Pelari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format ms word dan dilengkapi dengan gambar dan keterangan yang jelas untuk tiap langkah-langkahnya. Sertkan sumber sumber rujukan yang anda gunakan. Kumpulkan sebelum waktu batas akhir pengumpulan tugas.

**GATOT ARIFIANTO**
**NIM : 202420029**

# TUGAS 05
# ADVANCED DATABASE

Pelari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format ms word dan dilengkapi dengan gambar dan keterangan yang jelas untuk tiap langkah-langkahnya. Sertkan sumber sumber rujukan yang anda gunakan. Kumpulkan sebelum waktu batas akhir pengumpulan tugas.

**JAWAB :**
**PENGELOMPOKAN PERFORMA AKADEMIK MAHASISWA BERDASARKAN IP MENGGUNAKAN K-MEANS CLUSTERING.**

| No | Jurusan | NIM | Angkatan | IP1 | IP2 | IP3 | IP4 |
|---|---|---|---|---|---|---|---|
| 1 | SK | 09091001004 | 2009 | 3,21 | 3,5 | 2,96 | 3,32 |
| 2 | SK | 09091001005 | 2009 | 3,26 | 3,67 | 3,21 | 3,17 |
| 3 | SK | 09091001006 | 2009 | 3,11 | 3,62 | 3,04 | 3,05 |
| 4 | IF(BILINGUAL) | 09101402003 | 2010 | 3,63 | 3,42 | 3,58 | 3,48 |
| 5 | IF | 09101002068 | 2010 | 3,84 | 3,7 | 3,35 | 3,26 |
| 6 | SI | 09091003034 | 2009 | 3,47 | 3,71 | 3,62 | 2,89 |
| 7 | SI | 09111003036 | 2011 | 3,29 | 3,43 | 3,33 | 3,46 |
| 8 | SI (BILINGUAL) | 09111403015 | 2011 | 3,52 | 3,17 | 3,08 | 2,88 |
| 9 | IF | 09101002076 | 2010 | 3,11 | 3,05 | 3,2 | 2,37 |
| 10 | SI (BILINGUAL) | 09111403040 | 2011 | 2,95 | 3,1 | 2,48 | 2,28 |

pengelompokkan performa akademik siswa dilakukan dengan menggunakan Algoritma K-Means. Langkah-langkah dari algoritma K-Means adalah sebagai berikut :

1. Tentukan K sebagai jumlah *cluster* yang diinginkan.

2. Bangkitkan K *centroid* (titik pusat) awal secara acak.
3. Hitung masing-masing jarak setiap data ke masing-masing *centroid* menggunakan formula *Euclidean Distance*

$$d(x_j, c_j) = \sqrt{\sum_{j=1}^{n}(x_j - c_j)^2}$$

dengan:
d = jarak
c = *centroid*
x = data
j = nomor data (1…banyak data)

4. Tentukan posisi *centroid* baru dengan cara menghitung nilai rata-rata dari data yang berada pada *centroid* yang sama.

5. Hitung jarak setiap data dengan *centroid* yang baru menggunakan persamaan (1).

6. Apabila terjadi perubahan pada data *centroid* awal dan *centroid* yang baru maka kembali ke langkah 4. Apabila tidak terjadi perubahan pada data yang terletak dalam satu *centroid* maka algoritma berhenti.

*centroid* awal diambil secara acak di tiap *cluster* IP. Pembangkitan nilai acak pada *centroid* awal menggunakan metode pseudo random.Contoh nilai *centroid* acak yang diambil.

*Centroid* Awal pada pengulangan ke - 0

| C | a | b | c | d |
|---|---|---|---|---|
| c1 | 3,04 | 3,11 | 3,27 | 2,43 |
| c2 | 3,07 | 3,15 | 3,09 | 2,39 |
| c3 | 3,17 | 3,17 | 2,68 | 2,83 |
| c4 | 3,1 | 3,12 | 3,6 | 2,87 |

Keterangan:

c1 :*cluster* 1 a : nilai ip1 *cluster*
c2 :*cluster* 2 b : nilai ip2 *cluster*
c3 :*cluster* 3 c : nilai ip3 *cluster*
c4 :*cluster* 4 d : nilai ip4 *cluster*

selanjutnya adalah menghitung jarak masing – masing data terhadap *centroid* dengan menggunakan persamaan (1), jarak data 1 dengan tiap *cluster* adalah :

$$d(x_1, c_1) = \sqrt{(IP_1 - C1_a)^2 + (IP_2 - C1_b)^2 + (IP_3 - C1_c)^2 + (IP_4 - C1_d)^2}$$
$$= \sqrt{(3{,}21 - 3{,}04)^2 + (3{,}5 - 3{,}11)^2 + (2{,}96 - 3{,}27)^2 + (3{,}32 - 2{,}43)^2}$$
$$= 1{,}0340$$

$$d(x_1, c_2) = \sqrt{(IP_1 - C2_a)^2 + (IP_2 - C2_b)^2 + (IP_3 - C2_c)^2 + (IP_4 - C2_d)^2}$$
$$= \sqrt{(3{,}21 - 3{,}07)^2 + (3{,}5 - 3{,}15)^2 + (2{,}96 - 3{,}09)^2 + (3{,}32 - 2{,}39)^2}$$
$$= 1{,}0118$$

$$d(x_1, c_3) = \sqrt{(IP_1 - C3_a)^2 + (IP_2 - C3_b)^2 + (IP_3 - C3_c)^2 + (IP_4 - C3_d)^2}$$
$$= \sqrt{(3{,}21 - 3{,}17)^2 + (3{,}5 - 3{,}17)^2 + (2{,}96 - 2{,}68)^2 + (3{,}32 - 2{,}83)^2}$$
$$= 0{,}6549$$

$$d(x_1, c_4) = \sqrt{(IP_1 - C4_a)^2 + (IP_2 - C4_b)^2 + (IP_3 - C4_c)^2 + (IP_4 - C4_d)^2}$$
$$= \sqrt{(3{,}21 - 3{,}1)^2 + (3{,}5 - 3{,}12)^2 + (2{,}96 - 3{,}6)^2 + (3{,}32 - 2{,}87)^2}$$
$$= 0{,}8766$$

Berdasarkan perhitungan di atas data 1 masuk ke dalam *cluster* 3, karena data 1 memiliki jarak terpendek terhadap *centroid* pada *cluster* 3. Untuk hasil perhitungan selanjutnya dapat dilihat pada Tabel dibawah ini :

Hasil Perhitungan Jarak dan Pengelompokkan Data

| Data | dc1 | dc2 | dc3 | dc4 |
|---|---|---|---|---|
| 1 | 1,034021 | 1,011879 | 0,654981 | 0,876698 |
| 2 | 0,955615 | 0,964002 | 0,809074 | 0,755116 |
| 3 | 0,838033 | 0,812773 | 0,619758 | 0,772075 |
| 4 | 1,281718 | 1,347108 | 1,227436 | 0,862206 |
| 5 | 1,297459 | 1,31145 | 1,167733 | 1,048141 |
| 6 | 0,93755 | 1,002247 | 1,12641 | 0,696994 |
| 7 | 1,108783 | 1,152953 | 0,949421 | 0,743774 |
| 8 | 0,687459 | 0,665658 | 0,533854 | 0,670373 |
| 9 | 0,130384 | 0,155242 | 0,707107 | 0,644205 |
| 10 | 0,809197 | 0,633325 | 0,629126 | 1,274912 |

kolom data yang diwarnai menunjukan bahwa data tersebut memiliki jarak terpendek terhadap masing-masing *cluster*, sehingga data tersebut dikelompokan sesuai warna *cluster*-nya. Pada hasil perhitungan di atas diperoleh 1 data untuk *cluster* 1, 4 data untuk *cluster* 3, 5 data untuk *cluster* 4 dan tidak ada data yang masuk ke *cluster* 2.

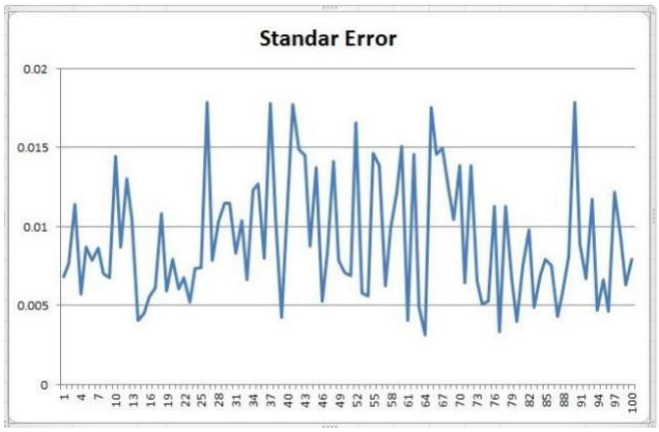
Pada perulangan ke empat diperoleh hasil akhir yaitu 1 data 6 untuk *cluster* 1, 1 data untuk *cluster* 2, 4 data untuk *cluster* 3 dan 4 data untuk *cluster* 4. Hasil akhir pada perulangan ke empat dapat dilihat pada Tabel dibawah ini :

Hasil Akhir Pengelompokkan

| No | dc1 | dc2 | dc3 | dc4 |
|---|---|---|---|---|
| 1 | 1,082867 | 1,240806 | 0,253377 | 0,6179 |
| 2 | 1,02323 | 1,321363 | 0,238747 | 0,416893 |
| 3 | 0,90161 | 1,096586 | 0,214243 | 0,656125 |
| 4 | 1,33559 | 1,792986 | 0,734166 | 0,286705 |
| 5 | 1,330413 | 1,69393 | 0,68775 | 0,343802 |
| 6 | 1,005982 | 1,521249 | 0,659545 | 0,442493 |
| 7 | 1,1755 | 1,52951 | 0,448553 | 0,374433 |
| 8 | 0,676018 | 1,024597 | 0,461952 | 0,676166 |
| 9 | 0 | 0,744715 | 0,876926 | 1,156114 |
| 10 | 0,744715 | 0 | 1,129159 | 1,591163 |

Sebanyak 100 sampel data mahasiswa S2 Universitas Bina Darma telah dikumpulkan sebagai data uji yang akan dikelompokkan. Data yang digunakan sebagai parameter performa akademik adalah data Indeks Prestasi Mahasiswa semester 1 sampai dengan semester 4.Data tersebut kemudian dikelompokkan sebanyak 50 kali.

rata-rata standard deviasi dari masing-masing data adalah sebesar 0.064926.Grafik nilai standard error dari 100 data dapat dilihat pada Gambar dibawah ini :




Berdasarkan pengujian, standard error maksimum dari 100 data adalah sebesar 0.017842 dan minimum sebesar 0.003124.Sedangkan, rata-rata standard error dari masing-masing data adalah sebesar 0.009182.

Hasil standard error dari 100 kali pengelompokkan data adalah **0.009182** atau **0.9182%.** Hal ini menunjukkan bahwa pengelompokkan yang dihasilkan oleh proses K-Means *clustering* tidak banyak mengalami perubahan. Hal ini sejalan dengan kebutuhan yang menginginkan hasil pengelompokkan yang konsisten.

**KESIMPULAN :**

1. Metode K-Means *Clustering* dapat digunakan untuk melakukan pengelompokkan **performa** akademik mahasiswa.

2. Standard Error rata rata dari data yang diuji adalah sebesar **0.92%** yang menunjukkan bahwa metode K–Means *Clustering* dapat melakukan **pengelompokkan performa akademik mahasiswa dengan baik.**

3. *Cluster* akhir dari suatu data berubah – ubah **sesuai dengan jarak terdekat data** terhadap *centroid*.

## SUMBER :

http://www.seminar.ilkom.unsri.ac.id/index.php/kntia/article/download/1146/611

SELESAI

Nama : Hairun Anisyah

NIM : 202420039

Kelas : MTI 23

**TUGAS 05:** Pelari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format ms word dan dilengkapi dengan gambar dan keterangan yang jelas untuk tiap langkah-langkahnya. Sertkan sumber sumber rujukan yang anda gunakan. Kumpulkan sebelum waktu batas akhir pengumpulan tugas.

MENGUKUR PERFORMA CLUSTERING MENGGUNAKAN V-MEASURE

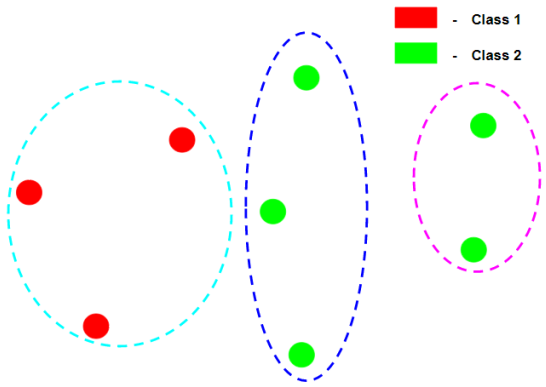
Kalkulasi v-measure dikembangkan berdasarkan dua pendekatan

1. Homogeneity
2. Completeness

Berikut penjelasan mengenai dua pendekatan tersebut.

1. Homogenitas

Homogenetas menjelaskan seberapa dekat algoritma dapat mengelompokkan data berdasarkan data-point dengan label pada kelas yang sama. Dalam kasus ketika jumlah pengelompokkan sama dengan data poin dan masing-masing poin memiliki satu cluster. Pada kasus yang ekstrem homogenitas lebih tinggi dibandingkan completeness berada pada minimum. Konsep dasar dalam homogeneity dimana cluster terkait memiliki tiap data poin di label kelas yang sama.

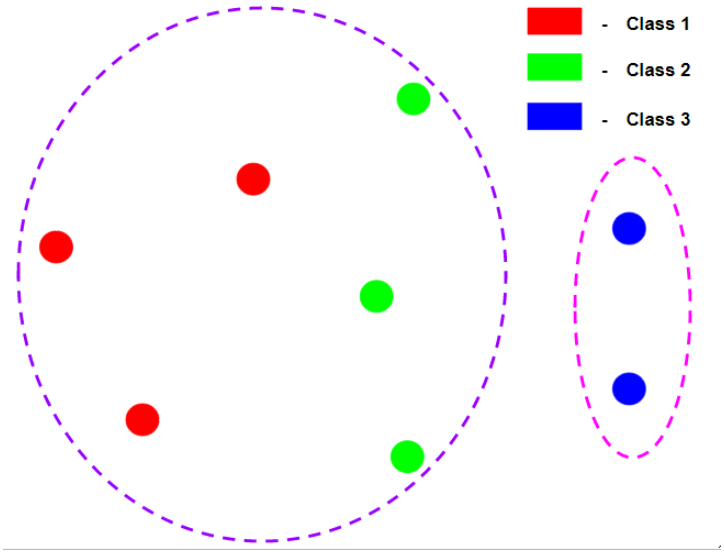



Pada gambar diatas, di tiap cluster memiliki data poin yang sama untuk tiap label kelas namun belum sempurna karena data poin yang sama tidak berada dalam cluster yang sama.

2. Completeness

Pengelompokkan sempurna ketika suatu data poin terkumpul dalam kelas yang sama pada satu cluster. Dijelaskan dengan seberapa dekat algoritma mampu mngelompokan data. Dalam beberapa kasus ketika data poin dikelompokkan ke satu cluster. Menjadi homogenitas minimum dan

completeness menjadi maksimum. Konsep dasar nya adalah label kelas terkait yang melakukan cek apakah data poin di tiap kelas label berada pada cluster yang sama.




Pada gambar diatas merupakan tipe completeness karena semua data poin berada pada label kelas yang sama berdasarkan cluster yang sama namun tidak homogeny dalam satu cluster.

**Rumus perhitungan V-Measure**

Jika N : data sampel

C : label kelas

K : cluster

$a_{ck}$ : jumlah data poin pada kelas C dan cluster K

Maka homogen nya berupa

$$h = 1 - \frac{H(C,K)}{H(C)}$$

Dimana

$$H(C,K) = -\sum_{k=1}^{K}\sum_{c=1}^{C}\frac{a_{ck}}{N}log(\frac{a_{ck}}{\sum_{c=1}^{C}a_{ck}})$$

dan

$$H(C) = -\sum_{c=1}^{C}\frac{\sum_{k=1}^{K}a_{ck}}{C}log(\frac{\sum_{k=1}^{K}a_{ck}}{C})$$

Semnetara completeness berupa:

$$c = 1 - \frac{H(K,C)}{H(K)}$$

Dimana

$$H(K,C) = -\sum_{c=1}^{C}\sum_{k=1}^{K}\frac{a_{ck}}{N}log(\frac{a_{ck}}{\sum_{k=1}^{K}a_{ck}})$$

Dan

$$H(K) = -\sum_{k=1}^{K}\frac{\sum_{c=1}^{C}a_{ck}}{C}log(\frac{\sum_{c=1}^{C}a_{ck}}{C})$$

Untuk persamaan v-measure

$$V_\beta = \frac{(1+\beta)hc}{\beta h+c}$$

Faktor β ditentukan untuk homogeny atau completeness pada algoritma clustering.

**Kelebihan V-measure dibandingkan algoritma lainnya**

Metriks evaluasi ini tidak bergantung pada jumlah label kelas, jumlah cluster dan ukuran data dan algoritma yang digunakan sehingga menjadi metriks yang dapat diandalkan.

**Langkah-langkah**

1. Import data dan libraries, disini bisa menggunakan library Sklearn
2. Muat data dan bersihkan data agar bisa terbaca oleh library
3. Buat model cluster yang berbeda untuk membandingkan nilai V-Measure
4. Visual kan hasil perhitungan untuk melihat performa

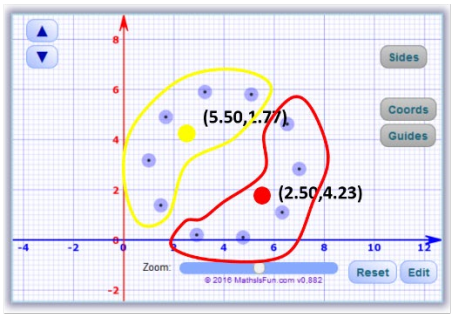
**Referensi**

**ML | V-Measure for Evaluating for Evaluating Clustering Performance** diakses 29 oktober 2020 (https://www.geeksforgeeks.org/ml-v-measure-for-evaluating-clustering-performance/)

## Pengukuran cluster menggukan :

1. **K-Means Clustering Algorithm (https://ekojunaidisalam.com/2017/02/09/k-means-clustering-algorithm/)**

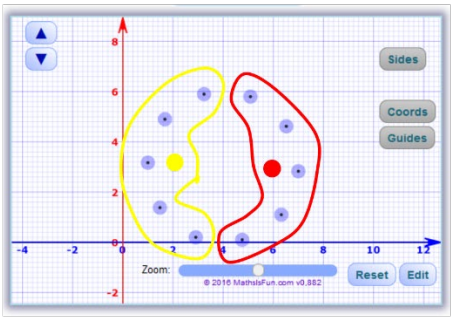
Centroid Awal
(1.48,1.38), (4.76,0.1)




(5.09,5.80), (3.24,5.90), (1.68,4.90), (1.00,3.17), (1.48,1.38), (2.91,0.20), (4.76,0.10), (6.32,1.10), (7.00,2.83), (6.52,4.62), **(2.50,4.23), (5.50,1.77)**

Centroid Akhir
**(2.50,4.23), (5.50,1.77)**

Centroid Awal
(5.09,5.8), (3.24,5.9)




(5.09,5.80), (3.24,5.90), (1.68,4.90), (1.00,3.17), (1.48,1.38), (2.91,0.20), (4.76,0.10), (6.32,1.10), (7.00,2.83), (6.52,4.62), **(5.94,2.89), (2.06,3.11)**

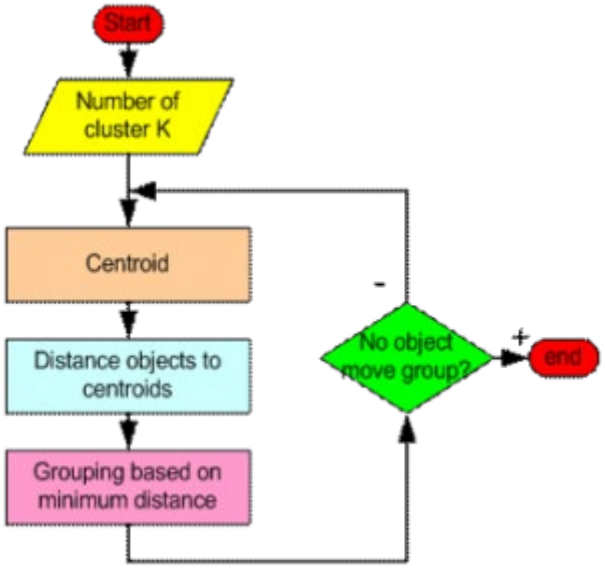
Centroid Akhir
**(5.94,2.89), (2.06,3.11)**

# K-Means Clustering

K-Means Clustering adalah salah satu algoritma dalam menentukan klasifikasi terhadap objek berdasarkan attribut / fitur dari objek tersebut kedalam **K** kluster/partisi. K adalah angka positif yang menyatakan jumlah grup/kluster/partisi terhadap objek. Pemartisian data dilakukan dengan mencari nilai jarak minimum antara data dan nilai ***centroid*** yang telah di set baik secara random atau pun dengan ***Initial Set of Centroids,*** kita juga dapat menentukan nilai centroid berdasarkan **K object** yang berurutan.

***Centroid*** adalah nilai rata-rata aritmetik dari sebuah bentuk objek dari seluruh titik dalam objek tersebut. Penerapan K-Means Clustering ini dapat dilakukan dengan prosedur *step by step* berikut :

- Siapkan data training berbentuk vector.
- Set nilai K cluster.
- Set nilai awal centroids.
- Hitung jarak antara data dan centroid menggunakan rumus ***Euclidean Distance***.
- Partisi data berdasarkan nilai minimum.
- Kemudian lakukan iterasi selama partisi data masih bergerak (tidak ada lagi objek yang bergerak ke partisi lain), bila masih maka ke poin 3.
- Bila grup data sekarang sama dengan grup data sebelumnya, maka hentikan iterasi.
- Data telah dipartisi sesuai nilai centroid akhir.




Contoh penerapan K-Means Cluster

| Data | Attribut / fitur | |
|---|---|---|
| | X | Y |
| A | 5,09 | 5,80 |
| B | 3,24 | 5,90 |
| C | 1,68 | 4,90 |
| D | 1,00 | 3,17 |
| E | 1,48 | 1,38 |
| F | 2,91 | 0,20 |
| G | 4,76 | 0,10 |
| H | 6,32 | 1,10 |
| I | 7,00 | 2,83 |
| J | 6,52 | 4,62 |

Akan dilakukan pemartisian data terhadap data diatas sebanyak 2 partisi, maka tahapannya dalah sebagai berikut :

- K = 2, (K = Jumlah Cluster)
- Nilai awal centroid (1.48,1.38) untuk partisi 0, dan centroid (4.76,0.10) untuk partisi 1

$$D(p,c)_n = \sqrt[2]{\sum_{i=0}^{n}(p_i - c_i)^2}$$

- Maka perhitungan Euclidean Distance adalah dimana ***p*** adalah data, ***c*** adalah centroid, n adalah jumlah data, i adalah iterasi.

- Hasil perhitungan jarak minimum A adalah

$$D(p,c)_0 = \sqrt{(5.09 - 1.48)^2 + (5.80 - 1.38)^2} \approx 5.707$$

$$D(p,c)_1 = \sqrt{(5.09 - 4.76)^2 + (5.80 - 0.10)^2} \approx 5.710$$

- Hasil perhitungan jarak minimum B adalah

$$D(p,c)_0 = \sqrt{(3.24 - 1.48)^2 + (5.90 - 1.38)^2} \approx 4.851$$

$$D(p,c)_1 = \sqrt{(3.24 - 4.76)^2 + (5.90 - 0.10)^2} \approx 5.996$$

| Med | X | Y |
|---|---|---|
| A | 5,707 | 5,710 |
| B | 4,851 | 5,996 |
| C | 3,526 | 5,703 |
| D | 1,853 | 4,854 |
| E | 0,000 | 3,521 |
| F | 1,854 | 1,853 |
| G | 3,521 | 0,000 |
| H | 4,848 | 1,853 |
| I | 5,707 | 3,531 |
| J | 5,992 | 4,851 |

- Hasil perhitungan jarak keseluruhan
- Hasil perhitungan partisi diambil dari jarak minimum, sehingga didapat sebagai berikut :

| Med | Cluster 0 | Cluster 1 |
|---|---|---|
| A | 1 | 0 |
| B | 1 | 0 |
| C | 1 | 0 |
| D | 1 | 0 |
| E | 1 | 0 |
| F | 0 | 1 |
| G | 0 | 1 |
| H | 0 | 1 |
| I | 0 | 1 |
| J | 0 | 1 |

Contoh pada A, X = 5.707 lebih kecil dari Y = 5.710, maka A termasuk ke dalam Cluster 0. Begitu juga dengan F, X = 1.854 lebih besar dari Y = 1.853, maka B masuk ke dalam Cluster 1.

