahulu. Hasil akhir *cluster* yang terbentuk mungkin tergantung pada kedua hal tersebut. Kelebihan metode ini yaitu, metode non-hirarki lebih cepat dan lebih unggul digunakan untuk ukuran data yang besar. Tahap-tahap pengelompokkan pada metode ini yaitu :

1. Partisi seluruh item ke dalam K inisial gerombol.
2. Tentukan centroid atau rata-rata dari setiap gerombol.

3. Hitung jarak antara setiap item dengan centroid. Tempatkan item ke dalam gerombol yang berdekatan nilai centroidnya.
4. Hitung kembali centroid untuk gerombol yang menerima item baru dan untuk gerombol yang kehilangan item.
5. Ulangi langkah 3-4 sampai tidak ada penempatan ulang item atau tidak ada lagi item yang berpindah tempat.



Gambar 2.1 Algoritma K-Means
(Sumber: https://www.dicoding.com/academies/184/tutorials/8417)

C. Distance Space Untuk Menghitung Jarak Antara Data dan Centroid

Jarak adalah pendekatan yang digunakan untuk menentukan kemiripan dua vektor fitur yang dinyatakan dengan *ranking*. Semakin kecil nilai *ranking* yang dihasilkan, semakin tinggi kemiripan antara kedua objek tersebut. beberapa jarak yang biasa digunakan dalam analisis *cluster* diantaranya :

1. *Euclidian Distance*

Jarak *Euclidian* merupakan jarak yang paling umum digunakan dalam analisis *cluster*. Jarak ini dapat digunakan apabila semua peubahnya berskala kontinu dengan asumsi peubah-peubah yang diamati tidak berkorelasi dan antar peubah memiliki satuan yang sama.

$$D(x_2, x_1) = \sqrt{\sum_{j=1}^{d} |x_{2j} - x_{1j}|^2}$$

2. *Manhattan Distance*

Jarak *Manhattan* digunakan jika peubah yang diamati tidak saling bebas (berkorelasi).

$$D(x, y) = \sum_{j=1}^{d} |x_j - y_j|$$

3. *Mahalonobis Distance*

Jarak *Mahalonobis* berfungsi untuk menghilangkan atau mengatasi perbedaan skala pada masing-masing peubah.

$$D(x, y) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{s_i^2}}$$

D. Contoh Soal Clustering Non-Hirarki *K-Means*

Diketahui 6 buah sample data yang ditunjukkan pada tabel dibawah ini. Misalkan cluster 1 adalah K1 dan cluster 2 adalah K2.

Sample Data	X	Y	Kelompok/Cluster
1	100	50	
2	40	60	
3	30	70	
4	90	10	
5	65	40	
6	25	35	

Lakukan pengelompokkan menggunakan metode K-Means Clustering!

Penyelesaian :

- Perhitungan pertama gunakan data ke-1 dan ke-2. Gunakan rumus E*uclidean Distance* untuk menghitung jarak minimum data terhadap centroid.

Cluster	X	Y
K1	100	50
K2	40	60

- Hitung centroid pertama.

- Cluster 1 (100,50) $= \sqrt{(100-100)^2 + (50-50)^2} = 0$ (jarak cluster 1 ke cluster 1)
- Jarak cluster 2 ke cluster 1 (40,60) → (100,50)

$$= \sqrt{(40-100)^2 + (60-50)^2} = \sqrt{(60)^2 + (10)^2} = 60{,}83$$

- Jarak cluster 1 ke cluster 2 (100,50) → (40,60)

$$= \sqrt{(100-40)^2 + (50-60)^2} = \sqrt{(60)^2 + (-10)^2} = 60{,}83$$

- Jarak cluster 2 ke cluster 2 (40,60) $= \sqrt{(40-40)^2 + (60-60)^2} = 0$

Hasil pengelompokan cluster data ke-1 masuk K1 dan data ke-2 masuk K2

Cluster	Centroid		Kelompok Cluster
	X	Y	
K1 (100,50)	0	60.83	1
K2 (40,60)	60.83	0	2

Dari tabel diatas, dapat dilihat bahwa : jarak minimum dari data 1 ke data 1 adalah 0 dan jarak minimum dari data 2 ke data 2 adalah 0. Jadi, centroid K1 adalah data 1 dan data centroid K2 adalah data 2.