- Setelah data di partisi, maka selanjutnya nilai centroid harus dihitung ulang untuk menentukan jarak minimum yang baru, berikut perhitungan centroid baru :

$$c_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{p_j \in S_i^{(t)}} p_j|$$

$$c_0 = \left(\frac{(5.09 + 3.24 + 1.68 + 1.00 + 1.48)}{5}, \frac{(5.80 + 5.90 + 4{,}90 + 3.17 + 1.38)}{5}\right) \approx (2.498, 4.23)$$

$$c_1 = \left(\frac{(2.91 + 4.76 + 6.32 + 7 + 6.52)}{5}, \frac{(0.2 + 0.1 + 1.1 + 2.83 + 4.62)}{5}\right) \approx (5.502, 1.77)$$

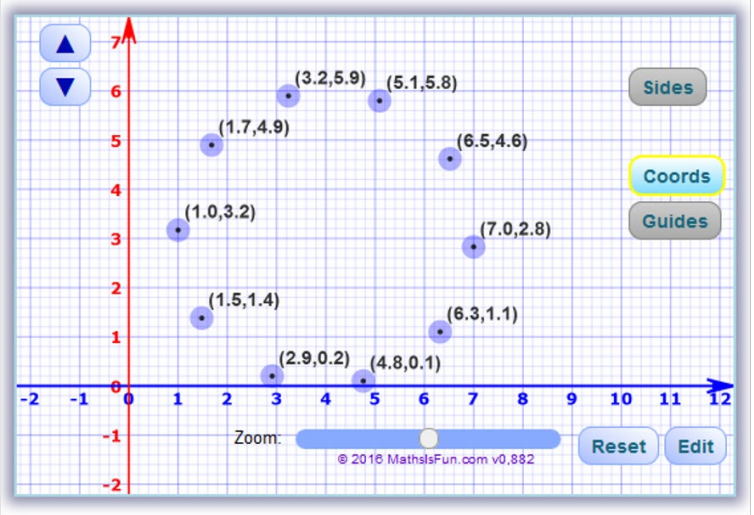
- Hitung jarak minimumnya kembali dengan menggunakan centroid yang baru, sehingga di dapat hasilnya sebagai berikut.

| Med | X | Y |
|---|---|---|
| A | 3,030 | 4,051 |
| B | 1,827 | 4,709 |
| C | 1,057 | 4,940 |
| D | 1,835 | 4,715 |
| E | 3,026 | 4,041 |
| F | 4,051 | 3,030 |
| G | 4,709 | 1,827 |
| H | 4,940 | 1,057 |
| I | 4,715 | 1,835 |
| J | 4,041 | 3,026 |

| Med | Cluster 0 | Cluster 1 |
|---|---|---|
| A | 1 | 0 |
| B | 1 | 0 |
| C | 1 | 0 |
| D | 1 | 0 |
| E | 1 | 0 |
| F | 0 | 1 |
| G | 0 | 1 |
| H | 0 | 1 |
| I | 0 | 1 |
| J | 0 | 1 |

- Klasifikasikan kembali data berdasar jarak minimum diatas.
- Karena tidak ada data yang berpindah ke cluster yang berbeda, sehingga iterasi kita cukupkan sampai dengan nilai centroid akhir : **(2.50,4.23), (5.50,1.77)**

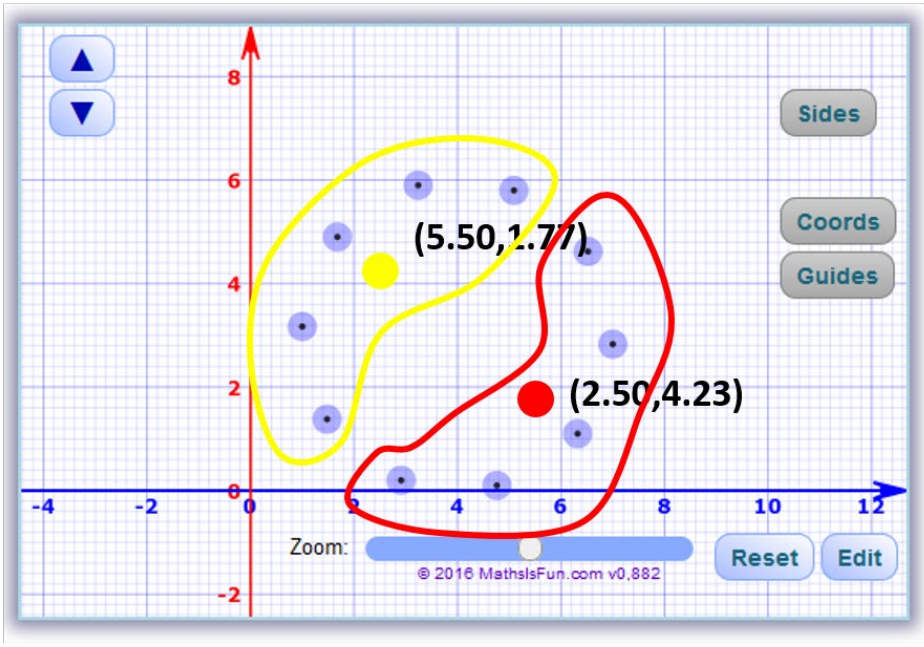
Berikut adalah tampilan kordinat cartesius dari sample data diatas :




(5.09,5.80), (3.24,5.90), (1.68,4.90), (1.00,3.17), (1.48,1.38), (2.91,0.20), (4.76,0.10), (6.32,1.10), (7.00,2.83), (6.52,4.62)

Berikut adalah tampilan setelah penerapan K-Means Clustering dengan centroid akhir:

Centroid Awal
(1.48,1.38), (4.76,0.1)




(5.09,5.80), (3.24,5.90), (1.68,4.90), (1.00,3.17), (1.48,1.38), (2.91,0.20), (4.76,0.10), (6.32,1.10), (7.00,2.83), (6.52,4.62), **(2.50,4.23), (5.50,1.77)**

Centroid Akhir
**(2.50,4.23), (5.50,1.77)**

Aplikasi K-Means Clustering sangat sering digunakan, mulai dari *unsupervised learning of neural network*, *Pattern Recognitions*, *Classifications Analysis*, *Artificial Intelligence*, *Image Processing*, *Computer Vision* dan banyak lainnya.

## 2. Metode Clustering Algoritma Fuzzy CMeans (https://edrianhadinata.wordpress.com/2013/12/19/metode-clustering-algoritma-fuzzy-cmeans/)

Logika *fuzzy* pertama kali dikembangkan oleh Lotfi A. Zadeh, seorang ilmuan Amerika Serikat berkebangsaan Iran dari universitas California di Barkeley, melalui tulisannya pada tahun 1965.( Munir, R. 2005).[5]

*Fuzzy* secara bahasa diartikan sebagai kabur atau samar-samar. Suatu nilai dapat bernilai benar atau salah secara bersamaan. Dalam fuzzy dikenal derajat keanggotaan yang memiliki rentang nilai 0 (nol) hingga 1(satu). Berbeda dengan himpunan tegas yang memiliki nilai 1 atau 0 (ya atau tidak).

*Logika fuzzy* adalah suatu cara yang tepat untuk memetakan suatu ruang *input* kedalam suatu ruang *output*, mempunyai nilai kontinyu. *Fuzzy* dinyatakan dalam derajat dari suatu keanggotaan dan derajat dari kebenaran. Oleh sebab itu sesuatu dapat dikatakan sebagian benar dan sebagian salah pada waktu yang sama (Kusumadewi. 2003). [4]

**Clustering**

Clustering adalah suatu metode pengelompokan berdasarkan ukuran kedekatan(kemiripan).Clustering beda dengan group, kalau group berarti kelompok yang sama,kondisinya kalau tidak ya pasti bukan kelompoknya.Tetapi kalau cluster tidak harus sama akan tetapi pengelompokannya berdasarkan pada kedekatan dari suatu karakteristik sample yang ada

Macam-macam metode clustering :

- Berbasis Metode Statistikk

1. Hirarchical clustering method : pada kasus untuk jumlah kelompok belum ditentukan terlebih dulu, contoh data-data hasil survey kuisioner. Macam-metode jenis ini: Single Lingkage,Complete Linkage,Average Linkage dll.
2. Non Hirarchical clustering method: Jumlah kelompok telah ditentukan terlebih dulu.Metode yang digunakan : K-Means.

- Berbasis Fuzzy : Fuzzy C-Means
- Berbasis Neural Network : Kohonen SOM, LVQ
- Metode lain untuk optimasi centroid atau lebar cluster : Genetik Algoritma (GA)

**Fuzzy C-Means**

Fuzzy clustering adalah proses menentukan derajat keanggotaan, dan kemudian menggunakannya dengan memasukkannya kedalam elemen data kedalam satu kelompok cluster atau lebih.

Hal ini akan memberikan informasi kesamaan dari setiap objek. Satu dari sekian banyaknya algoritma fuzzy clustering yang digunakan adalah algoritma fuzzy clustering c means. Vektor dari fuzzy clustering, $V=\{v_1, v_2, v_3, \ldots, v_c\}$, merupakan sebuah fungsi objektif yang di defenisikan dengan derajat keanggotaan dari data $X_j$ dan pusat cluster $V_j$.

Algoritma fuzzy clustering c means membagi data yang tersedia dari setiap elemen data berhingga lalu memasukkannya kedalam bagian dari koleksi cluster yang dipengaruhi oleh beberapa kriteria yang diberikan. Berikan satu kumpulan data berhingga. $X= \{x_1, \ldots, x_n\}$ dan pusat data.

$$J_m(X,U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c}(\mu_{ij})^m d^2(X_j, V_i) \ldots\ldots\ldots (1)$$

Dimana $\mu_{ij}$ adalah derajat keanggotaan dari $X_j$ dan pusat cluster adalah sebuah bagian dari keanggotaan matriks $[\mu_{ij}]$. $d^2$ adalahakar dari *Euclidean distance* dan m adalah parameter fuzzy yang rata-rata derajat kekaburan dari setiap data derajat keanggotaan tidak lebih besar dari 1,0 Ravichandran (2009).[9]

*Output* dari *Fuzzy C-Means* merupakan deretan pusat *cluster* dan beberapa derajat keanggotaan untuk tiap-tiap titik data. Informasi ini dapat digunakan untuk membangun suatu *fuzzy inference system*.

**Algoritma *Fuzzy Clustering Means* (FCM)**

Algoritma Fuzzy C-Means adalah sebagai berikut:

1. Input data yang akan dicluster $X$, berupa matriks berukuran $n \times m$ ($n$=jumlah sample data, $m$=atribut setiap data). $X_{ij}$=data sample ke-$i$ ($i$=1,2,…,$n$), atribut ke-$j$ ($j$=1,2,…,$m$).
2. Tentukan :
    1. Jumlah cluster = $c$
    2. Pangkat = $w$
    3. Maksimum iterasi = $MaxIter$
    4. Error terkecil yang diharapkan = $\xi$
    5. Fungsi obyektif awal = $Po$ =0
    6. Iterasi awal = $t$ =
3. Bangkitkan nilai acak μik, i=1,2,…,n; k=1,2,…,c sebagai elemen-elemen matriks partisi awal μik. μik adalah derajat keanggotaan yang merujuk pada seberapa besar kemungkinan suatu data bisa menjadi anggota ke dalam suatu cluster.Posisi dan nilai matriks dibangun secara random. Dimana nilai keangotaan terletak pada interval 0 sampai dengan 1. Pada posisi awal matriks partisi U masih belum akurat begitu juga pusat clusternya. Sehingga kecendrungan data untuk masuk suatu cluster juga

belum akurat.

$$Q_i = \sum_{k=1}^{c} \mu_{ik} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

4. Hitung pusat *Cluster* ke-k: $V_{kj}$ ,dengan $k$=1,2,…c dan $j$=1,2,…m. dimana $X_{ij}$adalah variabel fuzzy yang digunakan dan $w$ adalah bobot.

$$V_{kj} = \frac{\sum_{i=1}^{n}\left((\mu_{ik})^{w} * X_{ij}\right)}{\sum_{i=1}^{n}(\mu_{ik})^{w}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Fungsi objektif digunakan sebagai syarat perulangan untuk mendapatkan pusat cluster yang tepat. Sehingga diperoleh kecendrungan data untuk masuk ke *cluster* mana pada *step* akhir.

5. Hitung fungsi obyektif pada iterasi ke-$t$, $P_t$

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right](\mu_{ik})^{w}\right) \quad \ldots\ldots\ldots\ldots (5)$$

6. Perhitungan fungsi objektif $P_t$ dimana nilai variabel fuzzy $X_{ij}$ di kurang dengan dengan pusat cluster $V_{kj}$kemudian hasil pengurangannya di kuadradkan lalu masing-masing hasil kuadrad di jumlahkan untuk dikali dengan kuadrad dari derajat keanggotaan $\mu_{ik}$ untuk tiap *cluster*. Setelah itu jumlahkan semua nilai di semua *cluster* untuk mendapatkan fungsi objektif $P_t$.
7. Hitung perubahan matriks partisi:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}} \quad \ldots\ldots\ldots\ldots\ldots (6)$$

dengan: $i$=1,2,…n dan $k$=1,2,..$c$.

Untuk mencari perubahan matrik partisi $\mu_{ik}$,pengurangan nilai variabel fuzzy $X_{ij}$ di lakukan kembali terhadap pusat cluster $V_{kj}$lalu dikuadradkan. Kemudian dijumlahkan lalu dipangkatkan dengan -1/($w$-1) dengan bobot, $w$=2 hasilnya setiap data dipangkatkan dengan -1. Setelah proses perhitungan dilakukan, normalisasikan semua data derajat keanggotaan baru dengan cara menjumlahkan derajat keanggotaan baru $k$=1,…$c$, hasilnya kemudian dibagi dengan derajat keanggotaan yang baru. Proses ini dilakukan agar derajat keanggotaan yang baru mempunyai rentang antara 0 dan tidak lebih dari 1

8. Cek kondisi berhenti:a) jika:( $|Pt - Pt\text{-}1| < \xi$) atau ($t$>maxIter) maka berhenti.b) jika tidak, $t$=$t$+1, ulangi langkah ke-4.

Algoritma K-Means merupakan salah satu algoritma dengan partitional, karena K-Means didasarkan pada penentuan jumlah awal kelompok dengan mendefinisikan nilai centroid awalnya. Algoritma K-Means menggunakan proses secara berulang-ulang untuk mendapatkan basis data cluster. Dibutuhkan jumlah cluster awal yang diinginkan sebagai masukan dan menghasilkan jumlah cluster akhir sebagai output. Jika algoritma diperlukan untuk menghasilkan cluster K maka akan ada K awal dan K akhir. Metode K-Means akan memilih pola k sebagai titik awal centroid secara acak. Jumlah iterasi untuk mencapai cluster centroid akan dipengaruhi oleh calon cluster centroid awal secara random dimana jika posisi centroid baru tidak berubah. Nilai K yang dipilih menjadi pusat awal, akan dihitung dengan menggunakan rumus Euclidean Distance yaitu mencari jarak terdekat antara titik centroid dengan data/objek. Data yang memiliki jarak pendek atau terdekat dengan centroid akan membentuk sebuah cluster.

Algoritma K-Means

1. Tentukan k sebagai jumlah cluster yang akan dibentuk
2. Tentukan k Centroid (titik pusat cluster) awal secara random/acak.

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \quad ;i = 1,2,3,...n \quad \text{.....(1)}$$

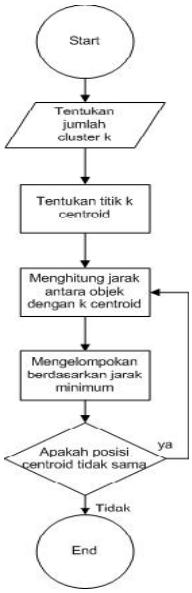
Dimana; v : centroid pada cluster
$x_i$ : objek ke-i
$n$ : banyaknya objek/jumlah objek yang menjadi anggota cluster

3. Hitung jarak setiap objek ke masing-masing centroid dari masing-masing cluster. Untuk menghitung jarak antara objek dengan centroid dapat menggunakan Euclidian Distance

$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad ;i = 1,2,3,...,n \quad \text{.....(2)}$$

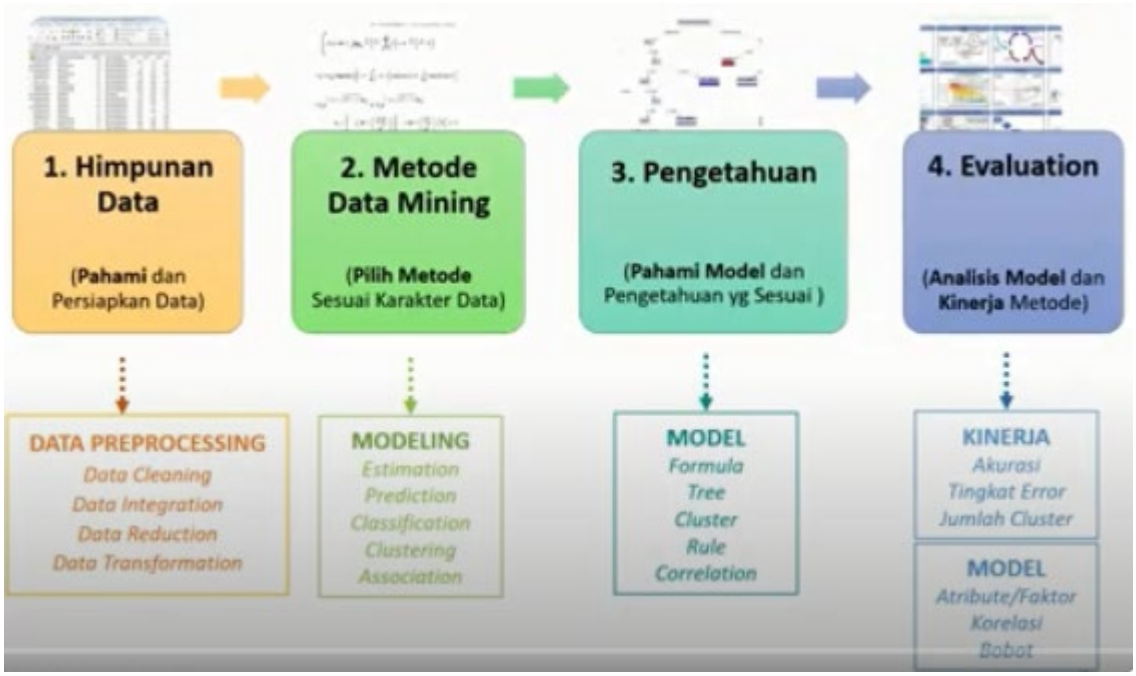
Dimana; $x_i$ : objek x ke-i
$y_i$: daya y ke-i
$n$ : banyaknya objek

4. Alokasikan masing-masing objek ke dalam centroid yang paling dekat
5. Lakukan iterasi, kemudian tentukan posisi centroid baru dengan menggunakan persamaan (1)
6. Ulangi langkah 3 jika posisi centroid baru tidak sama .

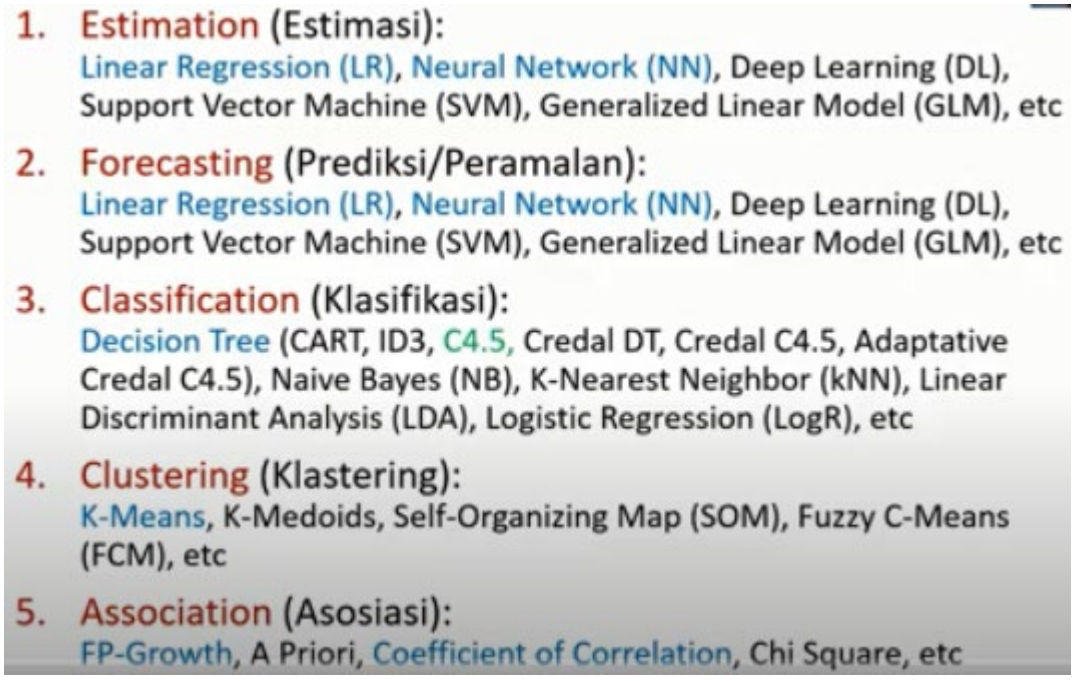

Start
Tentukan jumlah cluster k
Tentukan titik k centroid
Menghitung jarak antara objek dengan k centroid
Mengelompokan berdasarkan jarak minimum
Apakah posisi centroid tidak sama
ya
Tidak
End

NIM : 202420040 - ISTIARSO

**Data Mining Roadmap**

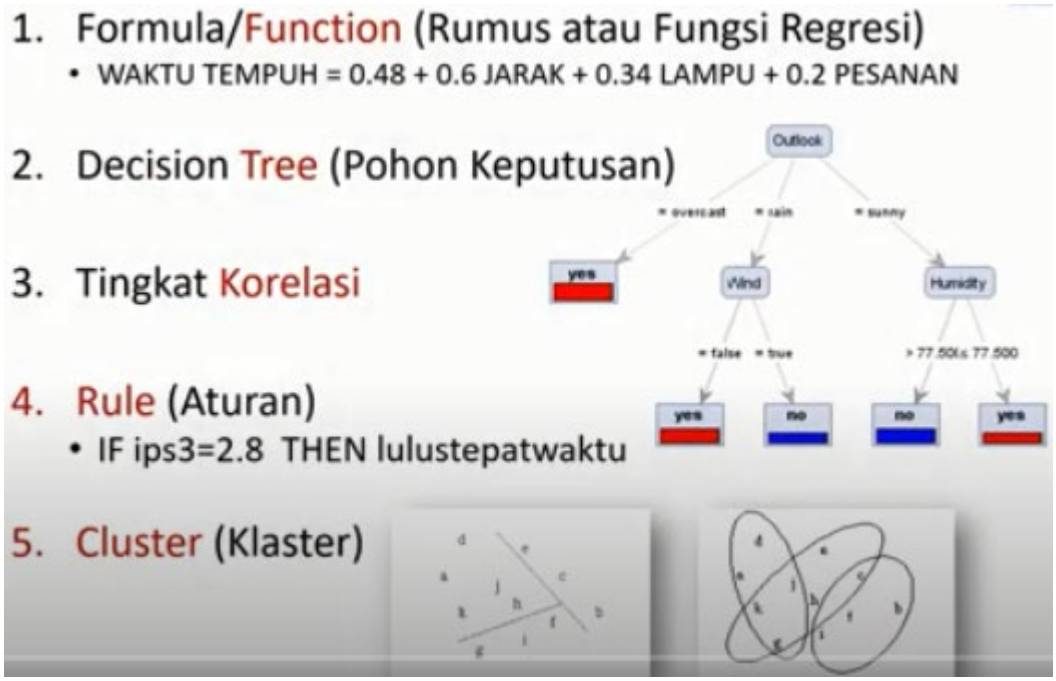



**Metode data Mining**

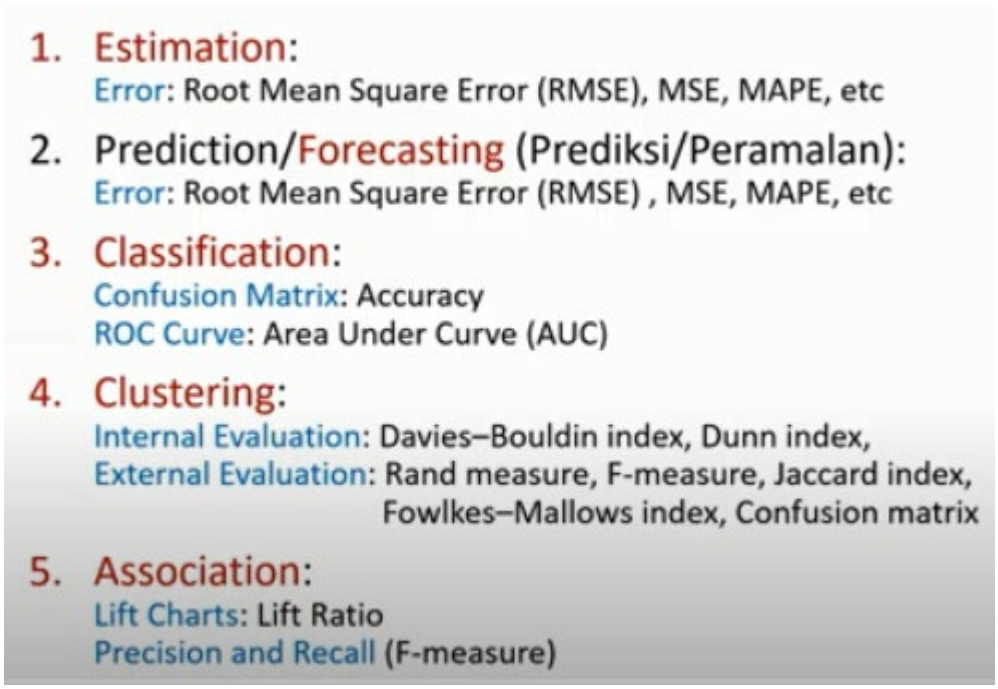


NIM : 202420040 - ISTIARSO

**Model/Knowledge data mining**




**Evaluasi Data Mining**



NIM : 202420040 - ISTIARSO

Salah satu metode pengukuran performance clustering adalah :

# Confusion matrix

juga sering disebut *error matrix*. Pada dasarnya *confusion matrix* memberikan informasi perbandingan hasil klasifikasi yang dilakukan oleh sistem (model) dengan hasil klasifikasi sebenarnya. *Confusion matrix* berbentuk tabel matriks yang menggambarkan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui. Gambar dibawah ini merupakan *confusion matrix* dengan 4 kombinasi nilai prediksi dan nilai aktual yang berbeda.

Terdapat 4 istilah sebagai representasi hasil proses klasifikasi pada *confusion matrix*. Keempat istilah tersebut adalah *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN). Agar lebih mudah memahaminya, saya menggunakan contoh kasus sederhana untuk memprediksi seorang pasien menderita kanker atau tidak.

- ***True Positive* (TP)**
  Merupakan data positif yang diprediksi benar. Contohnya, pasien menderita kanker (*class* 1) dan dari model yang dibuat memprediksi pasien tersebut menderita kanker (*class* 1).

- ***True Negative* (TN)**
  Merupakan data negatif yang diprediksi benar. Contohnya, pasien tidak menderita kanker (*class* 2) dan dari model yang dibuat memprediksi pasien tersebut tidak menderita kanker (*class* 2).

- ***False Postive* (FP) — Type I Error**
  Merupakan data negatif namun diprediksi sebagai data positif. Contohnya, pasien tidak menderita kanker (*class* 2) tetapi dari model yang telah memprediksi pasien tersebut menderita kanker (*class* 1).

- ***False Negative* (FN) — Type II Error**
  Merupakan data positif namun diprediksi sebagai data negatif. Contohnya, pasien menderita kanker (*class* 1) tetapi dari model yang dibuat memprediksi pasien tersebut tidak menderita kanker (*class* 2).

Pada beberapa kasus "Type II Error" lebih berbahaya, kita dapat menghubungkan pernyataan itu dengan contoh prediksi kanker diatas. Jika pasien tidak menderita kanker tetapi diprediksi menderita kanker (FP), maka pada diagnosa selanjutnya pasien tersebut dapat mengetahui keadaan sebenarnya bahwa pasien tersebut benar tidak menderita kanker. Tetapi jika ada pasien yang sebenarnya menderita kanker tetapi diprediksi tidak menderita kanker (FN), maka pasien tersebut akan mengetahui keadaan sebenarnya dengan sangat terlambat dan pasien tersebut tidak segera mengambil tindakan pencegahan medis untuk kanker itu. Sehingga dapat menyebabkan kondisi pasien yang semakin memburuk setiap harinya bahkan kematian. Jadi dapat dikatakan bahwa "Type II Error" lebih berbahaya.

Ada cara yang lebih mudah untuk mengingatnya, yaitu:

- Jika diawali dengan ***True*** maka prediksinya adalah benar, entah diprediksi terjadi atau tidak terjadi.
- Jika diawali dengan ***False*** maka prediksinya adalah salah.
- Positif dan negatif merupakan hasil prediksi dari model.

Tentunya kita ingin model yang telah kita buat memberikan 0 *false positive* dan 0 *false negative*. Tetapi pada prakteknya hal tersebut tidak akan pernah terjadi karena model mana pun tidak akan memberikan keakuratan 100%. Jika model anda memberikan nilai 100% maka ada masalah pada model yang anda buat atau data yang anda gunakan.

## Mengapa kita memerlukan confusion matrix ?

Seperti yang telah dijelaskan diatas, *confusion matrix* akan memberi tahu seberapa baik model yang kita buat. Secara khusus *confusion matrix* juga memberikan informasi tentang TP, FP, TN, dan FN. Hal ini sangat berguna karena hasil dari klasifikasi umumnya tidak dapat diekspresikan dengan baik dalam satu angka saja.

Dengan contoh yang sama untuk memprediksi kanker, anda akan mencoba memprediksi siapa yang akan mati karena kanker tahun ini berdasarkan perilaku seperti merokok dari seluruh populasi. Pada tahun tertentu, hanya 1% populasi yang mati karena kanker. Algoritma klasifikasi naif hanya akan memprediksi tidak ada yang mati karena kanker. Dengan *confusion matrix* memungkinkan kita untuk melihat dengan cepat, dari siapa yang akan diprediksi mati, berapa banyak yang mati dan yang tidak.

Berikut adalah beberapa manfaat dari *confusion matrix*:

1. Menunjukkan bagaimana model ketika membuat prediksi.
2. Tidak hanya memberi informasi tentang kesalahan yang dibuat oleh model tetapi juga jenis kesalahan yang dibuat.
3. Setiap kolom dari *confusion matrix* merepresentasikan *instance* dari kelas prediksi.
4. Setiap baris dari *confusion matrix* mewakili *instance* dari kelas aktual.

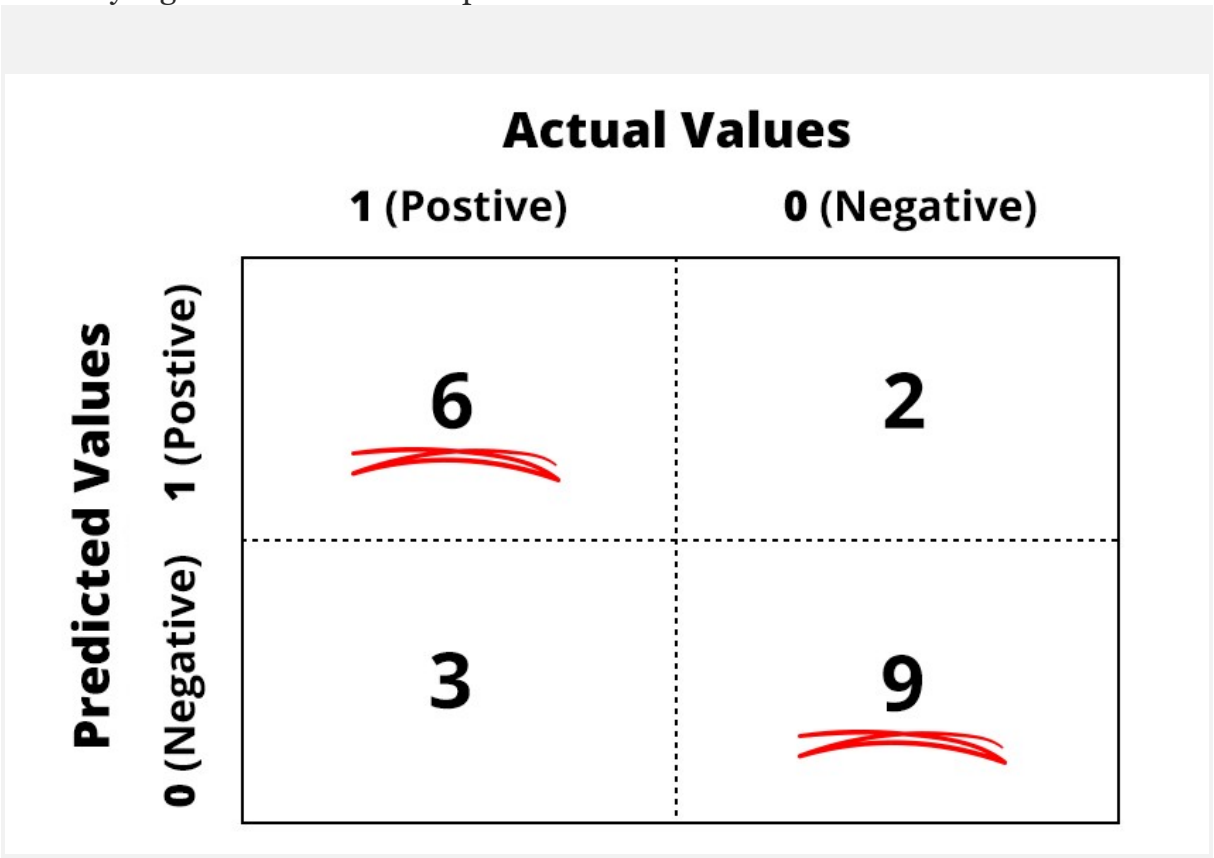
**Contoh confusion matrix pada klasifikasi biner**


Klasifikasi biner akan menghasilkan ouput dengan dua kelas (0 / 1) untuk data input yang diberikan.

*Confusion matrix dapat* digunakan untuk mengukur performa dalam permasalahan klasifikasi biner maupun permasalahan klasifikasi *multiclass*. Klasifikasi biner hanya menghasilkan dua *ouput* kelas (label), seperti “Ya” atau “Tidak”, “0” atau “1” untuk

setiap data *input* yang diberikan. Kelas utama biasanya dinotasikan sebagai data positif dan yang lainnya sebagai data negatif.

Sebagai contoh, sebuah model akan dilatih untuk memprediksi apakah seorang pasien sedang menderita kanker atau tidak. Dengan asumsi terdapat 20 pasien dengan 9 pasien positif kanker dan 11 pasien negatif kanker, maka contoh *confusion matrix* yang dihasilkan model seperti dibawah ini :




Contoh confusion matrix untuk klasifikasi biner

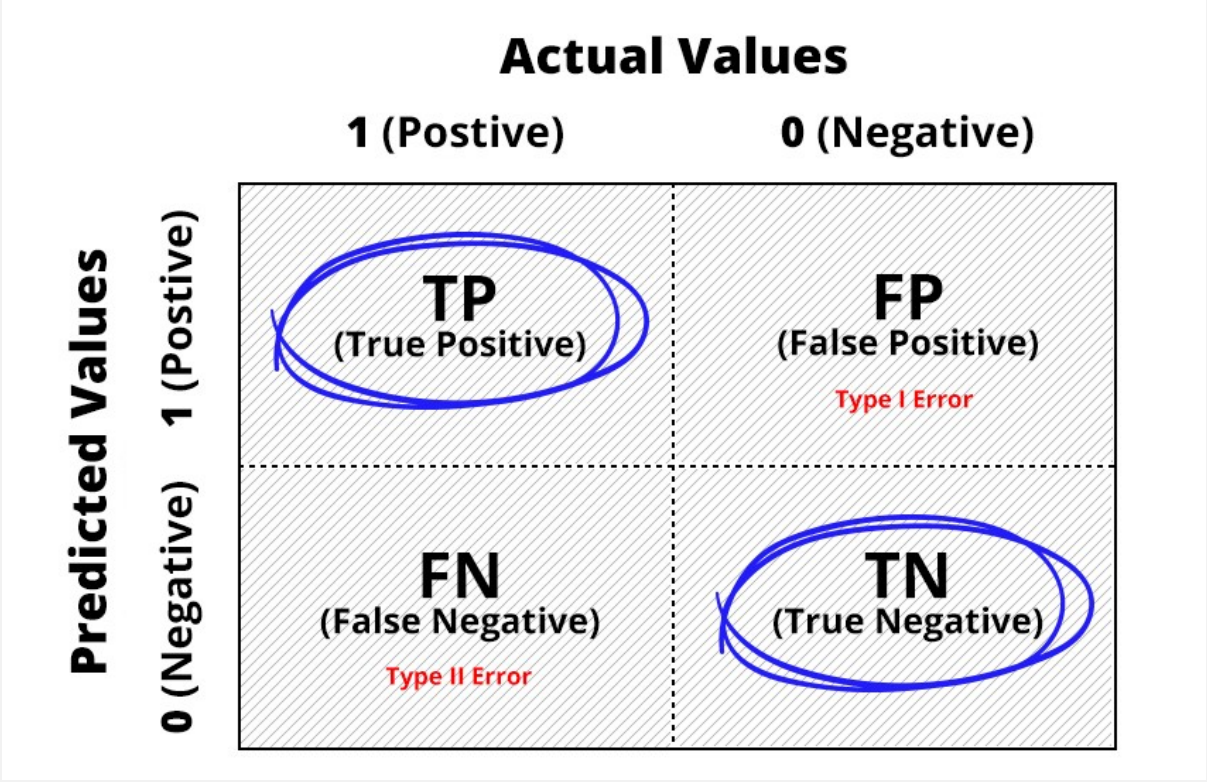
Jika dilihat dari *confusion matrix* diatas dari 9 pasien positif kanker, model memprediksi ada 3 pasien yang diprediksi negatif kanker (FN), dan dari 11 pasien negatif kanker, model memprediksi ada 2 pasien yang diprediksi positif kanker (FP). Prediksi yang benar terletak pada tabel diagonal (garis bawah merah), sehingga secara visual sangat mudah untuk melihat kesalahan prediksi karena kesalahan prediksi berada di luar tabel diagonal *confusion matrix*.

### Bagaimana mengukur performance metrics dari confusion matrix ?

Kita dapat menggunakan c*onfusion matrix* untuk menghitung berbagai *performance metrics* untuk mengukur kinerja model yang telah dibuat. Pada bagian ini mari kita pahami beberapa *performance metrics* populer yang umum dan sering digunakan: *accuracy*, *precission*, dan *recall*.

### Accuracy

*Accuracy* menggambarkan seberapa akurat model dapat mengklasifikasikan dengan benar. Maka, *accuracy* merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Dengan kata lain, *accuracy* merupakan tingkat kedekatan nilai prediksi dengan nilai aktual (sebenarnya). Nilai *accuracy* dapat diperoleh dengan persamaan (1).




Confusion matrix yang menggambarkan nilai accuracy.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Persamaan nilai accuracy.

Dari contoh *confusion matrix* klasifikasi biner diatas maka dengan menghitung nilai *accuracy* dapat menjawab pertanyaan “Berapa persen pasien yang benar diprediksi menderita kanker maupun yang tidak menderita kanker dari kesuluruhan pasien?”

$$accuracy = \frac{jumlah\ pasien\ dipredeksi\ benar\ (kanker + tidak\ kanker)}{jumlah\ pasien\ keseluruhan}$$

$$= \frac{6 + 9}{6 + 9 + 2 + 3} = \frac{15}{20} = 0{,}75$$

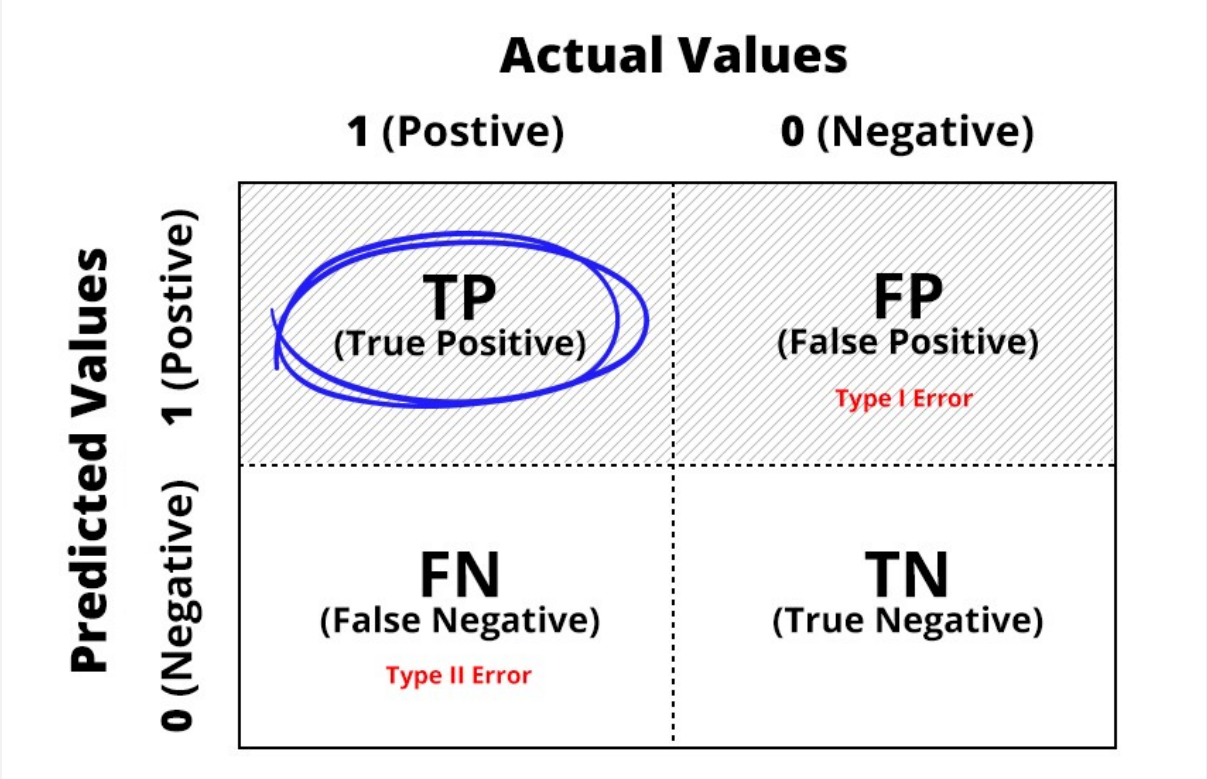
$$= 0{,}75 * 100\%$$

$$= 75\%$$

Menghitung nilai accuracy dari contoh confusion matrix klasifikasi biner.

## Precision **(Positive Predictive Value)**

*Precision* menggambarkan tingkat keakuratan antara data yang diminta dengan hasil prediksi yang diberikan oleh model. Maka, *precision* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positf. Dari semua kelas positif yang telah di prediksi dengan benar, berapa banyak data yang benar-benar positif. Nilai *precision* dapat diperoleh dengan persamaan (2).




Confusion matrix yang menggambarkan nilai precison.

$$precision = \frac{TP}{TP + FP} \quad (2)$$

Persamaan nilai precision.

Dari contoh *confusion matrix* klasifikasi biner diatas maka dengan menghitung nilai *precision* dapat menjawab pertanyaan “Berapa persen pasien yang benar menderita kanker dari keseluruhan pasien yang diprediksi menderita kanker?”

$$precision = \frac{jumlah\ pasien\ kanker\ diprediksi\ \mathbf{benar}}{jumlah\ pasien\ diprediksi\ kanker}$$

$$= \frac{6}{6+2} = \frac{6}{8} = 0{,}75$$

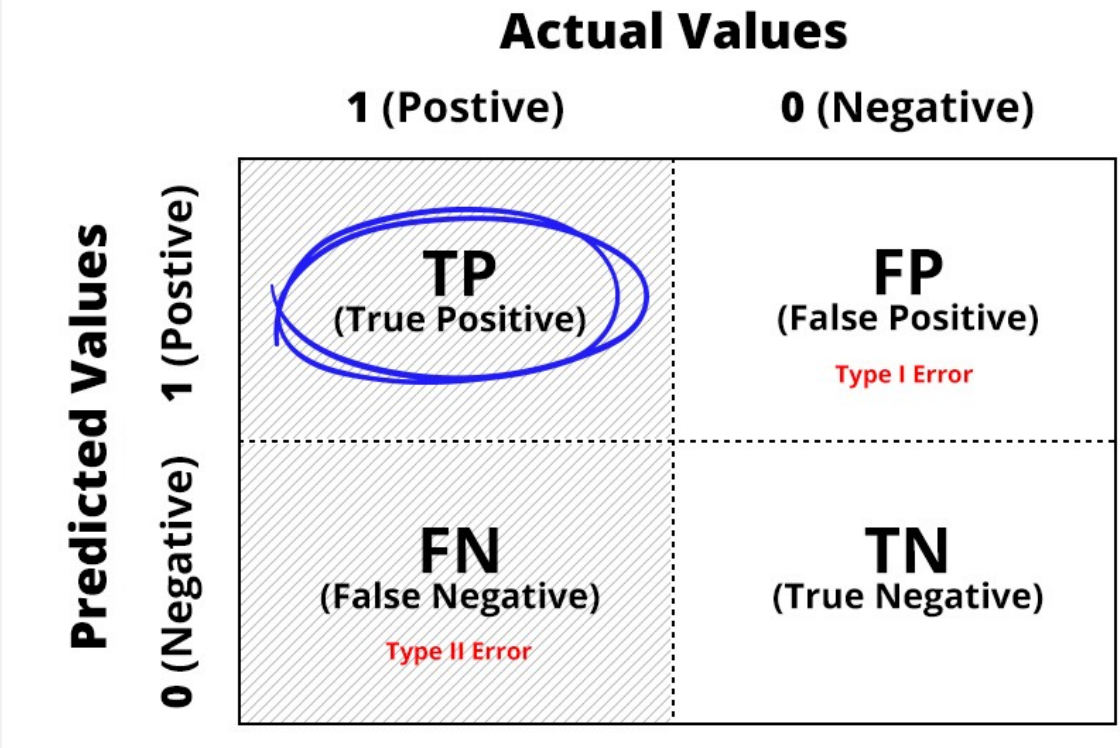
$$= 0{,}75 * 100\%$$

$$= 75\%$$

Menghitung nilai precision dari contoh confusion matrix klasifikasi biner.

### Recall atau Sensitivity (True Positive Rate)

*Recall* menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi. Maka, *recall* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Nilai *recall* dapat diperoleh dengan persamaan (3).



Confusion matrix yang menggambarkan nilai recall.

$$recall = \frac{TP}{TP + FN} \quad (3)$$

Persamaan nilai recall.

Dari contoh *confusion matrix* klasifikasi biner diatas maka dengan menghitung nilai *recall* dapat menjawab pertanyaan “Berapa persen pasien yang diprediksi kanker dibandingkan keseluruhan pasien yang sebenarnya menderita kanker”.

$$recall = \frac{jumlah\ pasien\ kanker\ diprediksi\ \mathbf{benar}}{jumlah\ pasien\ kanker}$$

$$= \frac{6}{6+3} = \frac{6}{9} = 0{,}66$$

$$= 0{,}66 * 100\%$$

$$= 66\%$$

Menghitung nilai recall dari contoh confusion matrix klasifikasi biner.

### Confusion Matrix pada Python

Pada bagian ini saya akan memberikan contoh bagaimana cara membuat model sederhana untuk prediksi dan menampilkan *confusion matrix*nya untuk menghitung beberapa *performance metrics* pada python. Kita membutuhkan library **scikit-learn** untuk menghasilkan *confusion matrix*, jadi pastikan anda telah menginstallnya terlebih dahulu. Fungsi `confusion_matrix()` akan menghitung *confusion matrix* pada model dan mengembalikan hasilnya dalam bentuk *array*.