- Berlanjut ke data ke-3 yaitu (30, 70) untuk mendapatkan centroid selanjutnya. Hitung jarak data ke-3 terhadap centroid 1 dan centroid 2. Jadi kita bisa menentukan apakah data 3 akan masuk ke cluster K1 atau cluster K2.

- Jarak data 3 ke cluster 1
  (100,50) → (30,70)

$$= \sqrt{(30-100)^2 + (70-50)^2} = \sqrt{(-70)^2 + (20)^2} = 72{,}80$$

- Jarak data 3 ke cluster 2
  (40,60) → (30,70)

$$= \sqrt{(30-40)^2 + (70-60)^2} = \sqrt{(-10)^2 + (10)^2} = 14{,}14$$

Hitung dengan menggunakan rumus *Euclidean Distance.* Hasil perhitungan disajikan dalam tabel berikut.

Dataset	Euclidean Distance		Kelompok/Cluster
	Cluster 1	Cluster 2	
Data 3(30,70)	70,8	14,14	2

Selanjutnya *update* nilai Centroid. Karena data 3 masuk ke cluster K2, maka centroidnya di*update* dengan rumus sebagai berikut :

$$X_{centroid\_baru} = \frac{(X_{K2} + X_{data3})}{2}$$

$$Y_{centroid\_baru} = \frac{(Y_{K2} + Y_{data3})}{2}$$

Berikut adalah hasil perhitungan centroid yang baru.

Cluster	X	Y
K1	100	50
K2	$\frac{40+30}{2} = 35$	$\frac{60+70}{2} = 65$

- Selanjutnya beralih ke dataset ke-4, hitung jarak antar data dan Centroid K1. Berikut ini adalah hasil perhitungannya.

  (100,50) → (90,10)

  $$= \sqrt{(90-100)^2 + (10-50)^2} = \sqrt{(-10)^2 + (-40)^2} = 41{,}23$$

  Selanjutnya hitung jarak antar data dan Centroid K2. Karena centroid K2 sudah diupdate, maka gunakan nilai centroid K2 yang baru (35, 65). Hal ini berlaku juga jika centroid K1 diupdate. Berikut ini perhitungannya.

  (35,65) → (90,10)

  $$= \sqrt{(90-35)^2 + (10-65)^2} = \sqrt{(55)^2 + (-55)^2} = 77{,}7$$

  Dari kedua perhitungan jarak dataset ke K1 dan K2, hasilnya sebagai berikut :

Dataset	Euclidean Distance		Kelompok/Cluster
	Cluster 1	Cluster 2	
Data 4(90,10)	41,23	77,7	1

Dari tabel tersebut, dapat dilihat bahwa data ke-4 masuk kedalam cluster K1. Sama seperti langkah sebelumnya, maka update kembali centroid K1 dengan data ke-3.

Cluster	X	Y
K1	$\frac{100+90}{2} = 95$	$\frac{50+10}{2} = 30$
K2	35	65

Cluster centroid yang baru dapat dilihat pada tabel berikut.

Cluster	X	Y
K1	95	30
K2	35	65

Perhitungan untuk dataset selanjutnya sama seperti langkah-langkah sebelumnya. Selalu gunakan centroid yang sudah di update serta lakukan update centroid baru. Semua data hasil perhitungan untuk menentukan cluster dapat dilihat pada tabel berikut.