## Dataset yang Digunakan

Saya menggunakan dataset diagnosa kanker payudara dari kaggle untuk melakukan prediksi penderita kanker. Disarankan anda untuk mengunduh dataset dan mengikuti langkah yang ada pada tulisan ini.

```
import pandas as pd
import numpy as np# membaca dataset
data = pd.read_csv("data/breast-cancer-wisconsin-data.csv")# mengahapus kolom yang tidak digunakan
data.drop(["Unnamed: 32","id"], axis=1, inplace=True)# merubah label M(ganas) = 1 dan B(jinak) = 0
data.diagnosis = [1 if each == "M" else 0 for each in data.diagnosis]data.head(3) # menampilkan sample data
```

| | diagnosis | radius_mean | texture_mean | perimeter_mean | area_mean | smoothness_mean |
|---|---|---|---|---|---|---|
| 0 | 1 | 17.99 | 10.38 | 122.8 | 1001.0 | 0.11840 |
| 1 | 1 | 20.57 | 17.77 | 132.9 | 1326.0 | 0.08474 |
| 2 | 1 | 19.69 | 21.25 | 130.0 | 1203.0 | 0.10960 |

Contoh dataset (menggunakan Pandas.head)

## Membagi Dataset dan Membuat Model Prediksi

Dataset dibagi ke dalam data latih dan data uji, data latih digunakan untuk melatih model yang telah dibuat sedangkan evaluasinya akan dilakukan pada data uji.

```
from sklearn.model_selection import train_test_splitx_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25, random_state=42)
```

Selanjutnya, kita akan membuat model sederhana menggunakan *Decision Tree* untuk melakukan prediksi.

```
from sklearn.tree import DecisionTreeClassifiermodel = DecisionTreeClassifier()
model.fit(x_train, y_train)
y_pred = model.predict(x_test) # prediksi
```

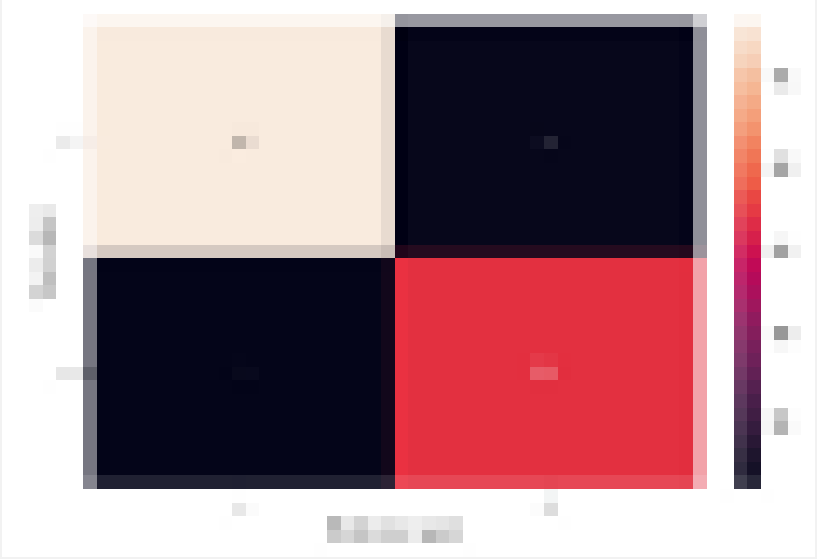
### Evaluasi Model menggunakan Confusion Matrix

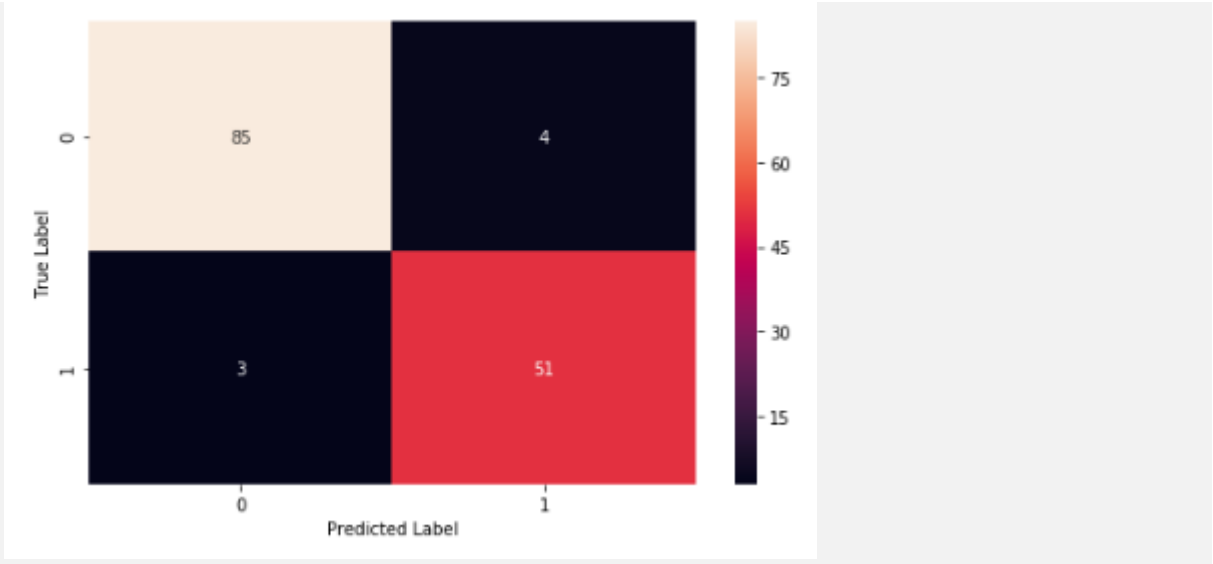
Kita akan menggunakan *confusion matrix* untuk mengevaluasi model yang sudah kita buat sebelumnya.

```
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred) # ouput
# array([[95,  3],
#        [ 2, 43]])
```

Kita dapat memvisualisasikan *confusion matrix* tersebut untuk memudahkan dalam mengevaluasi.

```
import seaborn as sns
import matplotlib.pyplot as pltf, ax =
plt.subplots(figsize=(8,5))
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True,
fmt=".0f", ax=ax)
plt.xlabel("y_head")
plt.ylabel("y_true")
plt.show()
```






Visualisasi dari confusion matrix.

### Seberapa baik model kita ?

Seperti yang sudah saya jelaskan diatas, dengan *confusion matrix* kita dapat mengetahui keakuratan dari model yang kita buat dengan *performance metrics* seperti: *accuracy*, *recall*, dan *precision*.

```
from sklearn.metrics import cclassification_report
print (classification_report(y_test, y_pred))# ouput
#              precision    recall  f1-score   support
0       0.98      0.97      0.97        98
           1       0.93      0.96      0.95        45
accuracy                           0.97       143
   macro avg       0.96      0.96      0.96       143
weighted avg       0.97      0.97      0.97       143
```

***Evaluasi cluster lain adalah :***

**Davies-Bouldin Index seperti dalam PDF file**

Reff :

Youtube : Romi Satria Wahono

https://medium.com/@ksnugroho/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f

NIM : 202420040 - ISTIARSO

# KLASTERISASI PROSES SELEKSI PEMAIN MENGGUNAKAN ALGORITMA K-MEANS
## (Study Kasus : Tim Hockey Kabupaten Kendal)

Alith Fajar Muhammad

*Jurusan Teknik Informatika FIK UDINUS, Jl. Nakula No. 5-11 Semarang-50131*

alithfm@gmail.com

Abstrak - Klasterisasi pemain hockey dilakukan untuk mengelompokkan pemain kedalam cluster tertentu yang memiliki kemiripan data menggunakan algoritma K-means. Data yang diambil meliputi pukulan hit, push, tapping serta data *Multi level running speed* dan Sprint 50 meter. Metode evaluasi cluster menggunakan DBI dan purity untuk mengukur seberapa bagus cluster yang dihasilkan. Dari hasil penelitian yang menggunakan 100 data pemain (59 Putra 41 Putri) menghasilkan 3 cluster dengan evaluasi nilai sebagai berikut : Untuk nilai DBI pada Cluster Putra adalah 0.95934206 dan Cluster Putri adalah 0.976979445. Sedangkan Nilai Purity untuk Cluster Putra adalah 0.879 dan Cluster Putri adalah 0.608. Ini membuktikan bahwa pengelompokan dengan k-means menghasilkan cluster yang masih belum cukup maksimal.

Kata kunci : Hockey, K-means, Cluster , Matlab, *davies-bouldin index* (DBI), Purity

## I. PENDAHULUAN

*Datamining* berisi pencarian trend atau pola yang diinginkan dalam database yang besar untuk membantu pengambilan keputusan diwaktu yang akan datang. Harapannya, perangkat *datamining* mampu mengenali pola-pola ini dalam data dengan masukan yang minimal. Pola-pola ini dikenali oleh perangkat tertentu yang dapat memberikan suatu anailsa data yang berguna dan berwawasan yang kemudian dapat dipelajari lebih teliti, yang mungkin saja menggunakan perangkat pendukung keputusan yang lainnya [1].

*Clustering* (Klasterisasi) adalah proses mengelompokkan atau penggolongan objek berdasarkan informasi yang diperoleh dari data yang menjelaskan hubungan antar objek dengan prinsip untuk memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas / *cluster*. *Clustering* dalam datamining berguna untuk menemukan pola distribusi di dalam sebuah data set yang berguna untuk proses analisa data. Kesamaan objek biasanya diperoleh dari kedekatan nilai-nilai atribut yang menjelaskan objek-objek data, sedangkan objek-objek data biasanya direpresentasikan sebagai sebuah titik dalam ruang multidimensi [2].

*Hockey* adalah olahraga permainan yang dilakukan oleh pria dan wanita dengan menggunakan alat pemukul (*stick*) dan bola. [3]. Dari uraian diatas olahraga *hockey* yang populer di jawa tengah ada di beberapa kota/kabupaten. Salah satunya adalah tim *hockey* kabupaten kendal. Tim *hockey* Kabupaten Kendal lahir pada tahun 2005. Seiring dengan pesatnya olahraga *hockey* di kabupaten kendal khususnya di kecamatan boja dan sekitarnya sekarang jumlah anggotanya lebih dari 50 orang, terdiri dari pria dan wanita.

Pada proses seleksinya Tim Kepelatihan tidak selalu menyeleksi pemain berdasarkan kriteria yang ditetapkan, sehingga terkadang proses seleksi tidak berjalan sesuai kemampuan yang diukur dengan tepat untuk setiap pemain.

Oleh karena itu diadakan penelitian yang bertujuan untuk menggali potensi data personal di setiap pemain yang akan di kelompokkan ke dalam *cluster* menggunakan algoritma K-Means.

## II. TEORI PENUNJANG

### 2.1 Metode Clustering

*Clustering* adalah suatu metode pengelompokan berdasarkan ukuran kedekatan (kemiripan). *Clustering* adalah proses mengelompokkan objek berdasarkan informasi yang diperoleh dari data yang menjelaskan hubungan antar objek dengan prinsip untuk memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas/*cluster*.

Tujuannya menemukan *cluster* yang berkualitas dalam waktu yang layak. *Clustering* dalam *datamining* berguna untuk menemukan pola distribusi di dalam sebuah data set yang berguna untuk proses analisa data. Kesamaan objek biasanya diperoleh dari kedekatan nilai-nilai atribut yang menjelaskan objek-objek data, sedangkan objek-objek data biasanya direpresentasikan sebagai sebuah titik dalam ruang multidimensi **[7]**.

Beberapa manfaat clustering adalah sebagai berikut :

a. Identifikasi Obyek (*Recognition*)
   Dalam bidang *image processing*, *Computer Vision* atau robot vision.
b. *Decission Support System* (Segementasi Pasar)

### 2.2 Algoritma *K-Means*

Algoritma *k-means* merupakan algoritma yang digunakan untuk pengelompokan iteratif, algoritma ini melakukan partisi set data ke dalam sejumlah K cluster yang sudah ditetapkan diawal. Partisi set data tersebut dilakukan untuk mengetahui karakteristik dari setiap *cluster*, sehingga *cluster* yang memiliki karakteristik sama dikelompokkan kedalam satu *cluster* dan yang memiliki karakteristik berbeda dikelompokkan kedalam *cluster* lain. Berikut langkah-langkah perhitungan dalam *k-means*, antara lain [1] [3] [9] [7] [10] [11]:

1. Tentukan jumlah *cluster* dan ambang batas perubahan fungsi objektif
2. Menentukan *centroid* awal yang digunakan
3. Menghitung jarak setiap data ke masing-masing centroid menggunakan jarak *euclidean* untuk mendapatkan jarak terdekat data dengan *centroid*nya
4. Menentukan centroid baru dengan menghitung nilai rata-rata dari data yang ada pada centroid yang sama
5. Ulangi langkah 3 dan 4 hingga kondisi konvergen tercapai, yaitu perubahan fungsi objektif sudah dibawah ambang batas yang diinginkan, atau tidak ada data yang berpindah *cluster*, atau perubahan posisi *centroid* sudah dibawah ambang batas yang sudah ditentukan.

Setelah perhitungan jarak dari setiap data terhadap *centroid* dihitung, kemudian dipilih jarak yang paling kecil atau yang mendekati nilai 0 sebagai *cluster* yang akan diikuti sebagai relokasi data pada *cluster* di sebuah iterasi. Relokasi sebuah data dalam cluster yang diikuti dapat dinyatakan dengan nilai keanggotaan *a* yang bernilai 0 atau 1. Nilai 0 jika tidak menjadi anggota sebuah cluster dan 1 jika menjadi anggota sebuah cluster. *K-means* mengelompokkan data secara tegas hanya pada satu cluster, maka nilai *a* sebuah data pada semua *cluster*, hanya satu yang bernilai 1. Perhitungan jarak antara data dan centroid dapat dilakukan dengan menggunakan persamaan *euclidean distance*, persamaannya sebagai berikut [1] [7] [10] [11] :

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \quad (2.1)$$

Keterangan :

$d_{ij}$ = Jarak objek antara objek i dan j

$P$ = $Dimensi\ data$

$x_{ik}$ = obyek i pada dimensi k

$x_{jk}$ = obyek j pada dimensi k

Dan dibawah ini persamaan untuk mencari nilai fungsi objektif setiap data [1] [7]:

$$\sum_{li=1}^{NK} \{x_{li-C_{li}}\}^2 \qquad (2.2)$$

NK adalah jumlah data yang tergabung dalam sebuah cluster

### 2.3 Evaluasi *Cluster*

Metode yang digunakan dalam menentukan evaluasi cluster ini menggunakan *davies-bouldin index (DBI)* dan *purity.*

#### 1. Davies-Bouldin Index

*Davies-bouldin index* merupakan salah satu metode evaluasi internal yang mengukur evaluasi cluster pada suatu metode pengelompokan yang didasarkan pada nilai kohesi dan separasi. Dalam suatu pengelompokan, kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap centroid dari cluster yang diikuti. Sedangkan separasi didasarkan pada jarak antar centroid dari clusternya.

*Sum of square within cluster* (SSW) merupakan persamaan yang digunakan untuk mengetahui matrik kohesi dalam sebuah cluster ke-i yang dirumuskan sebagai berikut :

$$SSW_i = \frac{1}{m_i} \sum_{j=i}^{m_i} d(x_j, c_i) \qquad (2.3)$$

Dari persamaan tersebut, $m_i$ merupakan jumlah data dalam cluster ke-i, $c_i$ adalah centroid cluster ke-i, dan *d()* merupakan jarak setiap data kecentroid yang dihitung menggunakan jarak *euclidean.*

*Sum of square between cluster* (SSB) merupakan persamaan yang digunakan untuk mengetahui separasi antar cluster yang dihitung menggunakan persamaan :

$$SSB_{i,j} = d(c_i, c_j) \qquad (2.4)$$

Setelah nilai kohesi dan separasi diperoleh, kemudian dilakukan pengukuran rasio ($R_{ij}$) untuk mengetahui nilai perbandingan antara cluster ke-i dan cluster ke-j. Cluster yang baik adalah cluster yang memiliki nilai kohesi sekecil mungkin dan separasi yang sebesar mungkin. Nilai rasio dihitung menggunakan persamaan sebagai berikut :

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \qquad (2.5)$$

Nilai rasio yang diperoleh tersebut digunakan untuk mencari nilai *davies-bouldin index* (DBI) dari persamaan berikut :

$$\text{DBI} = \frac{1}{K} \sum_{i=1}^{k} max_{i \neq j}\ (R_{i,j}) \qquad (2.6)$$

Dari persamaan tersebut, *k* merupakan jumlah cluster yang digunakan. Semakin kecil nilai DBI yang diperoleh (non-negatif >= 0), maka semakin baik *cluster* yang diperoleh dari pengelompokan *K-means* yang digunakan [7].

#### 2. Purity

*Purity* digunakan untuk menghitung kemurnian dari suatu *cluster* yang direpresentasikan sebagai anggota

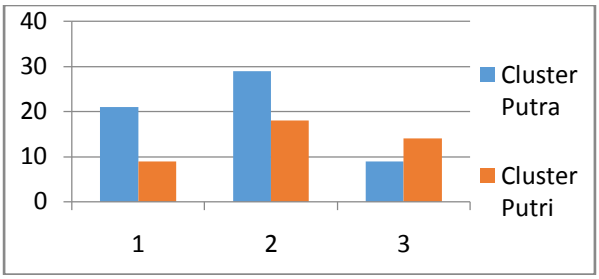
*cluster* yang paling banyak sesuai (cocok) disuatu kelas. Nilai *purity* yang semakin mendekati 1 menandakan semakin baik *cluster* yang diperoleh. Untuk menghitung nilai purity setiap *cluster* dapat menggunakan rumus berikut [13]:

$$Purity\ (j) = \frac{1}{nj}\max(n_{ij}) \quad (2.7)$$

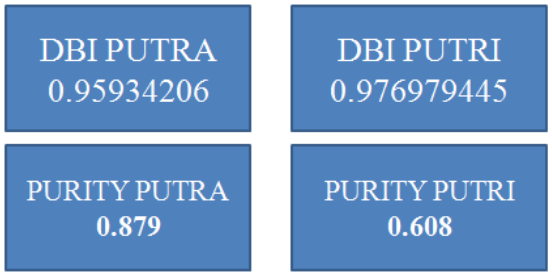
Sementara untuk menghitung purity keseluruhan jumlah K *cluster*, digunakan persamaan sebagai berikut:

$$Purity = \sum_{i=1}^{k} \frac{nj}{n}\ Purity(j) \quad (2.8)$$

## III. HASIL &IMPLEMENTASI




Hasil *cluster* untuk putra dan putri menunjukkan bahwa *Cluster* 1 terdapat 21 putra dan 9 putri, *Cluster* 2 terdapat 29 putra dan 18 putri sedangkan untuk *Cluster* 3 terdapat 9 putra dan 14 putri, ini menunjukkan bahwa *cluster* yang lebih dominan memiliki kemiripan data adalah cluster ke 2.




Dari perhitungan tersebut telah diperoleh nilai *DBI* untuk Cluster Putra adalah 0.95934206 dan untuk Cluster putri adalah 0.976979445. Semakin kecil nilai *DBI* atau semakin mendekati nilai 0 menunjukkan seberapa baik cluster yang diperoleh, sehingga nilai *DBI* yang telah diperoleh tersebut menunjukkan cluster dihitung dalam penelitian ini masih belum cukup bagus.

Sedangkan untuk Evaluasi *cluster* menggunakan Purity adalah sebagai berikut :

| Putra | | | | | |
|---|---|---|---|---|---|
| Cluster | Jumlah | Bagus | Sedang | Kurang | Purity |
| C1 | 21 | 0 | 4 | 17 | 0.809 |
| C2 | 29 | 3 | 26 | 0 | 0.896 |
| C3 | 9 | 9 | 0 | 0 | 1 |
| Total | 59 | | | | |
| Purity Putra | | | | | **0.879** |

| Putri | | | | | |
|---|---|---|---|---|---|
| Cluster | Jumlah | Bagus | Sedang | Kurang | Purity |
| C1 | 9 | 1 | 6 | 2 | 0.666 |
| C2 | 18 | 13 | 5 | 0 | 0.722 |
| C3 | 14 | 2 | 6 | 6 | 0.428 |
| Total | 41 | | | | |
| Purity Putri | | | | | **0.608** |

Dari perhitungan tersebut telah diperoleh nilai *purity* untuk Cluster Putra adalah 0.879 dan Cluster Putri adalah 0.608 . Hasil yang diperoleh ini menunjukkan nilai *purity* dari setiap data *custer* yang dihasilkan cukup bagus untuk cluster putra karena mendekati nilai 1 dan belum cukup bagus untuk cluster putri karena belum mendekati nilai 1.

## IV. PENUTUP

Berdasarkan penelitian yang dilakukan dalam klasterisasi proses seleksi pemain tim Hockey Kendal menggunakan algoritma K-means , menghasilkan Tiga Cluster yang telah tersusun untuk 100 Pemain (59 Putra dan 41 Putri) serta nilai evaluasi DBI serta Purity dari cluster tersebut. Sedangkan hasil dari penelitian yang diperoleh menunjukkan nilai *DBI* dan *purity* yang masih cukup rendah untuk dapat dikatakan mendekati nilai 0 atau 1. Sehingga dapat dikatakan proses Cluster yang dihasilkan belum cukup bagus untuk dikatakan sebagai cluster yang baik. Hasil ini dapat dipengaruhi berbagai hal diantaranya adalah proses pengukuran data

(tolok ukur) yang mengakibatkan nilai diantara pemain hampir sama.

## DAFTAR PUSTAKA


[1] T. Rismawan dan S. Kusumadewi, "Aplikasi K-Means Untuk Pengelompokan Mahasiswa Berdasarkan Nilai Body Mass Index & Ukuran Kerangka," *SNATI*, pp. 43-48, Juni 2008.

[2] H. A. Fajar, DATA MINING, vol. 2, Yogyakarta: ANDI, 2013.

[3] A. Elizabeth dan M. Sue, Field Hockey Step To Succes 2nd, vol. 15, New Zealand: Human Kinetics, 2008, pp. 73-82.

[4] Ediyanto, M. N. Mara dan N. Satyahadewi, "Pengklasifikasian Karakteristik dengan Metode K-Means Cluster Analysis," *Buletin Ilmiah, Mat Stat, dan Terapannya (Bimaster)*, vol. 02, pp. 133-136, 2013.

[5] P. D. S. Sani dan M. S. Dedy ST, Pengantar Data Mining Menggali Pengetahuan Dari Bongkahan Data, Jogjakarta: ANDI OFFSET, 2010.

[6] T. Primadi, Hockey dan Kreativitas dalam Olahraga, Bandung: ITB, 1985.

[7] S. C, M.CitraDevi dan G. Geetharamani, "An Analysis on The Performance of K-Means Clustering Algorithm For Cardiotogram Data Cluster," *International Journal on Computational Sciences & Aplications*, vol. 2, no. 5, pp. 11-20, Oktober 2012.

[8] Widiarina dan W. R. S., "Algoritma Cluster Dinamik untuk Optimasi Cluster pada Algoritma K-Means dalam Pemetaan Nasabah Potensial," *Journal Of Intelligent System*, vol. 1, no. 1, pp. 32-35, Februari 2015.

[9] R. Handoyo, R. R. M. dan S. M. Nasution, "Perbandingan Metode Clustering Menggunakan Metode Single Linkage dan K-Means pada Pengelompokan Dokumen," *JSM STMK Mikrosil*, vol. 15, no. 2, pp. 73-82, Oktober 2015.

[10] T. Khotimah, "Pengelompokan Surat Dalam Al Qur'an Menggunakan Algoritma K-Means," *Jurnal SIMETRIS*, vol. 5, no. 1, pp. 83-88, April 2014.

[11] J. Ong, "Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University," *Jurnal Ilmiah Teknik Industri*, vol. 12, no. 1, pp. 10-20, Juni 2013.

[12] P. Eko, DATA MINING. Mengolah Data Menjadi Informasi Menggunakan Matlab, Yogyakarta: ANDI, 2014.

[13] K. Prilianti dan H.Wijaya, "Aplikasi Text Mining Untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering," *Jurnal Cybermatika*, vol. 2, no. 1, Juni 2014.

[14] T. E. Purwoastuti dan W. S. Elishabet, Metodologi Penelitian, Yogyakarta: PUSTAKABARUPRESS, 2014.

Penelitian ini bertujuan untuk mengembangkan suatu perangkat lunak yang dapat mengelompokkan mahasiswa berdasarkan performa akademiknya. Keberhasilan pengelompokan data mahasiswa dihitung berdasarkan nilai standard eror dari proses pengelompokkan yang dilakukan secara berulang.

Langkah-langkah yang dilakukan pada penelitian ini adalah:

1. Studi literatur mengenai K-Means *Clustering*,
2. Menganalisis penelitian yang telah ada sebelumnya,
3. Mengumpulkan data untuk penelitian dengan cara mengumpulkan IP yang didapatkan dari bagian administrasi fakultas,
4. Mengubah format data yang telah dikumpulkan sehingga dapat memenuhi kebutuhan penelitian,
5. Melakukan pengembangan perangkat lunak dengan metode Rational Unified Process (RUP),
6. Melakukan analisis dan pembahasan hasil eksperimen masukan serta penggunaannya pada perangkat lunak,
7. Membuat kesimpulan terhadap hasil penelitian.

### *3.2 Pengelompokkan Performa Akademik Mahasiswa dengan K-Means*

Metode K-Means *Clustering* memiliki enam tahapan. Pada tahap pertama yaitu menentukan jumlah *cluster* (k) awal untuk menjadi acuan proses *clustering*. Banyaknya *cluster* (k) harus lebih kecil dari pada banyaknya data (k < n). Pada penelitian ini jumlah *cluster* adalah sebanyak 4 *cluster* (k = 4).

Data pada Tabel 1 adalah contoh data yang digunakan.Data yang digunakan pada penelitian ini adalah data performa akademik mahasiswa Fakultas Ilmu Komputer Universitas Sriwijaya.Data yang diambil berupa Indeks Prestasi mahasiswa dari semester 1 sampai dengan 4.