Sample Data	X	Y	Kelompok/Cluster
1	100	50	1
2	40	60	2
3	30	70	2
4	90	10	1
5	65	40	1

E. Contoh Penerapan Metode Non-Hirarki K-Means

Segmentasi customer merupakan proses pengelompokan customer ke dalam kelompok-kelompok kategori berdasarkan kedekatan karakteristik yang dimiliki oleh masing-masing customer. Segmentasi customer memungkinkan perusahaan dalam

memahami karakteristik konsumen pada masing-masing kelompok kategori sehingga perusahaan dapat membangun strategi pemasaran yang sesuai dengan karakteristik masing-masing kelompok kategori. Segmentasi customer umum digunakan sebagai dasar pertimbangan pelaksanaan strategi pemasaran yang banyak dilakukan oleh perusahaan didasarkan pada demografi dan perilaku customer sehingga memudahkan perusahaan untuk mengidentifikasi customer melalui karakteristik yang dimiliki.

Segmentasi dilakukan dengan menggunakan data yang merupakan data customer mall yang didapatkan dari *https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python*. Data terdiri dari 200 observasi dengan 5 variabel, yaitu Customer ID, jenis kelamin, usia (dalam tahun), penghasilan (dalam ribu $), dan nilai pengeluaran (dalam rentang 1-100). Segmentasi dilakukan dengan menggunakan metode K-means dan variabel yang menjadi dasar segmentasi yaitu kelompok usia, penghasilan, serta pengeluaran.

❖ Import dataset

Membaca dataset menggunakan fungsi `read.csv()` yang terdapat dalam package `readr` yang disimpan dalam `data` lalu menampilkan 6 data pertama.

```
library(readr)
data = read.csv("Mall_Customers.csv")
head(data)
```

	CustomerID <int>	Gender <chr>	Age <int>	Annual.Income..k.. <int>	Spending.Score..1.100. <int>
1	1	Male	19	15	39
2	2	Male	21	15	81
3	3	Female	20	16	6
4	4	Female	23	16	77
5	5	Female	31	17	40
6	6	Female	22	17	76

6 rows

Gambar 2.2 Data customer mall

Metode K-Means hanya bisa mengolah data numerik, untuk itu Kolom yang digunakan adalah kolom 3 sampai 5 yang berisi data numerik yang disimpan dalam variabel `data_num` lalu menampilkan 6 data pertama.

```
data_num = data[, 3:5]
head(data_num)
```

	Age <int>	Annual.Income..k.. <int>	Spending.Score..1.100. <int>
1	19	15	39
2	21	15	81
3	20	16	6
4	23	16	77
5	31	17	40
6	22	17	76

6 rows

Gambar 2.3 Data olah

❖ **Standarisasi data**

Selanjutnya melakukan standarisasi data dikarenakan skala atau range data yang belum sama. Standarisasi data menggunakan fungsi `scale()` dan hasilnya disimpan dalam `data_fix` lalu menampilkan kembali 6 data pertama.

```
data_fix = scale(data_num)
head(data_fix)
```

```
            Age Annual.Income..k.. Spending.Score..1.100.
[1,] -1.4210029          -1.734646             -0.4337131
[2,] -1.2778288          -1.734646              1.1927111
[3,] -1.3494159          -1.696572             -1.7116178
[4,] -1.1346547          -1.696572              1.0378135
[5,] -0.5619583          -1.658498             -0.3949887
[6,] -1.2062418          -1.658498              0.9990891
```

Gambar 2.4 Data standarisasi

❖ **Menentukan jumlah cluster terbaik (K optimal)**

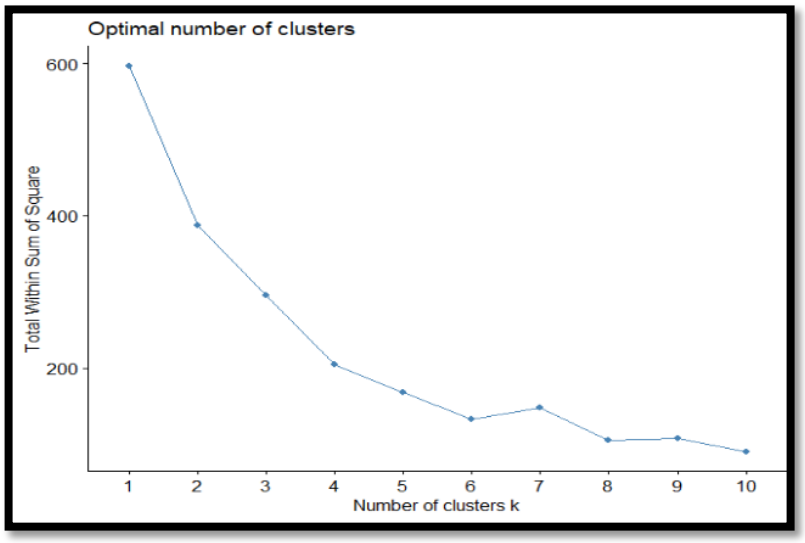
Pengujian model dilakukan untuk mengetahui seberapa dekat relasi antara objek dalam sebuah cluster dan seberapa jauh sebuah cluster terpisah dengan cluster lain. Terdapat beberapa metode yang dapat diterapkan dalam menentukan nilai K optimal.

1. Metode elbow

Elbow criterion adalah suatu modelling criterion yang bisa digunakan untuk menentukan jumlah cluster dengan melihat perubahan antara nilai RMSSTD.Metode elbow menggunakan nilai total wss (whitin sum square) sebagai penentu k optimalnya.

RMSSTD (Root Means Square Standard Deviation) merupakan alat ukur tingkat kemiripan (homogeneity) data yang terdapat di dalam cluster yang ditemukan (within clusters). Makin rendah nilai RMSSTD makin mirip data di dalam cluster yang ditemukan. Penentuan K optimal menggunakan fungsi `fviz_nbclust()` dari package `factoextra` dan menggunakan `method="wss"` untuk menerapkan metode elbow.

```
library(factoextra)
fviz_nbclust(data_fix, kmeans, method="wss")
```



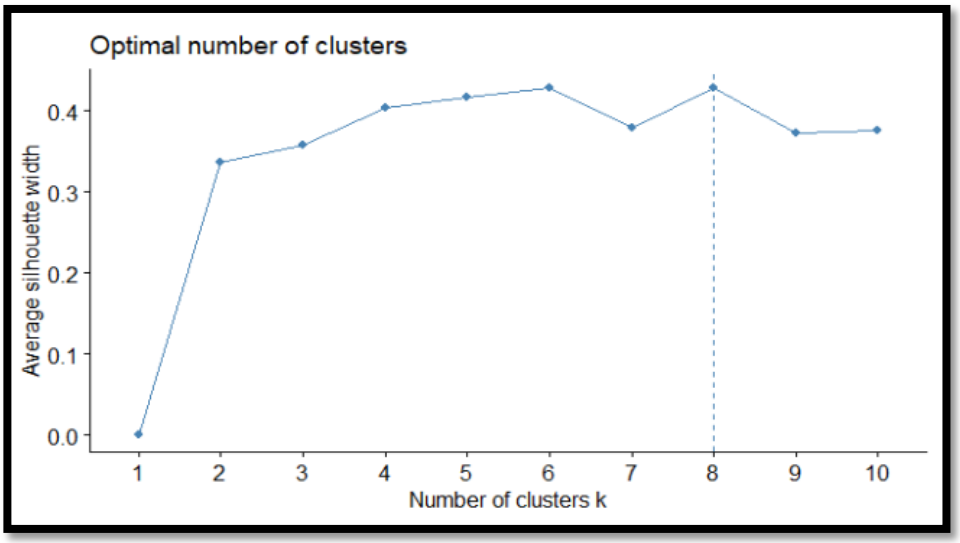
Gambar 2.5 Penentuan K optimal metode elbow

Dari gambar (2.5), garis mengalami patahan yang membentuk elbow atau siku pada saat K = 6. Maka dengan menggunakan metode ini diperoleh K optimal pada saat berada di K = 6. Untuk menjadi pembanding, dilakukan uji yang lainnya.