Tabel 1. Contoh Data Akademik Mahasiswa

Berdasarkan data tersebut, *centroid* awal diambil secara acak di tiap *cluster* IP. Pembangkitan nilai acak pada *centroid* awal menggunakan metode pseudo random.Contoh nilai *centroid* acak yang diambil dapat dilihat pada Tabel 2.

Tabel 2.*Centroid* Awal pada pengulangan ke - 0

| C | a | b | c | d |
|---|---|---|---|---|
| c1 | 3,04 | 3,11 | 3,27 | 2,43 |
| c2 | 3,07 | 3,15 | 3,09 | 2,39 |
| c3 | 3,17 | 3,17 | 2,68 | 2,83 |
| c4 | 3,1 | 3,12 | 3,6 | 2,87 |

c4 :*cluster* 4 d : nilai ip4 *cluster*

Untuk pengulangan berikutnya (pengulangan ke–1 sampai selesai), *centroid* baru ditentukan dengan menghitung rata – rata data pada sertiap *cluster*. Jika ditemukan nilai *centroid* baru yang berbeda dengan *centroid* sebelumnya, maka proses dilanjutkan ke langkah berikutnya. Namun, jika nilai *centroid* yang baru dihitung sama dengan *centroid* sebelumnya maka proses *clustering* berhenti.

Proses selanjutnya adalah menghitung jarak masing – masing data terhadap *centroid* dengan menggunakan persamaan (1), jarak data 1 dengan tiap *cluster* adalah :

$$d(x_1, c_1) = \sqrt{(IP_1 - C1_a)^2 + (IP_2 - C1_b)^2 + (IP_3 - C1_c)^2 + (IP_4 - C1_d)^2}$$
$$= \sqrt{(3{,}21 - 3{,}04)^2 + (3{,}5 - 3{,}11)^2 + (2{,}96 - 3{,}27)^2 + (3{,}32 - 2{,}43)^2}$$
$$= 1{,}0340$$

$$d(x_1, c_2) = \sqrt{(IP_1 - C2_a)^2 + (IP_2 - C2_b)^2 + (IP_3 - C2_c)^2 + (IP_4 - C2_d)^2}$$
$$= \sqrt{(3{,}21 - 3{,}07)^2 + (3{,}5 - 3{,}15)^2 + (2{,}96 - 3{,}09)^2 + (3{,}32 - 2{,}39)^2}$$
$$= 1{,}0118$$

$$d(x_1, c_3) = \sqrt{(IP_1 - C3_a)^2 + (IP_2 - C3_b)^2 + (IP_3 - C3_c)^2 + (IP_4 - C3_d)^2}$$
$$= \sqrt{(3{,}21 - 3{,}17)^2 + (3{,}5 - 3{,}17)^2 + (2{,}96 - 2{,}68)^2 + (3{,}32 - 2{,}83)^2}$$
$$= 0{,}6549$$

$$d(x_1, c_4) = \sqrt{(IP_1 - C4_a)^2 + (IP_2 - C4_b)^2 + (IP_3 - C4_c)^2 + (IP_4 - C4_d)^2}$$
$$= \sqrt{(3{,}21 - 3{,}1)^2 + (3{,}5 - 3{,}12)^2 + (2{,}96 - 3{,}6)^2 + (3{,}32 - 2{,}87)^2}$$
$$= 0{,}8766$$

Berdasarkan perhitungan di atas data 1 masuk ke dalam *cluster* 3, karena data 1 memiliki jarak terpendek terhadap *centroid* pada *cluster* 3. Untuk hasil perhitungan selanjutnya dapat dilihat pada Tabel 3.

Tabel 3 Hasil Perhitungan Jarak dan Pengelompokkan Data

| Data | dc1 | dc2 | dc3 | dc4 |
|---|---|---|---|---|
| 1 | 1,034021 | 1,011879 | 0,654981 | 0,876698 |
| 2 | 0,955615 | 0,964002 | 0,809074 | 0,755116 |
| 3 | 0,838033 | 0,812773 | 0,619758 | 0,772075 |
| 4 | 1,281718 | 1,347108 | 1,227436 | 0,862206 |
| 5 | 1,297459 | 1,31145 | 1,167733 | 1,048141 |
| 6 | 0,93755 | 1,002247 | 1,12641 | 0,696994 |
| 7 | 1,108783 | 1,152953 | 0,949421 | 0,743774 |
| 8 | 0,687459 | 0,665658 | 0,533854 | 0,670373 |
| 9 | 0,130384 | 0,155242 | 0,707107 | 0,644205 |
| 10 | 0,809197 | 0,633325 | 0,629126 | 1,274912 |

Pada Tabel 3 kolom data yang diwarnai menunjukan bahwa data tersebut memiliki jarak terpendek terhadap masing-masing *cluster*, sehingga data tersebut dikelompokan sesuai warna *cluster*-nya. Pada hasil perhitungan di atas diperoleh 1 data untuk *cluster* 1, 4 data untuk *cluster* 3, 5 data untuk *cluster* 4 dan tidak ada data yang masuk ke *cluster* 2.

Setelah beberapa kali perulangan, proses pengelompokkan data menunjukkan data terletak pada *cluster* yang sama seperti di perulangan sebelumnya. Jika *cluster* data tetap, maka nilai *centroid* juga tetap maka perulangan berhenti. Pada perulangan ke empat diperoleh hasil akhir yaitu 1 data 6 untuk *cluster* 1, 1 data untuk *cluster* 2, 4 data untuk *cluster* 3 dan 4 data untuk *cluster* 4. Hasil akhir pada perulangan ke empat dapat dilihat pada Tabel 4.

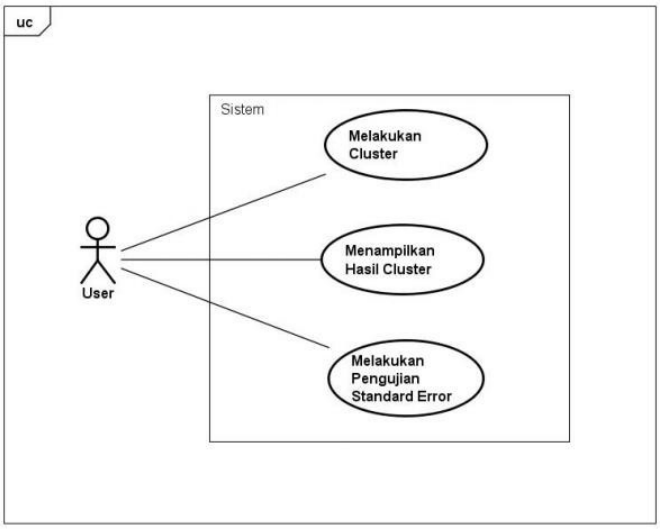
Tabel 4. Hasil Akhir Pengelompokkan

| No | dc1 | dc2 | dc3 | dc4 |
|---|---|---|---|---|
| 1 | 1,082867 | 1,240806 | 0,253377 | 0,6179 |
| 2 | 1,02323 | 1,321363 | 0,238747 | 0,416893 |
| 3 | 0,90161 | 1,096586 | 0,214243 | 0,656125 |
| 4 | 1,33559 | 1,792986 | 0,734166 | 0,286705 |
| 5 | 1,330413 | 1,69393 | 0,68775 | 0,343802 |
| 6 | 1,005982 | 1,521249 | 0,659545 | 0,442493 |
| 7 | 1,1755 | 1,52951 | 0,448553 | 0,374433 |
| 8 | 0,676018 | 1,024597 | 0,461952 | 0,676166 |
| 9 | 0 | 0,744715 | 0,876926 | 1,156114 |
| 10 | 0,744715 | 0 | 1,129159 | 1,591163 |

## I. HASIL DAN PEMBAHASAN

### *Perangkat Lunak Pengelompokkan Performa Akademik Mahasiswa*

Berdasarkan tujuan dari penelitian untuk membangun perangkat lunak, telah dilaksanakan proses analisis kebutuhan yang menghasilkan diagram use case yang dapat dilihat pada gambar 1.

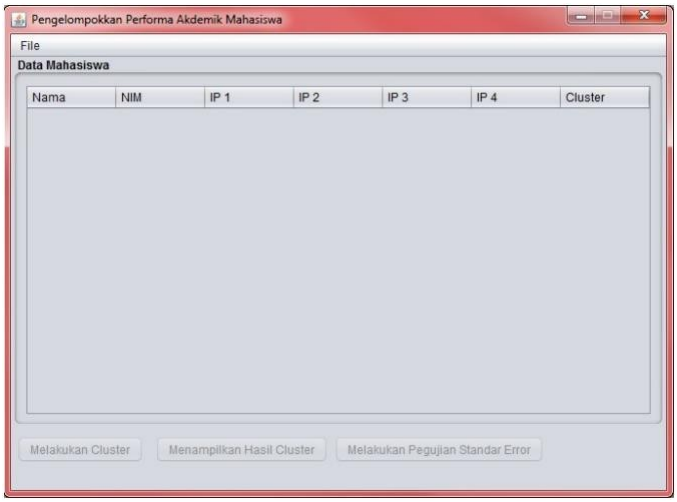



Gambar 1. Use Case Perangkat Lunak Pengelompokkan Performa Akademik Mahasiswa

Terdapat tiga use case utama, yaitu melakukan *cluster*, menampilkan hasil *cluster*, dan melakukan pengujian standardd error. Use case melakukan *cluster* adalah proses yang digunakan saat pengguna akan mengelompokkan data akademik mahasiswa. Data dengan format TXT diunggah pada perangkat lunak, dan kemudian dikelompokkan dengan menggunakan algoritma K-Means Clustering.

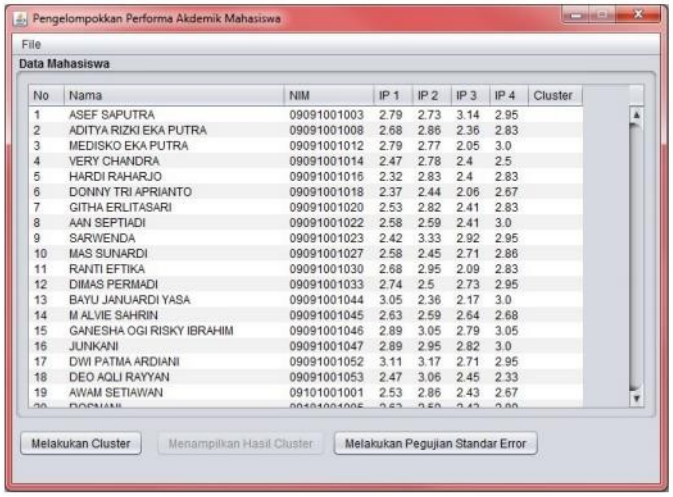
Selanjutnya, use case kedua, menampilkan hasil *cluster*, digunakan untuk menampilkan hasil pengelompokkan dalam bentuk grafis.Hal ini dilakukan untuk mempermudah pengguna dalam melihat hasil pengelompokkan. Pengguna diharapkan dapat mengambil keputusan dengan lebih mudah jika dapat melihat data secara grafik dan secara detil pada saat yang bersamaan. Use case terakhir adalah Melakukan Pengujian Standardd Error. Use case ini digunakan untuk mencari standard deviasi dan standardd error dari proses pengelompokkan. Jika fungsi ini digunakan oleh pengguna, akan dilakukan proses pengelompokkan sebanyak n kali yang kemudian dihitung nilai standardd errornya.

Berdasarkan use case yang telah diidentifikasi, telah dibangun perangkat lunak Pengelompokkan Performa Akademik Mahasiswa. Saat perangkat lunak dibuka, maka pengguna akan mendapatkan tampilan berupa antar muka awal dari perangkat lunak. Antarmuka ini dapat dilihat pada gambar 2.




Gambar 2. Antarmuka Awal Perangkat Lunak Pengelompokkan Performa Akademik Mahasiswa

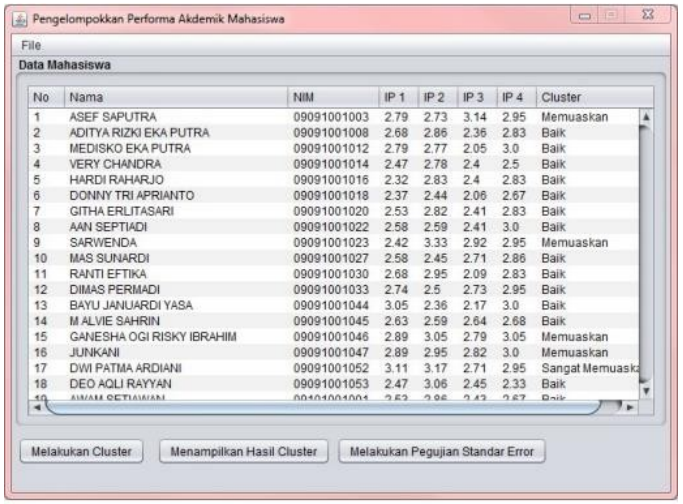
Pada antarmuka ini, hanya menu File yang aktif dan dapat diakses.Menu ini digunakan untuk mengunggah berkas berisi Nama, NIM, dan IP semester 1-4 dari sekelompok mahasiswa. Jika berkas telah diunggah, maka data-data tersebut akan ditampilkan pada panel yang berada di tengah, seperti pada Gambar 3.




Gambar 3. Tampilan Data yang Telah Diunggah pada Sistem

Setelah data diunggah, maka tombol melakukan *cluster* dan Melakukan Pengujian Standard Error menjadi aktif. Pengguna dapat menekan tombol melakukan *cluster* untuk mengelompokkan data menggunakan algoritma K-Means sebanyak 1 kali, sedangkan tombol pengujian akan mengulang pengelompokkan sebanyak n kali sebagai bagian dan proses pengukuran akurasi algoritma K-Means *Clustering*.

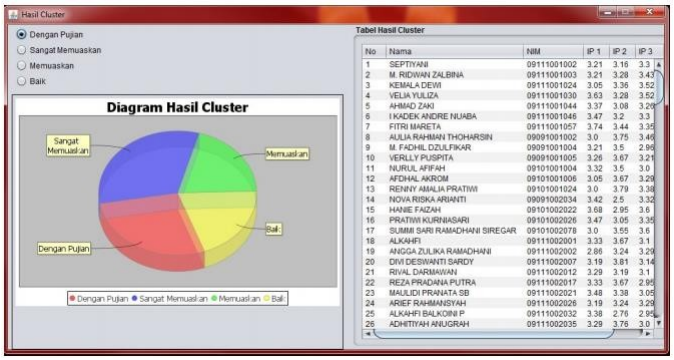
Jika pengguna menekan tombol Melakukan *cluster*, maka proses *clustering* seperti pada Use Case Melakukan *cluster*akan dijalankan. Diakhir proses, *cluster* dari masing-masing mahasiswa akan ditampilkan pada panel, seperti pada Gambar 4.




Gambar 4. Hasil Pengelompokkan Performa Akademik Mahasiswa.

Pada perangkat lunak yang dibangun, *centroid* pada masing-masing *cluster* terlebih dahulu diurutkan, dan diberi label “dengan pujian”, “sangat memuaskan”, “memuaskan”, dan “baik”. Predikat ini diberikan berdasarkan Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Tentang Standard Nasional Pendidikan Tinggi [6].

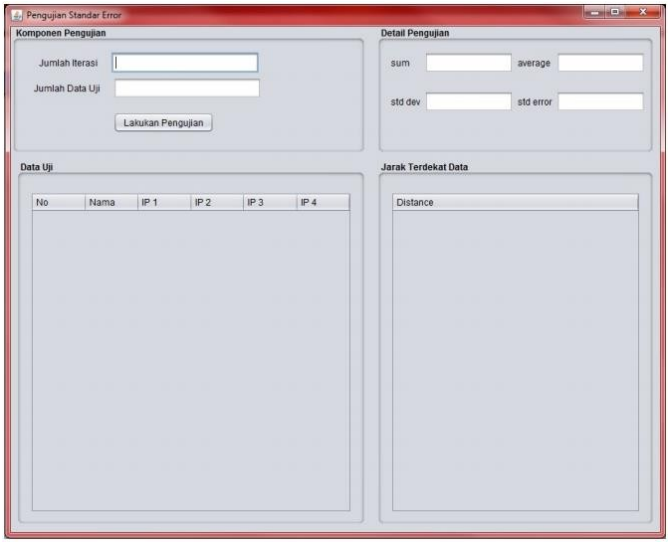
Selanjutnya, pengguna dapat memilih Menampilkan hasil *cluster*. Tombol ini adalah representasi dari Use Case “Menampilkan Hasil *Cluster*”, yang akan menampilkan hasil dalam bentuk *pie-chart*. Tampilan ini dapat dilihat pada gambar 5.




Gambar 5.Hasil pengelompokkan dalam bentuk grafik *pie-chart.*

Grafik persentase hasil pengelompokkan ditampilkan bersebelahan dengan hasil *clustering* dalam bentuk tabel.Hal ini dimaksudkan untuk mempermudah pengambilan keputusan mengenai mahasiswa dengan performa akademik tertentu.

Tombol lainnya yang dapat diakses adalah tombol Melakukan Pengujian Standard Eror yang merupakan implementasi dari use case “Melakukan Pengujian Standard Error. Tampilan dari fitur ini dapat dilihat pada Gambar 6.




Gambar 6. Tampilan Pengujian Standard Error

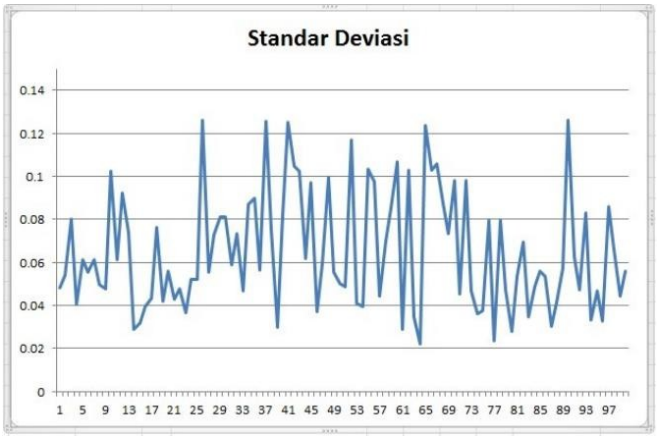
Pada fitur ini, pengguna dapat memilih jumlah iterasi dan banyak data uji yang ingin digunakan dari data yang telah diunggah sebelumnya. Jika Tombol lakukan pengujian ditekan, maka perangkat lunak akan melakukan *clustering* dari data uji yang dipilih sebanyak jumlah iterasi. Perangkat lunak kemudian menampilkan nilai sum, rata-rata, standard deviasi, dan standard error dari pengujian.

*Pengujian Performa K-Means Clustering*

Pada penelitian ini, performa dari K-Means *clustering* dinilai dengan menghitung standard deviasi dan standard error dari proses pengelompokkan yang dilakukan berkali-kali. Standard deviasi adalah ukuran yang digunakan untuk mengukur sebaran data. Sebuah standard deviasi yang rendah menunjukkan bahwa titik data cenderung dekat dengan mean dari himpunan, sementara standard deviasi yang tinggi menunjukkan bahwa titik data tersebar lebih besar dari rentang nilai. Standardd error adalah standard deviasi dari distribusi sampling statistik yang mengukur akurasi sampel yang mewakili populasi.

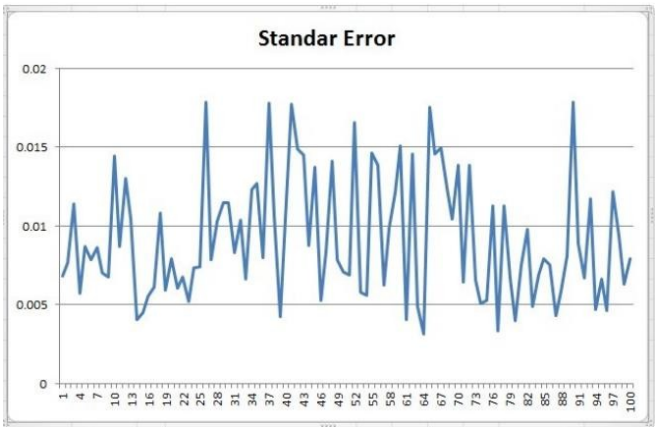
Sebanyak 100 sampel data mahasiswa semester 5 fakultas ilmu komputer Universitas Sriwijaya telah dikumpulkan sebagai data uji yang akan dikelompokkan pada penelitian ini. Data yang digunakan sebagai parameter performa akademik adalah data Indeks Prestasi Mahasiswa semester 1 sampai dengan semester 4.Data tersebut kemudian dikelompokkan sebanyak 50 kali.

Grafik standard deviasi dari 100 data dapat dilihat paga Gambar 7.




Gambar 7. Standard Deviasi Hasil Pengujian K-Means *Clustering*

Berdasarkan pengujian, standard deviasi maksimum dari 100 data adalah sebesar 0.012616 dan minimum sebesar 0.022089.Sedangkan, rata-rata standard deviasi dari masing-masing data adalah sebesar 0.064926.Grafik nilai standard error dari 100 data dapat dilihat pada Gambar 8.




Gambar 8. Standard Error Hasil Pengujian K-Means *Clustering*

Berdasarkan pengujian, standard error maksimum dari 100 data adalah sebesar 0.017842 dan minimum sebesar 0.003124.Sedangkan, rata-rata standard error dari masing-masing data adalah sebesar 0.009182.

Hasil standard error dari 100 kali pengelompokkan data adalah 0.009182 atau 0.9182%. Hal ini menunjukkan bahwa pengelompokkan yang dihasilkan oleh proses K-Means *clustering* tidak banyak mengalami perubahan. Hal ini sejalan dengan kebutuhan yang menginginkan hasil pengelompokkan yang konsisten.

## II. KESIMPULAN

Berdasarkan hasil dan pembahasan dari penelitian yang telah dilakukan, dapat disimpulkan bahwa:

1. Metode K-Means *Clustering* dapat digunakan untuk melakukan pengelompokkan performa akademik mahasiswa.

2. Standard Error rata rata dari data yang diuji adalah sebesar 0.92% yang menunjukkan bahwa metode K–Means *Clustering* dapat melakukan pengelompokkan performa akademik mahasiswa dengan baik.

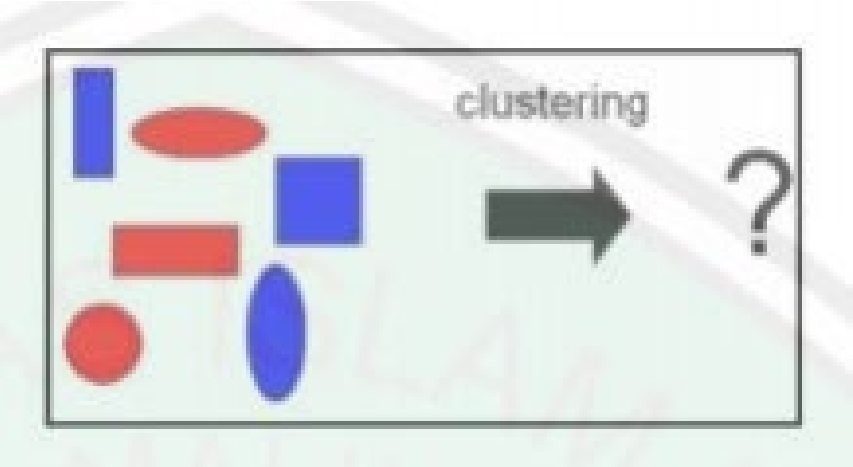
3. *Cluster* akhir dari suatu data berubah – ubah sesuai dengan jarak terdekat data terhadap *centroid*.

Tugas 05

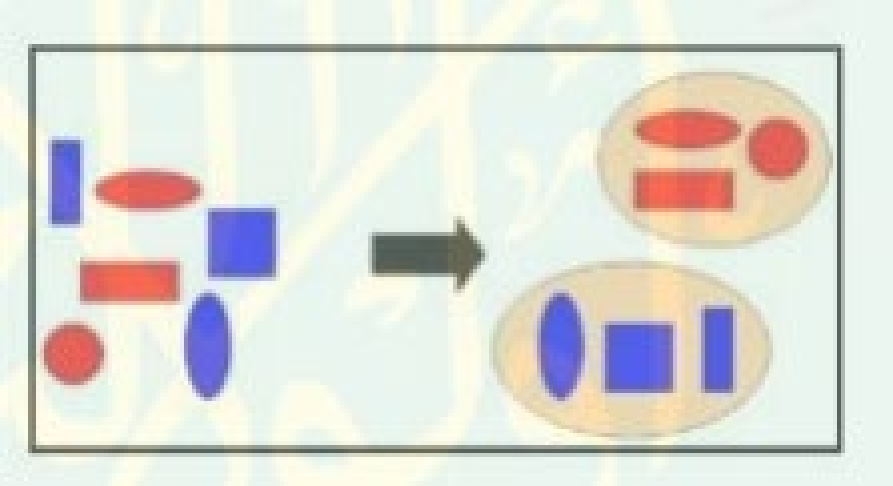
**Nama : Juminovario**
**NIM : 202420018**
**Kelas : MTI 23 Reguler A**
**MK : Advanced Database**

Pengklusteran merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Cluster adalah kumpulan yang memiliki record kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan record-ricord dalam clusterlain.