2. Metode silhouette

Metode silhouette merupakan gabungan dari dua metode yaitu metode cohesion yang berfungsi untuk mengukur seberapa dekat relasi antara objek dalam sebuah cluster, dan metode separation yang berfungsi untuk mengukur seberapa jauh sebuah cluster terpisah dengan cluster lain. Pendekatan rata-rata nilai metode silhoutte untuk menduga kualitas dari klaster yang terbentuk. Semakin tinggi nilai rata-rata nya maka akan semakin baik. Penentuan K optimal menggunakan fungsi `fviz_nbclust()` dari package `factoextra` dan menggunakan `method="silhouette"` untuk menerapkan metode silhouette.

```
fviz_nbclust(data_fix, kmeans, method="silhouette")
```



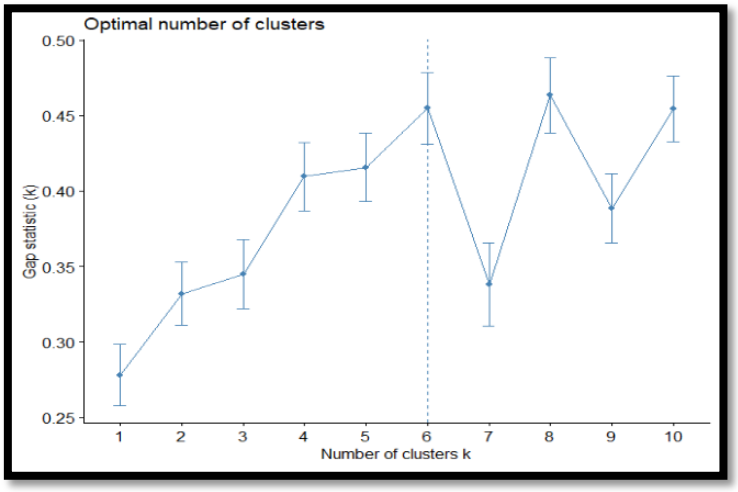
Gambar 2.6 Penentuan K optimal metode silhouette

Berdasarkan hasil keluaran diatas diperoleh banyak klaster optimal yang terbentuk pada K = 8. Sedangkan opsi keduanya pada K = 6. Karena nilai rata-rata silhoutte pada K = 8 dan K = 6 merupakan yang tertinggi dari yang lain.

3. Metode gap statistics

Kriteria banyak cluster optimum diberikan oleh nilai gap statistik (k) yang paling tinggi pada jumlah cluster tertentu. Penentuan K optimal menggunakan fungsi `fviz_nbclust()` dari package `factoextra` dan menggunakan `method="gap_stat"` untuk menerapkan metode gap statistics.

```
fviz_nbclust(data_fix, kmeans, method="gap_stat")
```



Gambar 2.7 Penentuan K optimal metode gap statistics

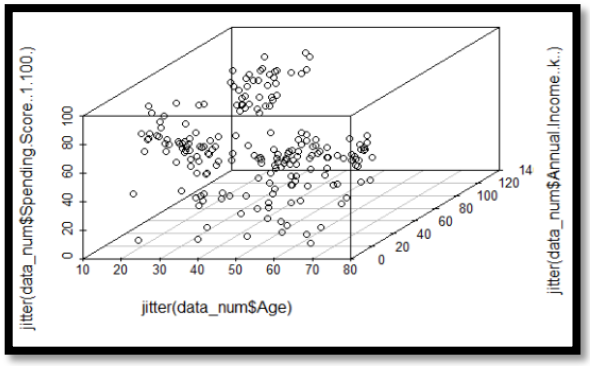
Dari hadil metode gap statistics diperoleh K = 6 sebagai nilai K yang optimal untuk membentuk klaster.

Sehingga jika dibandingkan dengan metode sebelumnya maka dapat ditarik keputusan nilai k yang optimal untuk membentuk klaster adalah 6.