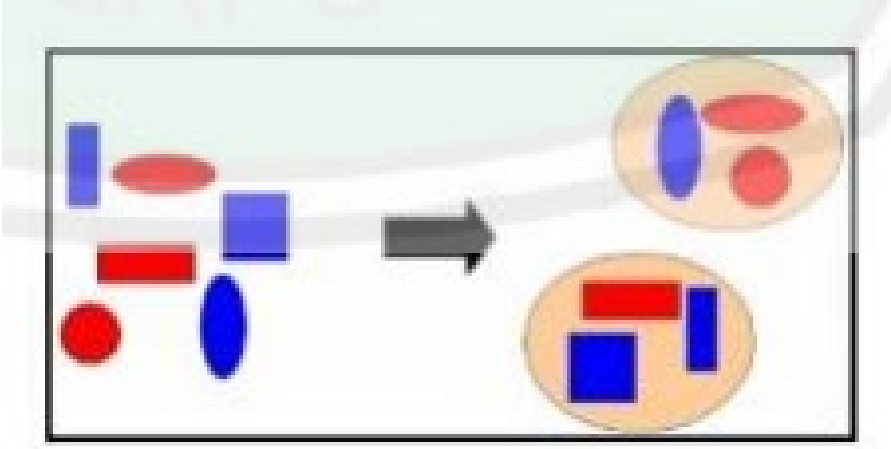
Dalam algoritma clustering, cluster adalah sekumpulan ojek yang mempunyai kesamaan diantara anggotanya dan memiliki ketidak samaan dengan objek lain pada cluster lainnya dengan kata lain sebuah cluster adalah sekumpulan objek yang digabungkan bersama karena persamaan atau kedekatannya. Clustering adalah proses membuat pengelompokan sehingga semua anggota dari setiap  partision mempunyai persamaan berdasarkan matrik tertentu . gambar berikut menunjukkan contoh data yang akan dilakukan klasterisasi :



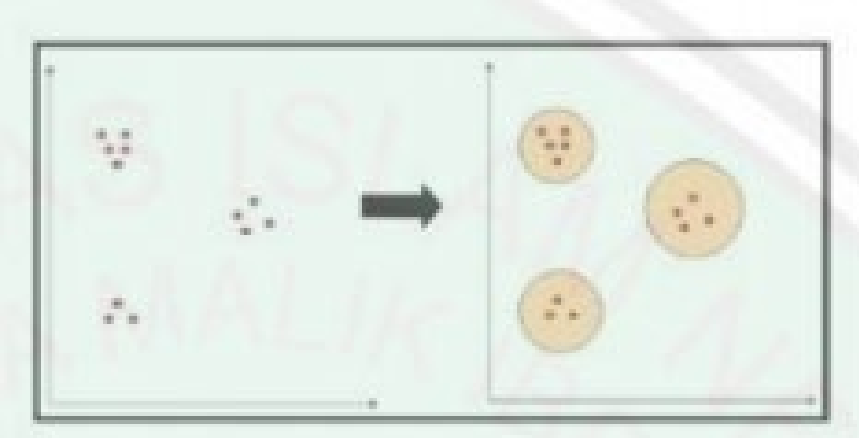
Jika data dilakukan clustering (pengelompokan) berdasarkan warna, maka pengelompokannya seperti gambar berikut ini.



Jika data dilakukan clustering (pengelompokan) berdasarkan bentuk, maka pengelompokannya seperti gambar berikut ini.



Selain menggunakan similaritas (kesamaan) berdasarkan bentuk dan warna clustering juga bisa dilakukan dengan menggunakan similaritas berdasarkan jarak, artinya data yang memiliki jarak berdekatan akan membentuk satu cluster, contohnya dapat dilihat pada gambar berikut ini.



Ada perbedaan antara metode klasifikasi dan metode clustering, dimana pada dasarnya terdapat tiga poin perbedaan yaitu: data, label dan analisis hasil. Perbedaan tersebut dapat ditabelkan seperti tabel berikut:

| Perbedaan | Klasifikasi | Klasterisasi |
|---|---|---|
| Data | Supervised | Unsupervised |
| Label | Ya | Tidak |
| Analisis hasil | Error ratio | Variance |

Data supervised pada klarifikasi artinya data melalui pembelajaran terbimbing, sedangkan data unsupervised pada klasterisasi dinyatakan dengan variance yang menunjukkan variasi data dalam satu cluster, sedangkan klasifikasi analisa hasil diukur menggunkan rasio kesalahan (error ratio). Pada dataset yang digunakan oleh klasifikasi terdapat satu atribut (label) yang berfungsi sebagai atribut target, sedangkan dataset pada klasterisasi tidak terdapat atribut (label) sebagai attribut target.

*Clustering* atau klasterisasi adalah metode pengelompokan data. Menurut Tan, 2006 *clustering* adalah sebuah proses untuk mengelompokan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang maksimum dan data antar *cluster* memiliki kemiripan yang minimum.

*Clustering* merupakan proses partisi satu *set* objek data ke dalam himpunan bagian yang disebut dengan *cluster.* Objek yang di dalam *cluster* memiliki kemiripan karakteristik antar satu sama lainnya dan berbeda dengan *cluster* yang lain. Partisi tidak dilakukan secara manual melainkan dengan suatu algoritma *clustering*. Oleh karena itu, *clustering* sangat berguna dan bisa menemukan *group* atau kelompokyang tidak dikenal dalam data. *Clustering* banyak digunakan dalam berbagai aplikasi seperti misalnya pada *business inteligence,* pengenalan pola citra*, web search,* bidang ilmu biologi, dan untuk keamanan (*security*). Di dalam *business inteligence*, *clustering* bisa mengatur banyak *customer* ke dalam banyaknya kelompok. Contohnya mengelompokan *customer* ke dalam beberapa *cluster* dengan kesamaan karakteristik yang kuat. *Clustering* juga dikenal sebagai data segmentasi karena *clustering* mempartisi banyak data set ke dalam banyak *group* berdasarkan kesamaannya. Selain itu *clustering* juga bisa sebagai *outlier detection*.

Salah satu langkah-langkah untuk mengevaluasi performance sebuah clustering yaitu dapat kita lakukan dengan menerapkan beberapa karakteristik dari clustering yaitu, partitioning clustering, hierarchical clustering, overlapping clustering dan hybrid (merupakan kombinasi dari karakteristik hierarchical, overlapping, partitioning).

1. Partitioning clustering
    a. Disebut juga exclusive clustering
    b. Setiap data harus termasuk dalam cluster tertentu
    c. Memungkinkan bagi setiap data yang termasuk cluster tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah ke cluster yang lain.

    Contoh : k-maens, residual analysis.

2. Hierarchical clustering
    a. Setiap data harus termasuk  dalam cluster tertentu
    b. Memungkinkan bagi setiap data yang termasuk cluster tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah ke cluster yang lain

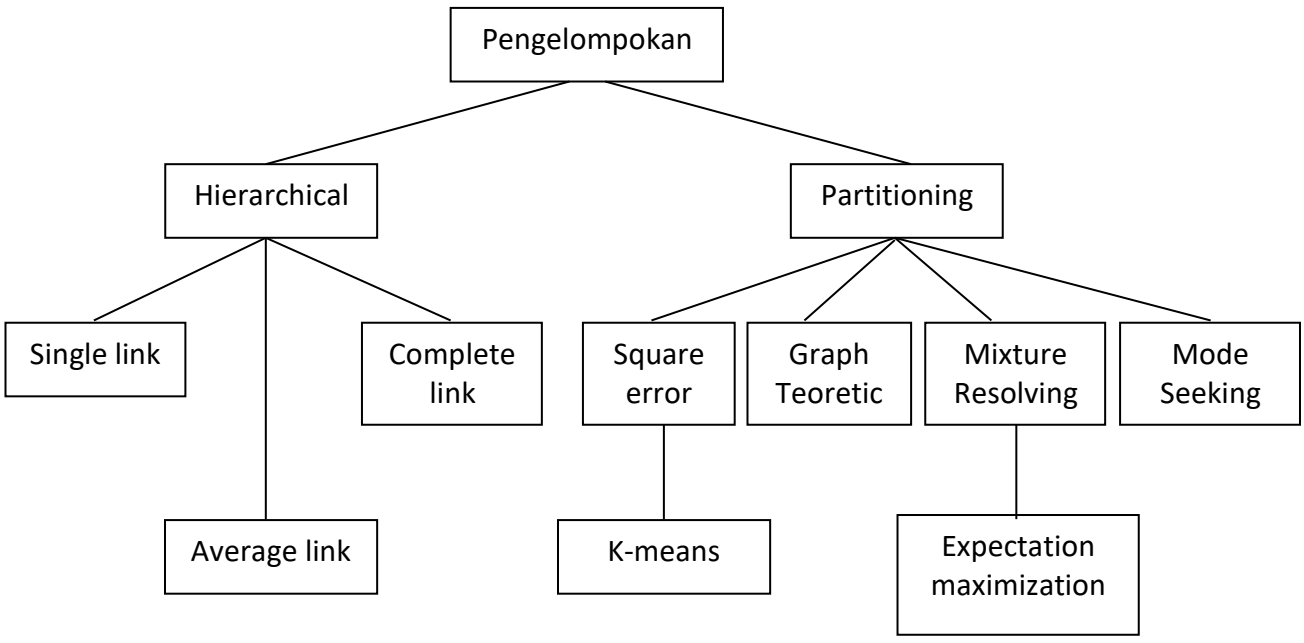
    Contoh : single linkage, centroid linkage, complete linkage.

3. Overlapping clustering
    a. Setiap data harus termasuk  ke beberapa cluster
    b. Data mempunyai nilai keanggotaan (membership) pada beberapa cluster

    Contoh : fuzzy  c-means, gaussian mixture.

4. Hybrid

    Merupakan kombinasi dari karakteristik partitioning, overlapping dan hierarchical.




Sumber : http://etheses.uin-malang.ac.id/7360/1/06550032.pdf

NNnan

Nama : Nanda Tri Haryati
NIM/Kelas : 202420016/MTI23-REG-A

Tugas 05

Pelari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format ms word dan dilengkapi dengan gambar dan keterangan yang jelas untuk tiap langkah-langkahnya. Sertkan sumber sumber rujukan yang anda gunakan. Kumpulkan sebelum waktu batas akhir pengumpulan tugas.

Jawaban :

Pada tugas ini saya coba mengambil sample untuk mengukur performance Clustering dengan menggunakan K-Means, pengukuran performance clustering saat ini menggunakan clusterisasi sampai didapatkan nilai stabil dan dengan rapidminer saya coba untuk bandingkan hasilnya dengan sumber data yang sama, adapun datanya sebagai berikut :

Data Komsumsi Listrik bulan Mei 2020 di Prov Lampung :

| | Kabupaten/Kota | Latitude | Longitude | Mei 2020 | | |
|---|---|---|---|---|---|---|
| | | | | Rumah Tangga | Industri | Bisnis |
| | Kabupaten Tulang Bawang Barat | -4.52569 | 105.07912 | 8647851 | 1424419 | 771600 |
| | Kabupaten Way Kanan | -4.49636 | 104.56552 | 8922355 | 904997 | 580164 |
| | Kabupaten Lampung Barat | -5.10952 | 104.1466 | 7870512 | 908299 | 1508881 |
| | Kabupaten Lampung Selatan | -5.56226 | 105.54743 | 39811088 | 28596132 | 6923204 |
| | Kabupaten Lampung Tengah | -4.8008 | 105.31311 | 43819603 | 17071293 | 5099183 |
| | Kabupaten Lampung Timur | -5.11349 | 105.68817 | 17150724 | 4651063 | 3006046 |
| | Kabupaten Lampung Utara | -4.81339 | 104.75209 | 21682207 | 3629202 | 2188612 |
| | Kabupaten Pringsewu | -5.33311 | 104.98561 | 17504659 | 701278 | 1618426 |
| | Kabupaten Tanggamus | -5.30274 | 104.56552 | 14831519 | 2175713 | 1911240 |
| | Kabupaten Tulang Bawang | -4.31765 | 105.50054 | 13723285 | 1574286 | 1730416 |
| | Kota Bandar Lampung | -5.39713 | 105.26678 | 60305690 | 7951584 | 18855425 |
| | Kota Metro | -5.11783 | 105.30726 | 17774258 | 753564 | 2565855 |

Dari data diatas saya coba untuk mengukur bagaimana performance dari clustering yang menggunakan K-Means dengan membandingkan 2 cara yaitu dengan Coding PhP dan Rapidminer.

NNnan

Nama : Nanda Tri Haryati
NIM/Kelas : 202420016/MTI23-REG-A

- Berbasiskan

  Web Stage

  Awal Data

Proses Iterasi Selanjutnya

| | Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rumah Tangga | Industri | Bisnis | 4035984 | 396909 | 14840900 | 396909 | 4035984 | 14840900 | 14840900 | 396909 | 4035984 | | | |
| | Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 14841505 | | | 16517975 | | | 7075721 | | | 0 | 0 | 1 |
| | Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 15083214 | | | 16907244 | | | 6872410 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 13881925 | | | 15600637 | | | 7431963 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 46235745 | | | 47110171 | | | 37776216 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 44222985 | | | 46371873 | | | 33450421 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 18170250 | | | 20521510 | | | 4949129 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 21952614 | | | 24765091 | | | 7788708 | | | 0 | 0 | 1 |
| | Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 18876749 | | | 21877595 | | | 3610102 | | | 0 | 0 | 1 |
| | Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 16937645 | | | 19467785 | | | 2771059 | | | 0 | 0 | 1 |
| | Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 16343647 | | | 18855689 | | | 2819738 | | | 0 | 0 | 1 |
| | Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 56916336 | | | 60170677 | | | 48412148 | | | 0 | 0 | 1 |
| | Kota Metro | 17774258 | 753564 | 2565855 | 18426722 | | | 21527268 | | | 3300465 | | | 0 | 0 | 1 |

Iterasi 1

**Iterasi Ke-1**

| | Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rumah Tangga | Industri | Bisnis | 0 | 0 | 0 | 0 | 0 | 0 | 22670312 | 5861819 | 3896587 | | | |
| | Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 8798276 | | | 8798276 | | | 15036139 | | | 1 | 1 | 0 |
| | Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 8986880 | | | 8986880 | | | 14985829 | | | 1 | 1 | 0 |
| | Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 8065152 | | | 8065152 | | | 15788369 | | | 1 | 1 | 0 |
| | Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 49503456 | | | 49503456 | | | 28632421 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 47303153 | | | 47303153 | | | 23966456 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 18022653 | | | 18022653 | | | 5720564 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 22092515 | | | 22092515 | | | 2979616 | | | 0 | 0 | 1 |
| | Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 17593299 | | | 17593299 | | | 7648868 | | | 0 | 0 | 1 |
| | Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 15111602 | | | 15111602 | | | 8886825 | | | 0 | 0 | 1 |
| | Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 13921252 | | | 13921252 | | | 10155025 | | | 0 | 0 | 1 |
| | Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 63683051 | | | 63683051 | | | 40553120 | | | 0 | 0 | 1 |
| | Kota Metro | 17774258 | 753564 | 2565855 | 17974307 | | | 17974307 | | | 7199754 | | | 0 | 0 | 1 |

Nama : Nanda Tri Haryati
NIM/Kelas : 202420016/MTI23-REG-A

Iterasi 2

**Iterasi Ke-2**

| Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rumah Tangga | Industri | Bisnis | 8480239 | 1079238 | 953548 | 8480239 | 1079238 | 953548 | 27400337 | 7456012 | 4877600 | | | |
| Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 424674 | | | 424674 | | | 20122005 | | | 1 | 1 | 0 |
| Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 604352 | | | 604352 | | | 20070365 | | | 1 | 1 | 0 |
| Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 842248 | | | 842248 | | | 20871868 | | | 1 | 1 | 0 |
| Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 42124082 | | | 42124082 | | | 24599103 | | | 0 | 0 | 1 |
| Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 39010289 | | | 39010289 | | | 19028794 | | | 0 | 0 | 1 |
| Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 9599374 | | | 9599374 | | | 10790042 | | | 1 | 1 | 0 |
| Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 13502579 | | | 13502579 | | | 7387296 | | | 0 | 0 | 1 |
| Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 9056769 | | | 9056769 | | | 12416646 | | | 1 | 1 | 0 |
| Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 6515994 | | | 6515994 | | | 13951919 | | | 1 | 1 | 0 |
| Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 5323356 | | | 5323356 | | | 15217136 | | | 1 | 1 | 0 |
| Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 55259240 | | | 55259240 | | | 35754544 | | | 0 | 0 | 1 |
| Kota Metro | 17774258 | 753564 | 2565855 | 9438452 | | | 9438452 | | | 11955265 | | | 1 | 1 | 0 |

Iterasi 3

**Iterasi Ke-3**

| Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rumah Tangga | Industri | Bisnis | 13303145 | 1636702 | 1711578 | 13303145 | 1636702 | 1711578 | 41404647 | 14312052 | 8266606 | | | |
| Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 4753986 | | | 4753986 | | | 35989913 | | | 1 | 1 | 0 |
| Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 4583319 | | | 4583319 | | | 35971236 | | | 1 | 1 | 0 |
| Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 5484994 | | | 5484994 | | | 36740518 | | | 1 | 1 | 0 |
| Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 38165991 | | | 38165991 | | | 14435341 | | | 0 | 0 | 1 |
| Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 34365049 | | | 34365049 | | | 4845408 | | | 0 | 0 | 1 |
| Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 5056271 | | | 5056271 | | | 26631954 | | | 1 | 1 | 0 |
| Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 8625908 | | | 8625908 | | | 23238759 | | | 1 | 1 | 0 |
| Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 4305393 | | | 4305393 | | | 28295952 | | | 1 | 1 | 0 |
| Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 1632887 | | | 1632887 | | | 29896697 | | | 1 | 1 | 0 |
| Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 425168 | | | 425168 | | | 31164567 | | | 1 | 1 | 0 |
| Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 50428448 | | | 50428448 | | | 22579372 | | | 0 | 0 | 1 |
| Kota Metro | 17774258 | 753564 | 2565855 | 4636870 | | | 4636870 | | | 27833908 | | | 1 | 1 | 0 |

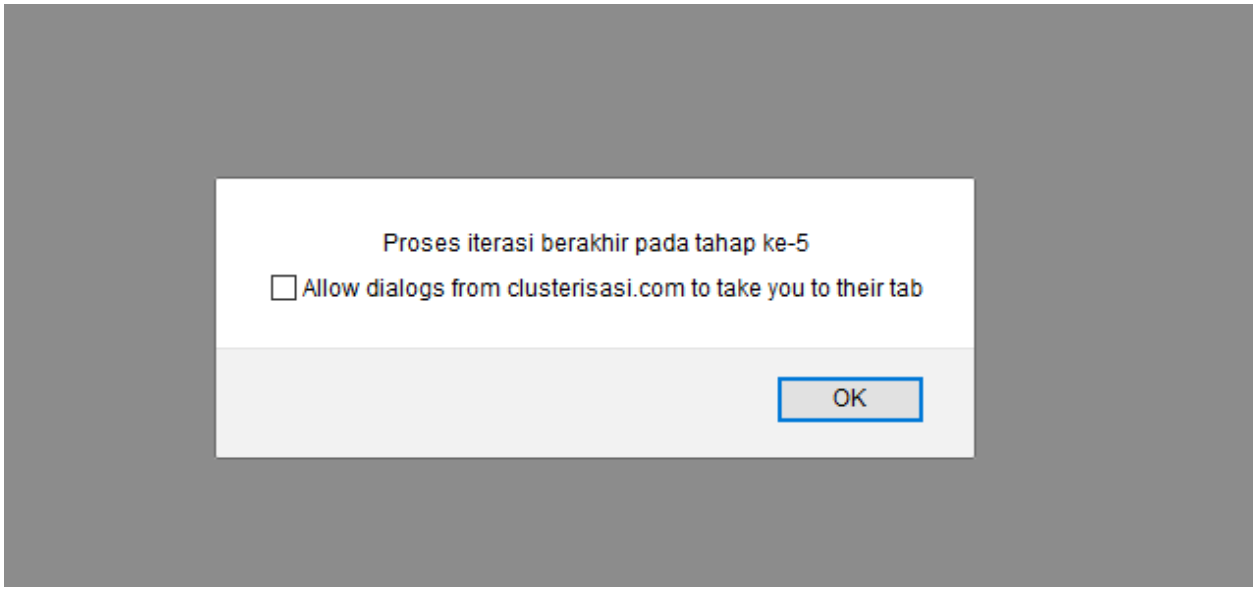
Nama : Nanda Tri Haryati
NIM/Kelas : 202420016/MTI23-REG-A

Iterasi 4

**Iterasi Ke-4**

| Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rumah Tangga | Industri | Bisnis | 14234152 | 1858091 | 1764582 | 14234152 | 1858091 | 1764582 | 47978793 | 17873003 | 10292604 | | | |
| Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 5690417 | | | 5690417 | | | 43682130 | | | 1 | 1 | 0 |
| Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 5525072 | | | 5525072 | | | 43676654 | | | 1 | 1 | 0 |
| Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 6439208 | | | 6439208 | | | 44425547 | | | 1 | 1 | 0 |
| Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 37359253 | | | 37359253 | | | 13894235 | | | 0 | 0 | 1 |
| Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 33434412 | | | 33434412 | | | 6701733 | | | 0 | 0 | 1 |
| Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 4224726 | | | 4224726 | | | 34326134 | | | 1 | 1 | 0 |
| Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 7667473 | | | 7667473 | | | 30985012 | | | 1 | 1 | 0 |
| Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 3472145 | | | 3472145 | | | 36038623 | | | 1 | 1 | 0 |
| Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 692271 | | | 692271 | | | 37621722 | | | 1 | 1 | 0 |
| Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 585404 | | | 585404 | | | 38889575 | | | 1 | 1 | 0 |
| Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 49515797 | | | 49515797 | | | 17991910 | | | 0 | 0 | 1 |
| Kota Metro | 17774258 | 753564 | 2565855 | 3793990 | | | 3793990 | | | 35568129 | | | 1 | 1 | 0 |

Iterasi 5




Pada saat iterasi ke 5 data sudah didapatkan sampai dengan semua nilainya stabil atau sama, sehingga didapat nilai cluster yang sesuai sebagai berikut :

Iterasi ke-1

| C1 | C2 | C3 |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |

Nama : Nanda Tri Haryati
NIM/Kelas : 202420016/MTI23-REG-A

Iterasi ke-2

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |

Iterasi ke-3

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |

Nama : Nanda Tri Haryati
NIM/Kelas : 202420016/MTI23-REG-A

Iterasi ke-4

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |

Iterasi ke-5

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |

Nama : Nanda Tri Haryati
NIM/Kelas : 202420016/MTI23-REG-A

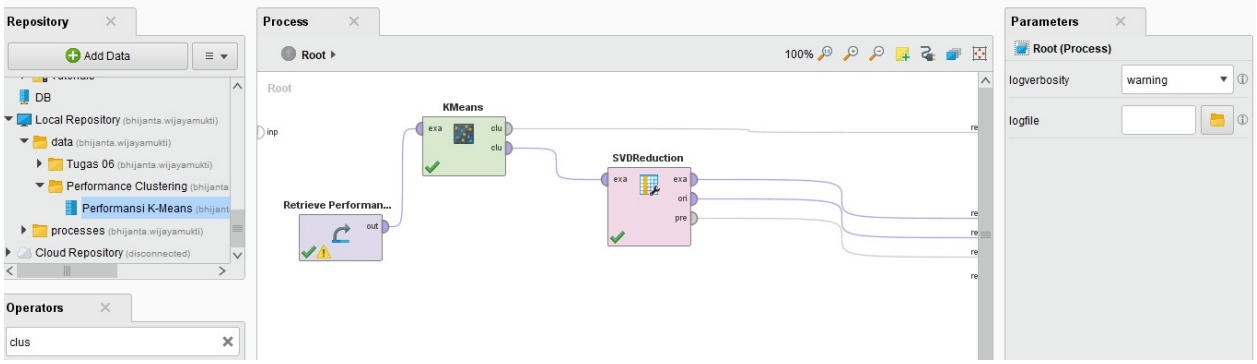
Dari Hasil Iterasi ke 5 didaptkan kesimpulan sebagai berikut :

| | Kabupaten/Kota | Bulan | Zona |
|---|---|---|---|
| | Kabupaten Tulang Bawang Barat | Mei | Kuning |
| | Kabupaten Way Kanan | Mei | Kuning |
| | Kabupaten Lampung Barat | Mei | Kuning |
| | Kabupaten Lampung Selatan | Mei | Hijau |
| | Kabupaten Lampung Tengah | Mei | Hijau |
| | Kabupaten Lampung Timur | Mei | Kuning |
| | Kabupaten Lampung Utara | Mei | Kuning |
| | Kabupaten Pringsewu | Mei | Kuning |
| | Kabupaten Tanggamus | Mei | Kuning |
| | Kabupaten Tulang Bawang | Mei | Kuning |
| | Kota Bandar Lampung | Mei | Hijau |
| | Kota Metro | Mei | Kuning |

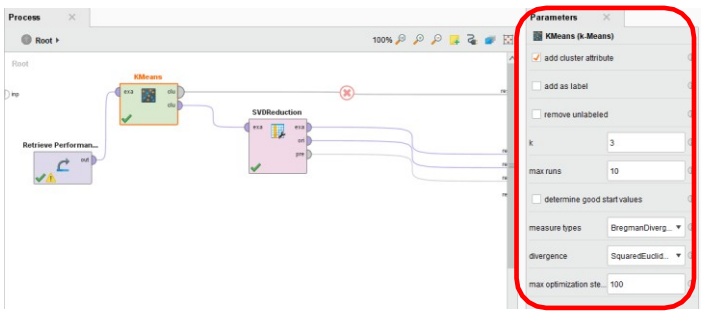
- Berbasiskan Rapidminer

  Dengan data yang sama seperti diatas saya coba mengukur bagaimana performance dari Clustering K-Means menggunakan tool rapidminer, adapun hasilnya sebagai berikut :

  Desain Mapping Rapidminer

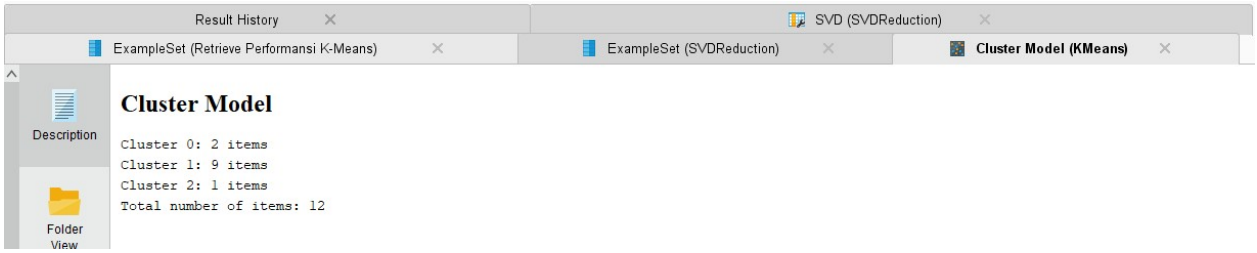


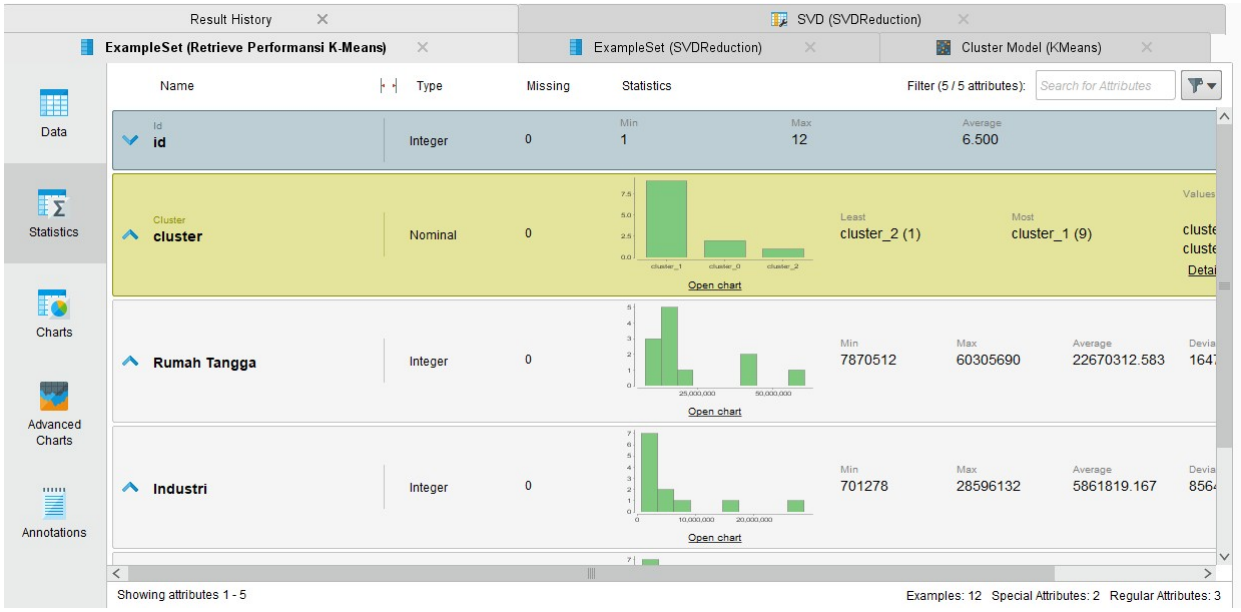

Parameter K-Means

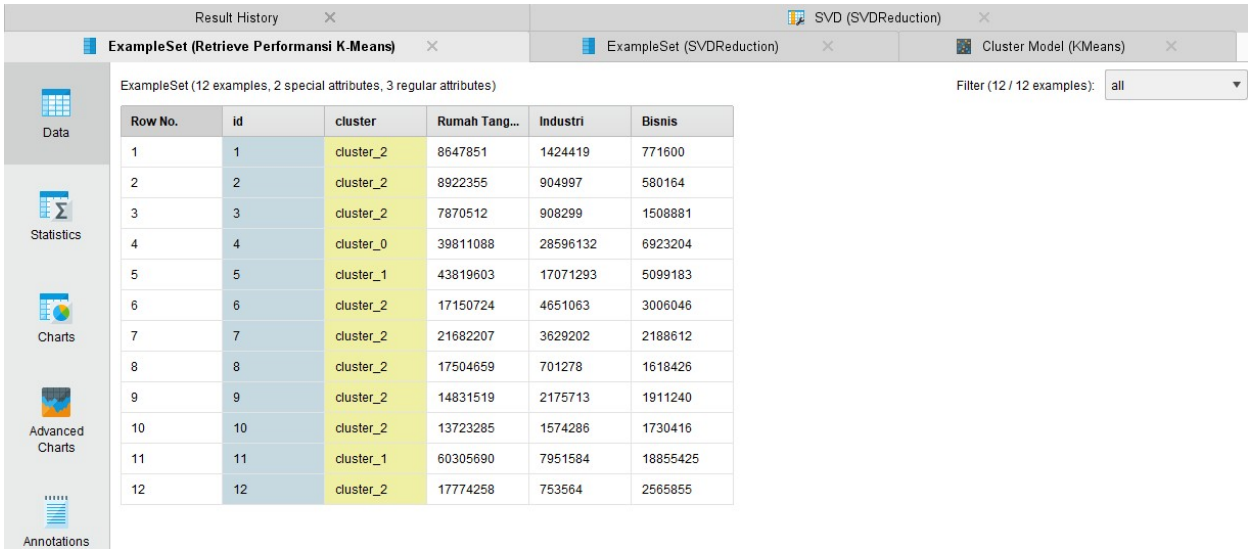


Nama : Nanda Tri Haryati
NIM/Kelas : 202420016/MTI23-REG-A

# Hasil Clustering

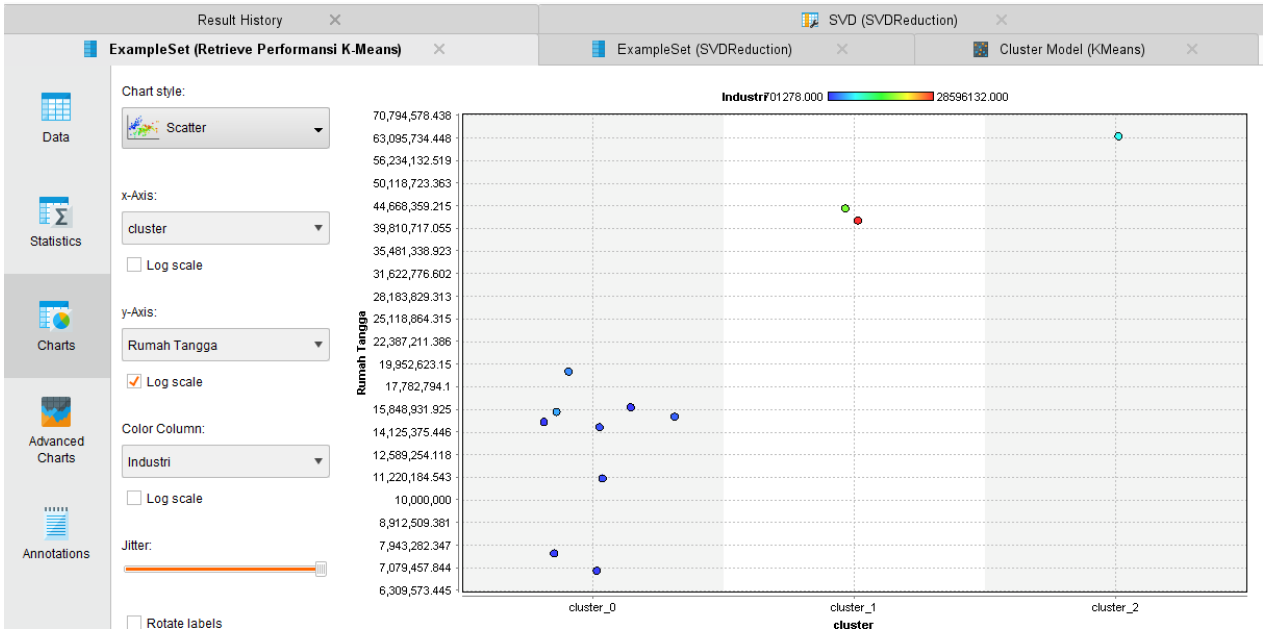



| Row No. | id | cluster | Rumah Tang... | Industri | Bisnis |
|---|---|---|---|---|---|
| 1 | 1 | cluster_2 | 8647851 | 1424419 | 771600 |
| 2 | 2 | cluster_2 | 8922355 | 904997 | 580164 |
| 3 | 3 | cluster_2 | 7870512 | 908299 | 1508881 |
| 4 | 4 | cluster_0 | 39811088 | 28596132 | 6923204 |
| 5 | 5 | cluster_1 | 43819603 | 17071293 | 5099183 |
| 6 | 6 | cluster_2 | 17150724 | 4651063 | 3006046 |
| 7 | 7 | cluster_2 | 21682207 | 3629202 | 2188612 |
| 8 | 8 | cluster_2 | 17504659 | 701278 | 1618426 |
| 9 | 9 | cluster_2 | 14831519 | 2175713 | 1911240 |
| 10 | 10 | cluster_2 | 13723285 | 1574286 | 1730416 |
| 11 | 11 | cluster_1 | 60305690 | 7951584 | 18855425 |
| 12 | 12 | cluster_2 | 17774258 | 753564 | 2565855 |

Nama : Nanda Tri Haryati
NIM/Kelas : 202420016/MTI23-REG-A




Kesimpulan

*Sumber : https://clusterisasi.com/k-means/admin/iterasi_kmeans_hasil*

1. K-means merupakan salah satu algoritma clustering. Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan supervised learning yang menerima masukan berupa vektor (-x-1 , y1) , (-x-2 , y2) , …, (-x-i , yi), di mana xi merupakan data dari suatu data pelatihan dan yi merupakan label kelas untuk xi
2. Kelebihan pada algoritma k-means, yaitu [2]:
    - Mudah untuk diimplementasikan dan dijalankan.
    - Waktu yang dibutuhkan untuk menjalankan pembelajaran ini relatif cepat.
    - Mudah untuk diadaptasi.
    - Umum digunakan.
3. Secara performance dari Clustering dengan K-Means cukup dengan baik dapat menyelesaikan clustering baik secara metode Web dan Rapidminer.

**NAMA** : **OMAN ARROHMAN**

**MATA KULIAH** : **ADVANCED DATABASE**

**NIM** : **202420042**

## TUGAS 5

Pelajari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format word dan dilengkapi dengan gambar dan keterangan yang jelas untuk tiap langkah-langkahnya. Sertakan sumber-sumber rujukan yang anda gunakan

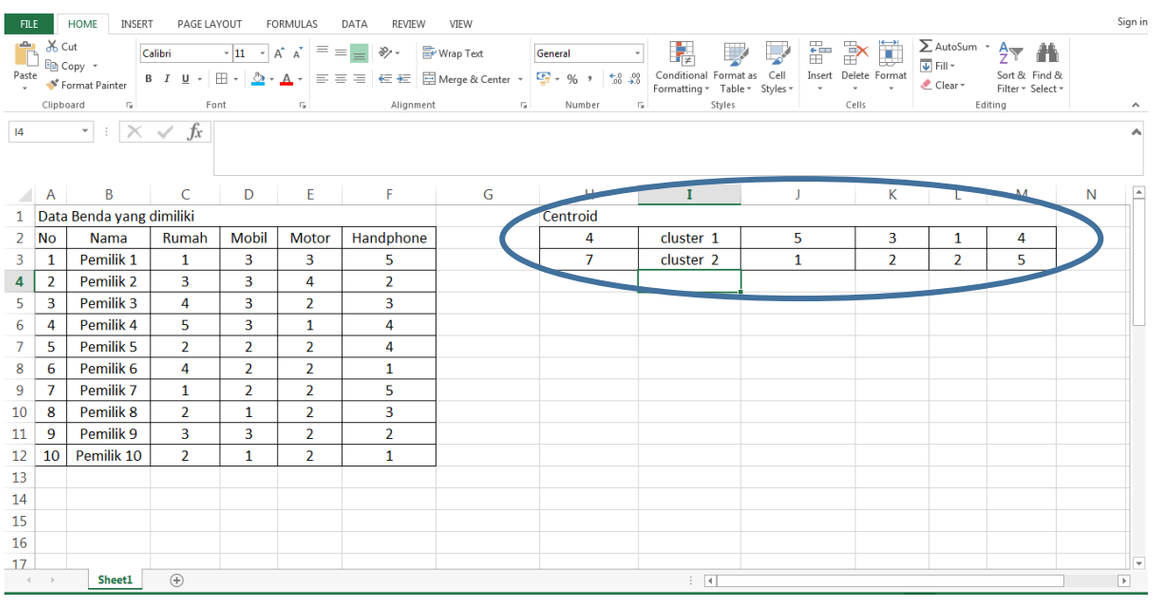
Langkah-langkah yang harus dilakukan :

1. Siapakan data yang akan diolah untuk mengukur clustering dari cluster 1, cluster 2, jarak terdekat dan kelompok data. Contoh seperti pada tabel dibawah ini, perhitungan akan menggunakan Ms.Excel

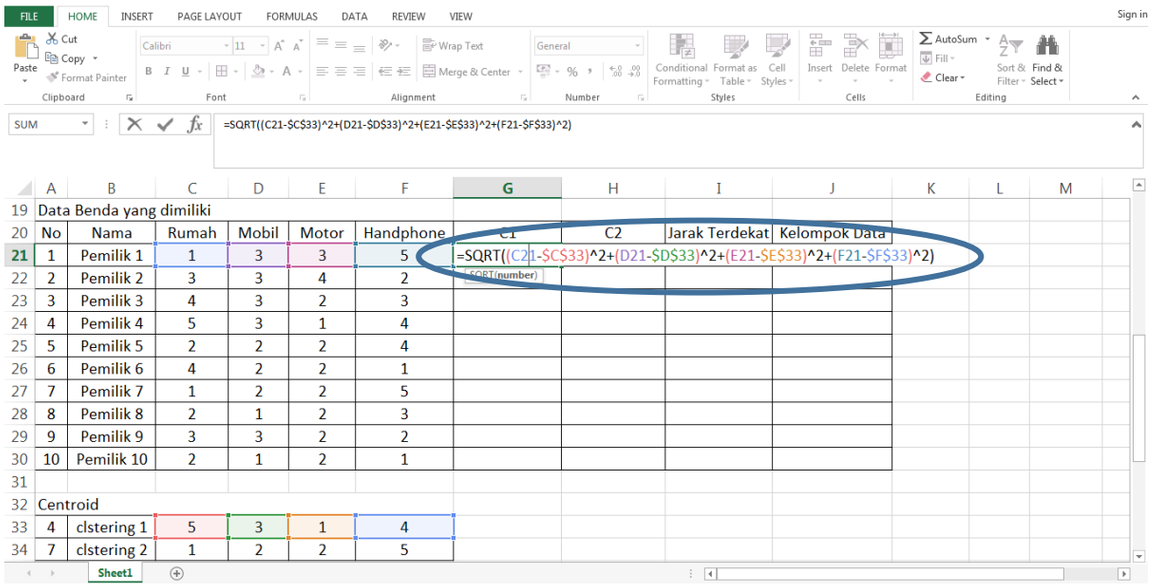
Data Benda yang dimiliki

| No | Nama | Rumah | Mobil | Motor | Handphone |
|---|---|---|---|---|---|
| 1 | Pemilik 1 | 1 | 3 | 3 | 5 |
| 2 | Pemilik 2 | 3 | 3 | 4 | 2 |
| 3 | Pemilik 3 | 4 | 3 | 2 | 3 |
| 4 | Pemilik 4 | 5 | 3 | 1 | 4 |
| 5 | Pemilik 5 | 2 | 2 | 2 | 4 |
| 6 | Pemilik 6 | 4 | 2 | 2 | 1 |
| 7 | Pemilik 7 | 1 | 2 | 2 | 5 |
| 8 | Pemilik 8 | 2 | 1 | 2 | 3 |
| 9 | Pemilik 9 | 3 | 3 | 2 | 2 |
| 10 | Pemilik 10 | 2 | 1 | 2 | 1 |

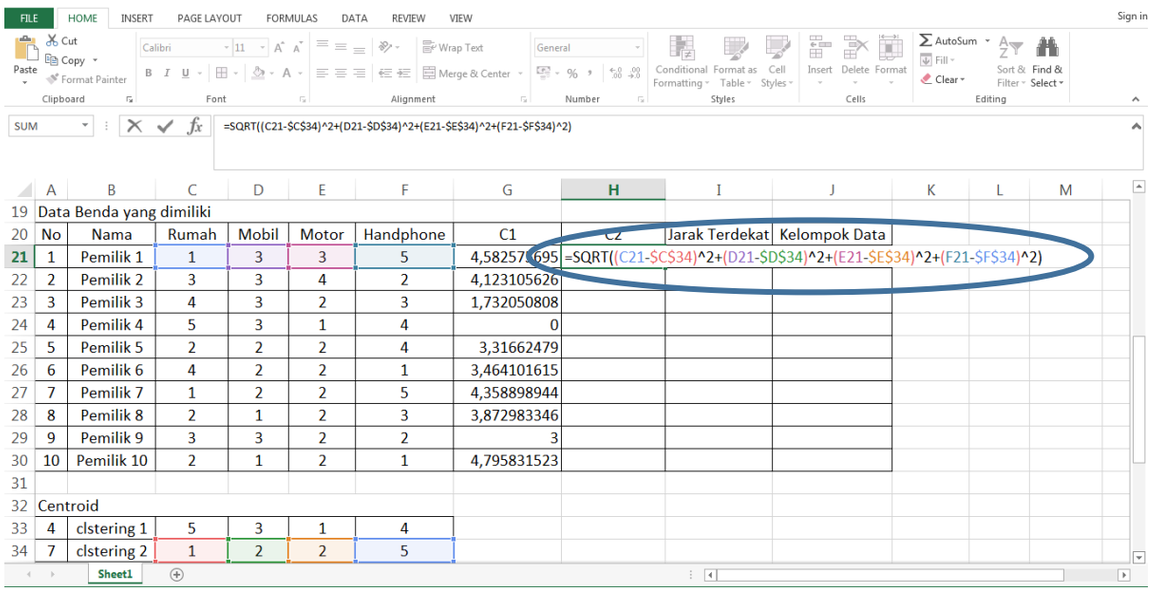
2. Untuk mengukur clustering diperlukan centroid atau data pusat yang terlebih dahulu harus ditentukan, pada pengukuran ini centroid diambil dari data 4 dan 7



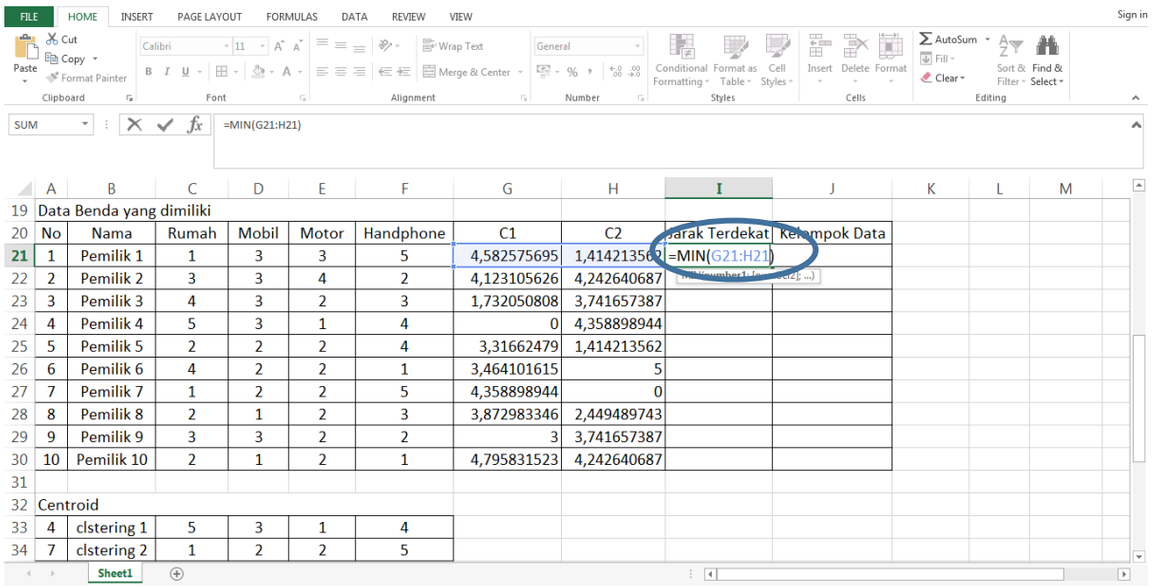
3. Selanjutnya menghitung nilai cluster 1 dengan menggunakan rumus seperti pada gambar dengan menghitung nilai pada tabel data benda dan centroid cluster 1




4. Langkah berikutnya menghitung nilai cluster 2 dengan menggunakan rumus sama seperti langkah ke 3 tapi dengan nilai yang berbeda pada tabel data benda dan centroid cluster 2

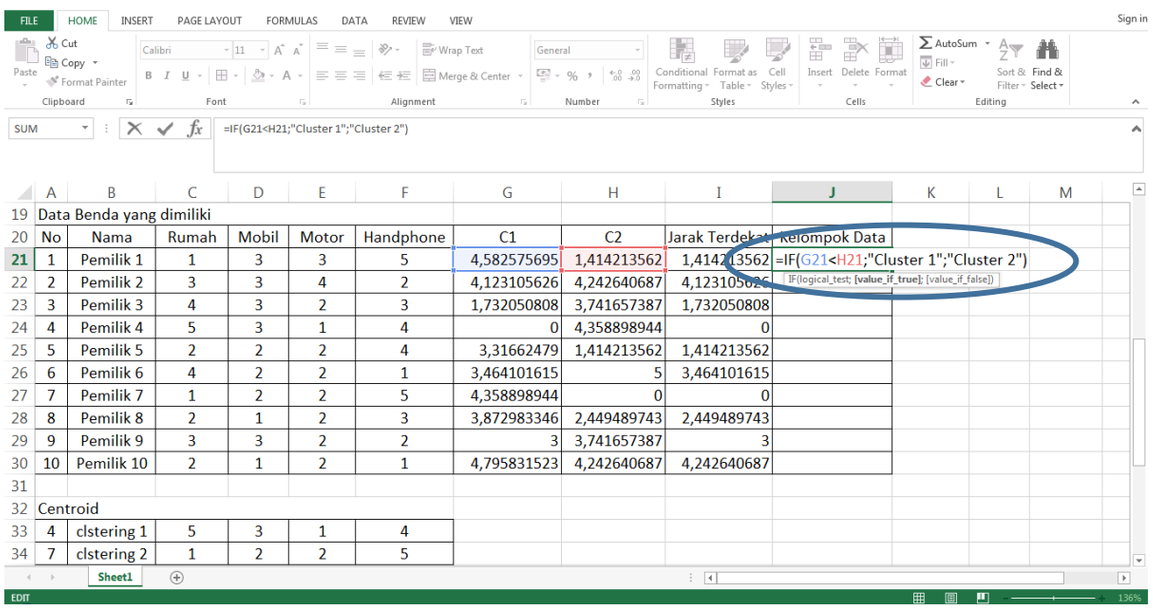


5. Berikutnya menghitung jarak terdekat dari cluster dengan rumus seperti pada gambar yaitu dari nilai cluster 1 dan cluster 2 lalu didapat jarak terdekat