❖ **Plot dan korelasi data**

Kita juga dapat melihat plot data dalam ruang tiga dimensi sesuai jumlah variabel yang kita gunakan sebagai dimensinya. Plot data menggunakan fungsi `scatterplot3d()` dari package `scatterplot3d`.

```
library(scatterplot3d)
scatterplot3d(jitter(data_num$Age),
jitter(data_num$Annual.Income..k..),
jitter(data_num$Spending.Score..1.100.))
```



Gambar 2.8 Plot 3D data

Selanjutnya, kita dapat melihat terlebih dahulu korelasi antar variabel dengan fungsi cor yang terdapat pada package corrplot. Dari output diketahui bahwa tidak terdapat korelasi yang kuat antar variabel, dengan begitu kita dapat menggunakan penghitungan jarak dengan Euclidian Distance karena telah memenuhi asusmsi bahwa peubah-peubah yang diamati tidak berkorelasi dan antar peubah memiliki satuan yang sama. Plot data menggunakan fungsi `cor()` dan `corrplot()` dari package `corrplot`.

```
library(corrplot)
cor_data = cor(data_num)
corrplot(cor_data, method = "color", type = "upper")
```

Gambar 2.9 Plot korelasi data

❖ **Analisis K-Means dengan kluster optimal**

K optimal yang telah didaptkan sebelumnya dapat kita gunakan untuk melakukan analisis cluster K-Means. Baris kedua dapat digunakan untuk memilih metode penghitungan jarak yang diinginkan, diantaranya "euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski", "pearson", "spearman" atau "kendall". Namun baris tersebut tidak kita eksekusi, karena kita akan menggunakan Euclidian Distance yang sudah secara default dipakai tanpa perlu kita atur dalam fungsi `kmeans()`.

```
set.seed(123)
#data_fix = get_dist(data_num, method = "euclidean", stand
= TRUE)
clus_cust = kmeans(data_fix,6)
clus_cust
```

```
K-means clustering with 6 clusters of sizes 13, 57, 22, 22, 40, 46

Cluster means:
         Age Annual.Income..k..
1 -0.6005052          1.1003939
2  1.1950641         -0.4802119
3 -0.9719569         -1.3262173
4  0.6777992          1.0257105
5 -0.4277326          0.9724070
6 -0.7985068         -0.4177864
  Spending.Score..1.100.
1             -1.4256531
2             -0.3216162
3              1.1293439
4             -1.1853182
5              1.2130414
6             -0.2266218

Clustering vector:
  [1] 6 3 6 3 6 3 6 3 2 3 2 3 2 3 6 3
 [17] 6 3 2 3 6 3 2 3 2 3 2 3 6 3 2 3
 [33] 2 3 2 3 2 3 6 3 2 3 2 6 2 3 2 6
 [49] 6 6 2 6 6 2 2 2 2 2 6 2 2 6 2 2
 [65] 2 6 2 2 6 6 2 2 2 2 2 6 2 6 6 2
 [81] 2 6 2 2 6 2 2 6 6 2 2 6 2 6 6 6
 [97] 2 6 2 6 6 2 2 6 2 6 2 2 2 2 2 6
[113] 6 6 6 6 2 2 2 2 6 6 5 5 6 5 4 5
[129] 4 5 4 5 6 5 1 5 4 5 1 5 4 5 6 5
[145] 1 5 4 5 1 5 4 5 4 5 4 5 1 5 1 5
[161] 4 5 1 5 4 5 4 5 1 5 4 5 1 5 4 5
[177] 4 5 4 5 1 5 4 5 4 5 4 5 4 5 1 5
[193] 1 5 4 5 4 5 1 5

Within cluster sum of squares by cluster:
[1] 10.965875 55.486846  8.191823
[4] 14.374768 23.915440 41.345810
 (between_SS / total_SS =  74.2 %)

Available components:

[1] "cluster"      "centers"
[3] "totss"        "withinss"
[5] "tot.withinss" "betweenss"
[7] "size"         "iter"
[9] "ifault"
```

Gambar 2.10 Analisis K-Means

Dari output dapat diketahui bahwa data customer telah dibagi ke dalam 6 kategori kelompok di mana masing-masing kelompok berujmlah 13, 57, 22, 22, 40, dan 46 serta inf0rmasi-informasi tambahan lainnya. Berikut kita lihat hasil clustering dari 200 customer ke dalam 6 kategori customer dengan metode K-Means dalam 6 data pertama.

```
data["Kategori"]=clus_cust$cluster
head(data)
```

	CustomerID <int>	Gender <chr>	Age <int>	Annual.Income..k.. <int>	Spending.Score..1.100. <int>	Kategori <int>
1	1	Male	19	15	39	6
2	2	Male	21	15	81	3
3	3	Female	20	16	6	6
4	4	Female	23	16	77	3
5	5	Female	31	17	40	6
6	6	Female	22	17	76	3

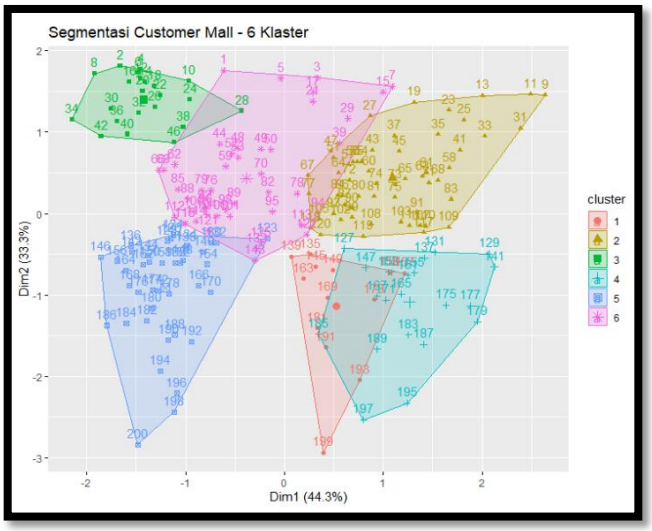
6 rows

Gambar 2.11 Hasil klastering 6 data pertama

❖ **Visualisasi area pengelompokkan**

Visualisasi area dari hasil 6 kategori tersebut juga bisa kita lihat pada gambar (2.12).

```
fviz_cluster(clus_cust, geom=c("text", "point"), data=data_num) + ggtitle("Segmentasi Customer Mall - 6 Klaster")
```



Gambar 2.12 Visualisasi klastering

❖ **Centroid klaster**

Untuk merepresentasikan karakteristik tiap klaster kita dapat menggunakan acuan nilai means masing-masing variabel pada tiap kelompok yang terbentuk.

```
aggregate(data_num, by=list(clus_cust$cluster),FUN=mean)
```

Group.1 <int>	Age <dbl>	Annual.Income..k.. <dbl>	Spending.Score..1.100. <dbl>
1	30.46154	89.46154	13.38462
2	55.54386	47.94737	41.89474
3	25.27273	25.72727	79.36364
4	48.31818	87.50000	19.59091
5	32.87500	86.10000	81.52500
6	27.69565	49.58696	44.34783

6 rows

Gambar 2.13 Centroid cluster

Didapatkan hasil segmentasi sebagai berikut:

1. Kategori 1 memiliki karakteristik penghasilan yang besar namun dengan pengeluaran yang sangat kecil.
2. Kategori 2 merupakan kumpulan customer yang sudah berumur dengan penghasilan dan pengeluaran yang standar.
3. Kategori 3 merupakan customer dengan kelompok umur muda yang memiliki penghsilan rendah namun pengeluaran yang cukup tinggi.
4. Kategori 4 memiliki karakteristik yang cukup mirip seperti kategori 1 dalam hal penghasilan dan pengeluaran, hanya saja dalam kelompok umur yang lebih tua.
5. Kategori 5 merupakan kelompok customer yang memiliki penghasilan tinggi dan diimbangi dengan pengeluaran yang tinggi pula.
6. Kategori 6 memiliki karakteristik yang cukup mirip seperti kategori 2 dalam hal penghasilan dan pengeluaran, namun dalam kelompok umur yang lebih muda.

F. Referensi Video Ekologi Populasi

Untuk memahami lebih lanjut mengenai materi tentang clustering non-hirarki K-means, disajikan video yang dapat memberikan lebih banyak wawasan terkait materi tersebut melalui video berikut:

- Judul : Belajar Data Science: Clustering with k-means
- Link Sumber : *https://www.youtube.com/watch?v=6QV4vPpDxKQ*
- Ulasan :

Dalam video ini menjelaskan bagaimana caranya melakukan clustering dengan K-means dengan analogi yang mudah untuk dimengerti. Dengan kasus sederhana yaitu pemetaan customer dalam 1 dimensi, yaitu usia, ditampilkan bagaimana algoritma K-means dalam menyelesaikan masalah tersebut ditambah visualisasi yang mudah dipahami.

DAFTAR PUSTAKA

Husain A, Ahmad. (2018). K-Means Clustering. https://rstudio-pubs-static.s3.amazonaws.com/355692_27b22c3b8ca34570854372280eb86a87.html

Ketutrare. (2019). *Algoritma K-Means Clustering dan Contoh Soal*. https://www.ketutrare.com/2018/11/algoritma-k-means-clustering-dan-contoh.html

Fathia, A. N., Rahmawati, R. and Tarno (2016) 'Analisis Klaster Kecamatan Di Kabupaten Semarang Berdasarkan Potensi Desa Menggunakan Metode Ward Dan Single Linkage', Jurnal Gaussian, 5(4), pp. 801–810.

Sitepu, R., Irmeilyana, I. and Gultom, B. (2011) 'Analisis Cluster terhadap Tingkat Pencemaran Udara pada Sektor Industri di Sumatera Selatan', Jurnal Penelitian Sains, 14(3), p. 168311.

Yulianto, S. and Hidayatullah, K. H. (2016) 'Analisis Klaster Untuk Pengelompokan Kabupaten/Kota Di Provinsi Jawa Tengah Berdasarkan Indikator Kesejahteraan Rakyat', Statistika, 2(1), pp. 56–63. Available at: https://jurnal.unimus.ac.id/index.php/statistik/article/view/1115.

Choudhary, Vijay. (2018). Data Segmentasi Pelanggan Mall. https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

LAMPIRAN

Syntax Penerapan K-Means Menggunakan RSTUDIO:

```
# Import dataset dan preprocess data
library(readr)
data = read.csv("Mall_Customers.csv")
head(data)
data_num = data[, 3:5]
head(data_num)
# Standarisasi data
data_fix = scale(data_num)
head(data_fix)
# Menentukan K optimal
# 1. Metode elbow
library(factoextra)
fviz_nbclust(data_fix, kmeans, method="wss")
#  2. Metode silhouette
fviz_nbclust(data_fix, kmeans, method="silhouette")
# 3. Metode gap statistics
fviz_nbclust(data_fix, kmeans, method="gap_stat")
# Plot data
library(scatterplot3d)
scatterplot3d(jitter(data_num$Age),
jitter(data_num$Annual.Income..k..),
jitter(data_num$Spending.Score..1.100.))
# Plot korelasi data
library(corrplot)
cor_data = cor(data_num)
corrplot(cor_data, method = "color", type = "upper")
# Analisis K-Means dengan kluster optimal
set.seed(123)
#data_fix = get_dist(data_num, method = "euclidean", stand
= TRUE)
```

```
clus_cust = kmeans(data_fix,6)
clus_cust
data["Kategori"]=clus_cust$cluster
head(data)
# Visualisasi area pengelompokkan
fviz_cluster(clus_cust,      geom=c("text",      "point"),
data=data_num)  +  ggtitle("Segmentasi  Customer  Mall  -  6
Klaster")
# Centroid klaster
aggregate(data_num, by=list(clus_cust$cluster),FUN=mean)
```