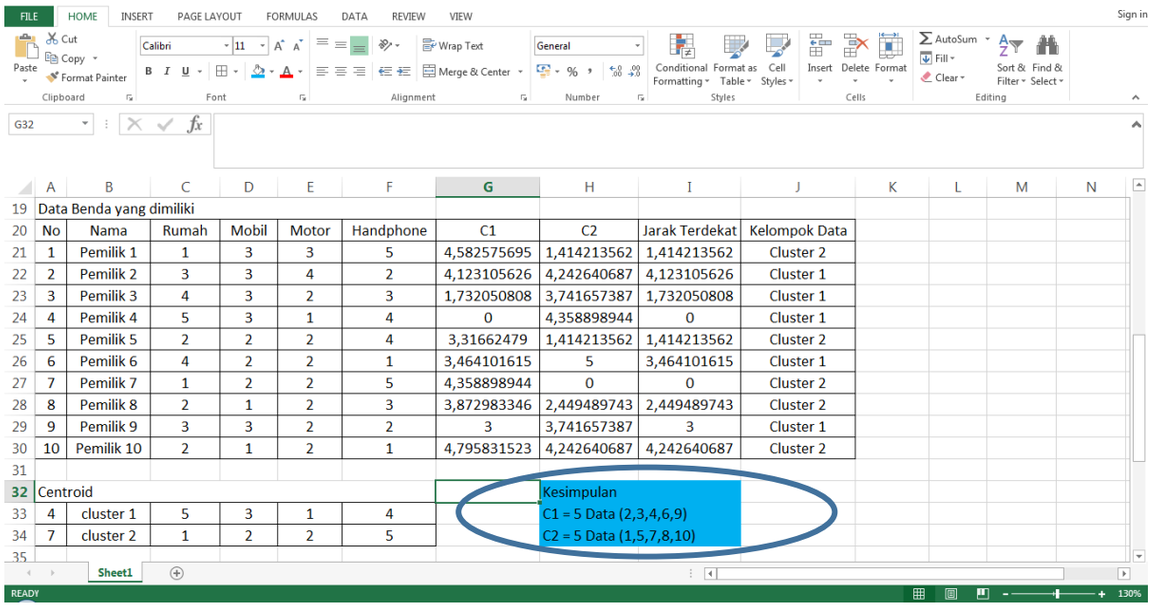
| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 19 | Data Benda yang dimiliki | | | | | | | | | |
| 20 | No | Nama | Rumah | Mobil | Motor | Handphone | C1 | C2 | Jarak Terdekat | Kelompok Data |
| 21 | 1 | Pemilik 1 | 1 | 3 | 3 | 5 | 4,582575695 | 1,414213562 | =MIN(G21:H21) | |
| 22 | 2 | Pemilik 2 | 3 | 3 | 4 | 2 | 4,123105626 | 4,242640687 | | |
| 23 | 3 | Pemilik 3 | 4 | 3 | 2 | 3 | 1,732050808 | 3,741657387 | | |
| 24 | 4 | Pemilik 4 | 5 | 3 | 1 | 4 | 0 | 4,358898944 | | |
| 25 | 5 | Pemilik 5 | 2 | 2 | 2 | 4 | 3,31662479 | 1,414213562 | | |
| 26 | 6 | Pemilik 6 | 4 | 2 | 2 | 1 | 3,464101615 | 5 | | |
| 27 | 7 | Pemilik 7 | 1 | 2 | 2 | 5 | 4,358898944 | 0 | | |
| 28 | 8 | Pemilik 8 | 2 | 1 | 2 | 3 | 3,872983346 | 2,449489743 | | |
| 29 | 9 | Pemilik 9 | 3 | 3 | 2 | 2 | 3 | 3,741657387 | | |
| 30 | 10 | Pemilik 10 | 2 | 1 | 2 | 1 | 4,795831523 | 4,242640687 | | |
| 31 | | | | | | | | | | |
| 32 | Centroid | | | | | | | | | |
| 33 | 4 | clstering 1 | 5 | 3 | 1 | 4 | | | | |
| 34 | 7 | clstering 2 | 1 | 2 | 2 | 5 | | | | |

6. Setelah didapatkan jarak terdekat, langkah selanjutnya adalah mengelompokkan data menjadi cluster 1 atau cluster 2



| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 19 | Data Benda yang dimiliki | | | | | | | | | |
| 20 | No | Nama | Rumah | Mobil | Motor | Handphone | C1 | C2 | Jarak Terdekat | Kelompok Data |
| 21 | 1 | Pemilik 1 | 1 | 3 | 3 | 5 | 4,582575695 | 1,414213562 | 1,414213562 | =IF(G21<H21;"Cluster 1";"Cluster 2") |
| 22 | 2 | Pemilik 2 | 3 | 3 | 4 | 2 | 4,123105626 | 4,242640687 | 4,123105626 | |
| 23 | 3 | Pemilik 3 | 4 | 3 | 2 | 3 | 1,732050808 | 3,741657387 | 1,732050808 | |
| 24 | 4 | Pemilik 4 | 5 | 3 | 1 | 4 | 0 | 4,358898944 | 0 | |
| 25 | 5 | Pemilik 5 | 2 | 2 | 2 | 4 | 3,31662479 | 1,414213562 | 1,414213562 | |
| 26 | 6 | Pemilik 6 | 4 | 2 | 2 | 1 | 3,464101615 | 5 | 3,464101615 | |
| 27 | 7 | Pemilik 7 | 1 | 2 | 2 | 5 | 4,358898944 | 0 | 0 | |
| 28 | 8 | Pemilik 8 | 2 | 1 | 2 | 3 | 3,872983346 | 2,449489743 | 2,449489743 | |
| 29 | 9 | Pemilik 9 | 3 | 3 | 2 | 2 | 3 | 3,741657387 | 3 | |
| 30 | 10 | Pemilik 10 | 2 | 1 | 2 | 1 | 4,795831523 | 4,242640687 | 4,242640687 | |
| 31 | | | | | | | | | | |
| 32 | Centroid | | | | | | | | | |
| 33 | 4 | clstering 1 | 5 | 3 | 1 | 4 | | | | |
| 34 | 7 | clstering 2 | 1 | 2 | 2 | 5 | | | | |

7. Dari pengelompokan data diperoleh kesimpulan yaitu pada cluster 1 memiliki 5 data dan cluster 2 memiliki 5 data



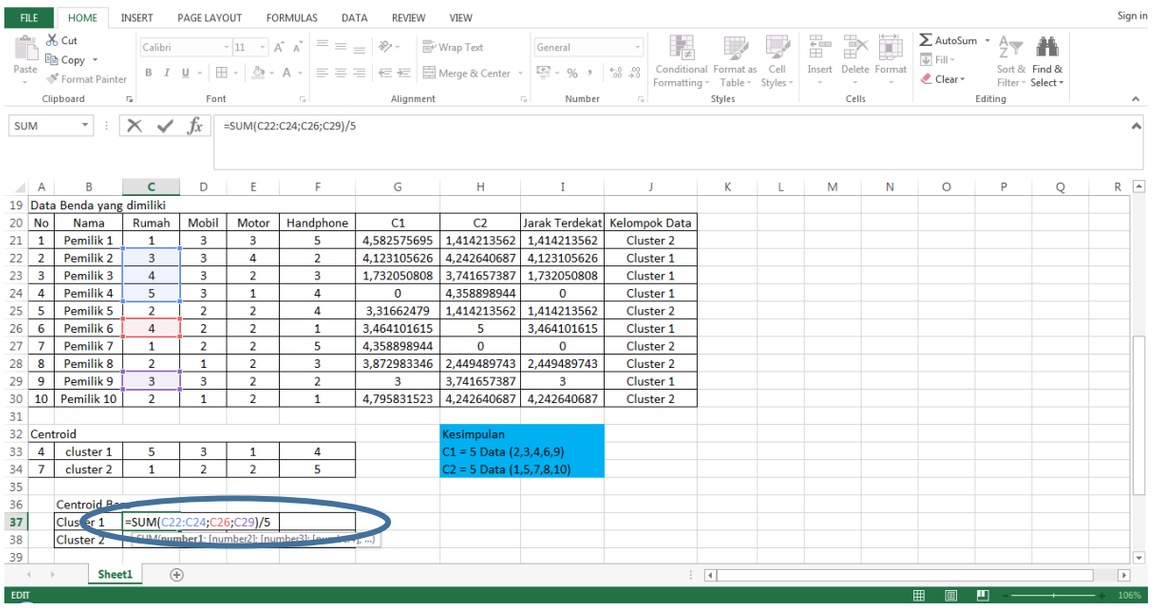
| No | Nama | Rumah | Mobil | Motor | Handphone | C1 | C2 | Jarak Terdekat | Kelompok Data |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Pemilik 1 | 1 | 3 | 3 | 5 | 4,582575695 | 1,414213562 | 1,414213562 | Cluster 2 |
| 2 | Pemilik 2 | 3 | 3 | 4 | 2 | 4,123105626 | 4,242640687 | 4,123105626 | Cluster 1 |
| 3 | Pemilik 3 | 4 | 3 | 2 | 3 | 1,732050808 | 3,741657387 | 1,732050808 | Cluster 1 |
| 4 | Pemilik 4 | 5 | 3 | 1 | 4 | 0 | 4,358898944 | 0 | Cluster 1 |
| 5 | Pemilik 5 | 2 | 2 | 2 | 4 | 3,31662479 | 1,414213562 | 1,414213562 | Cluster 2 |
| 6 | Pemilik 6 | 4 | 2 | 2 | 1 | 3,464101615 | 5 | 3,464101615 | Cluster 1 |
| 7 | Pemilik 7 | 1 | 2 | 2 | 5 | 4,358898944 | 0 | 0 | Cluster 2 |
| 8 | Pemilik 8 | 2 | 1 | 2 | 3 | 3,872983346 | 2,449489743 | 2,449489743 | Cluster 2 |
| 9 | Pemilik 9 | 3 | 3 | 2 | 2 | 3 | 3,741657387 | 3 | Cluster 1 |
| 10 | Pemilik 10 | 2 | 1 | 2 | 1 | 4,795831523 | 4,242640687 | 4,242640687 | Cluster 2 |

Centroid

| | | | | | |
|---|---|---|---|---|---|
| 4 | cluster 1 | 5 | 3 | 1 | 4 |
| 7 | cluster 2 | 1 | 2 | 2 | 5 |

Kesimpulan
C1 = 5 Data (2,3,4,6,9)
C2 = 5 Data (1,5,7,8,10)

8. Dari kesimpulan yang didapat selanjunya adalah membuat interasi baru tapi sebelum itu kita harus menentukan centroid yang baru terlebih dahulu dengan menghitung jumlah data pada kesimpulan cluster 1 seluruh nilai yang menjadi cluster 1 kan dibagi jumlah data yaitu 5, begitu pula cluster 2 dicari dengan cara yang sama



| No | Nama | Rumah | Mobil | Motor | Handphone | C1 | C2 | Jarak Terdekat | Kelompok Data |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Pemilik 1 | 1 | 3 | 3 | 5 | 4,582575695 | 1,414213562 | 1,414213562 | Cluster 2 |
| 2 | Pemilik 2 | 3 | 3 | 4 | 2 | 4,123105626 | 4,242640687 | 4,123105626 | Cluster 1 |
| 3 | Pemilik 3 | 4 | 3 | 2 | 3 | 1,732050808 | 3,741657387 | 1,732050808 | Cluster 1 |
| 4 | Pemilik 4 | 5 | 3 | 1 | 4 | 0 | 4,358898944 | 0 | Cluster 1 |
| 5 | Pemilik 5 | 2 | 2 | 2 | 4 | 3,31662479 | 1,414213562 | 1,414213562 | Cluster 2 |
| 6 | Pemilik 6 | 4 | 2 | 2 | 1 | 3,464101615 | 5 | 3,464101615 | Cluster 1 |
| 7 | Pemilik 7 | 1 | 2 | 2 | 5 | 4,358898944 | 0 | 0 | Cluster 2 |
| 8 | Pemilik 8 | 2 | 1 | 2 | 3 | 3,872983346 | 2,449489743 | 2,449489743 | Cluster 2 |
| 9 | Pemilik 9 | 3 | 3 | 2 | 2 | 3 | 3,741657387 | 3 | Cluster 1 |
| 10 | Pemilik 10 | 2 | 1 | 2 | 1 | 4,795831523 | 4,242640687 | 4,242640687 | Cluster 2 |

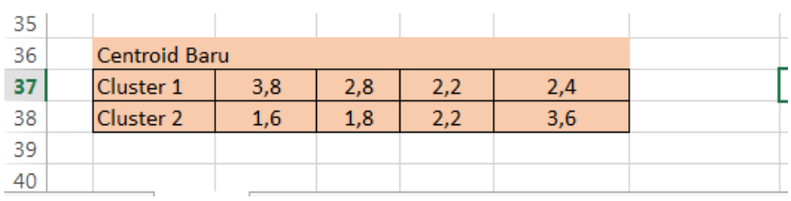
Centroid

| | | | | | |
|---|---|---|---|---|---|
| 4 | cluster 1 | 5 | 3 | 1 | 4 |
| 7 | cluster 2 | 1 | 2 | 2 | 5 |

Kesimpulan
C1 = 5 Data (2,3,4,6,9)
C2 = 5 Data (1,5,7,8,10)

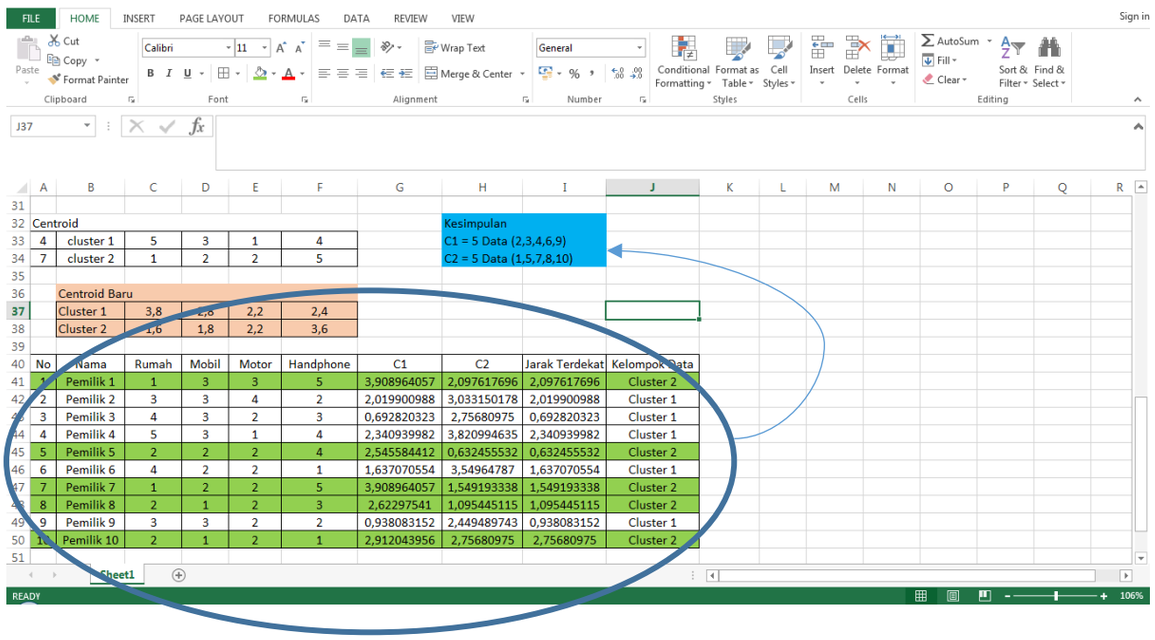
Centroid Ba
Cluster 1 =SUM(C22:C24;C26;C29)/5
Cluster 2

9. Setelah mencari nilai centroid baru maka akan diperoleh nilai centroid baru seperti pada gambar dibawah



| Centroid Baru | | | | |
|---|---|---|---|---|
| Cluster 1 | 3,8 | 2,8 | 2,2 | 2,4 |
| Cluster 2 | 1,6 | 1,8 | 2,2 | 3,6 |

10. Dari data centroid yang baru akan dilakukan kembali perhitungan untuk iterasi ke 2 dengan mengulang langkah yang sama seperti langka k3 hingga langkah ke 7 dengan hasil seperti pada gambar



| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 31 | | | | | | | | | | |
| 32 | Centroid | | | | | | Kesimpulan | | | |
| 33 | 4 | cluster 1 | 5 | 3 | 1 | 4 | C1 = 5 Data (2,3,4,6,9) | | | |
| 34 | 7 | cluster 2 | 1 | 2 | 2 | 5 | C2 = 5 Data (1,5,7,8,10) | | | |
| 35 | | | | | | | | | | |
| 36 | | Centroid Baru | | | | | | | | |
| 37 | | Cluster 1 | 3,8 | 2,8 | 2,2 | 2,4 | | | | |
| 38 | | Cluster 2 | 1,6 | 1,8 | 2,2 | 3,6 | | | | |
| 39 | | | | | | | | | | |
| 40 | No | Nama | Rumah | Mobil | Motor | Handphone | C1 | C2 | Jarak Terdekat | Kelompok Data |
| 41 | 1 | Pemilik 1 | 1 | 3 | 3 | 5 | 3,908964057 | 2,097617696 | 2,097617696 | Cluster 2 |
| 42 | 2 | Pemilik 2 | 3 | 3 | 4 | 2 | 2,019900988 | 3,033150178 | 2,019900988 | Cluster 1 |
| 43 | 3 | Pemilik 3 | 4 | 3 | 2 | 3 | 0,692820323 | 2,75680975 | 0,692820323 | Cluster 1 |
| 44 | 4 | Pemilik 4 | 5 | 3 | 1 | 4 | 2,340939982 | 3,820994635 | 2,340939982 | Cluster 1 |
| 45 | 5 | Pemilik 5 | 2 | 2 | 2 | 4 | 2,545584412 | 0,632455532 | 0,632455532 | Cluster 2 |
| 46 | 6 | Pemilik 6 | 4 | 2 | 2 | 1 | 1,637070554 | 3,54964787 | 1,637070554 | Cluster 1 |
| 47 | 7 | Pemilik 7 | 1 | 2 | 2 | 5 | 3,908964057 | 1,549193338 | 1,549193338 | Cluster 2 |
| 48 | 8 | Pemilik 8 | 2 | 1 | 2 | 3 | 2,62297541 | 1,095445115 | 1,095445115 | Cluster 2 |
| 49 | 9 | Pemilik 9 | 3 | 3 | 2 | 2 | 0,938083152 | 2,449489743 | 0,938083152 | Cluster 1 |
| 50 | 10 | Pemilik 10 | 2 | 1 | 2 | 1 | 2,912043956 | 2,75680975 | 2,75680975 | Cluster 2 |
| 51 | | | | | | | | | | |

11. Setelah selesai dilakukan mengukur clustering pada iterasi ke 2 diperoleh nilai yang sama dengan iterasi yang pertama jadi perhitungan dapat distop atau diselesaikan karena nilai clusternya telah sama.

Sumber :


1. https://youtu.be/8mXL2z3lg_o
2. https://www.youtube.com/watch?v=5o_JQ6AHA0Y
3. https://www.youtube.com/watch?v=MayK_kwJgv4

Nama : Puspita Dewi S
Nim : 202420011
Kelas : MTI 23 Reguler A

## LINK RUJUKAN :

https://youtu.be/0IovbG6E0Go
https://www.youtube.com/watch?v=F5NHk6c94-8

SAMPEL DATA YANG AKAN DIKELOMPOKKAN

| NO | DAERAH | LUAS LAHAN | PRODUKSI |
|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 |

Langkah Langkah Perhitungan
1. Menentukan Jumlah Cluster data
2. Tentukan titik Pusat Cluster
3. menentukan jarak obyek dengan centroid
4. pengelompokan obyek
5. jika kelompok data hasil perhitungan baru sama dengan hasil perhitungan kelompok data baru maka perhitungan selesai

Langkah Penyelesaian

**1. Tentukan cluster atau kategori pusat dengan asumsi (Max, Mid, Min) namun dalam konsepnya dipilih secara acak**

| | | |
|---|---|---|
| Cluster I | 162 | 964 |
| Cluster II | 72 | 387 |
| Cluster III | 31 | 182 |

**2. Penentuan Jarak Terdekat**

**CLUSTER = SQRT(((C8-$B$3)^2)+(D8-$C$3)^2)**

| NO | DAERAH | LUAS | PRODUKSI | C1 | C2 | C3 | JARAK |
|---|---|---|---|---|---|---|---|

| | | LAHAN | | | | | TERPENDEK |
|---|---|---|---|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 570 | 16 | 223 | 16 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 793 | 209 | 0 | 0 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 714 | 130 | 79 | 79 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 684 | 100 | 109 | 100 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 692 | 108 | 101 | 101 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 509 | 77 | 284 | 77 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 584 | 0 | 209 | 0 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 18 | 566 | 775 | 18 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 263 | 322 | 530 | 263 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 643 | 59 | 150 | 59 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 684 | 100 | 109 | 100 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 661 | 77 | 132 | 77 |

3. Pengelompokan Objek

| NO | DAERAH | LUAS LAHAN | PRODUKSI | KELOMPOK |
|---|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 2 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 3 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 3 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 2 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 3 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 2 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 2 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 1 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 1 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 2 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 2 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 2 |

DENGAN KESIMPULAN DATA 1

cluster 1 : sebanyak 2 data

cluster 2 : sebanyak 7 adata

cluster 3 : sebanyak 3 data

| | | | | | cluster baru | | |
|---|---|---|---|---|---|---|---|
| **NO** | **DAERAH** | **LUAS LAHAN** | **PRODUKSI** | **KELOMPOK** | **C1** | **C2** | **C3** |
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 2 | 137,5 | 60,14286 | 43,66666667 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 3 | 826 | 353,1429 | 240,6666667 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 3 | | | |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 2 | | | |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 3 | | | |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 2 | | | |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 2 | | | |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 1 | | | |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 1 | | | |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 2 | | | |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 2 | | | |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 2 | | | |

penentuan cluster baru

| cluster | luas lahan | produksi |
|---|---|---|
| I | 137,5 | 826 |
| II | 60,14286 | 353,1429 |
| III | 43,66667 | 240,6667 |

| **NO** | **DAERAH** | **LUAS LAHAN** | **PRODUKSI** | **C1** | **C2** | **C3** | **JARAK TERPENDEK** |
|---|---|---|---|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 430 | 49 | 163 | 49 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 653 | 174 | 60 | 60 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 574 | 95 | 19 | 19 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 544 | 65 | 49 | 49 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 552 | 73 | 41 | 41 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 369 | 111 | 224 | 111 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 444 | 36 | 149 | 36 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 122 | 602 | 715 | 122 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 122 | 357 | 470 | 122 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 503 | 24 | 90 | 24 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 543 | 64 | 50 | 50 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 521 | 42 | 72 | 42 |

| NO | DAERAH | LUAS LAHAN | PRODUKSI | KELOMPOK |
|---|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 2 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 3 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 3 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 3 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 3 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 2 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 2 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 1 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 1 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 2 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 3 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 2 |

DENGAN KESIMPULAN DATA 2
cluster 1 : sebanyak 2 data
cluster 2 : sebanyak 5 adata
cluster 3 : sebanyak 5 data

| | | | | | cluster baru | | |
|---|---|---|---|---|---|---|---|
| **NO** | **DAERAH** | **LUAS LAHAN** | **PRODUKSI** | **KELOMPOK** | **C1** | **C2** | **C3** |
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 2 | 137,5 | 64,6 | 45,8 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 3 | 826 | 378,6 | 260,2 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 3 | | | |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 3 | | | |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 3 | | | |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 2 | | | |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 2 | | | |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 1 | | | |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 1 | | | |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 2 | | | |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 3 | | | |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 2 | | | |

penentuan cluster baru

| cluster | luas lahan | produksi |
|---|---|---|
| I | 137,5 | 826 |
| II | 64,6 | 378,6 |
| III | 45,8 | 260,2 |

| **NO** | **DAERAH** | **LUAS LAHAN** | **PRODUKSI** | **C1** | **C2** | **C3** | **JARAK TERPENDEK** |
|---|---|---|---|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 4688 | 25 | 550 | 25 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 10698 | 932 | 141 | 141 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 7265 | 124 | 9 | 9 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 7119 | 124 | 46 | 46 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 6937 | 87 | 48 | 48 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 4894 | 86 | 572 | 86 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 3851 | 63 | 813 | 63 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 720 | 10054 | 14188 | 720 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 480 | 2670 | 4962 | 480 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 5355 | 37 | 300 | 37 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 7474 | 187 | 35 | 35 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 5647 | 36 | 225 | 36 |

| NO | DAERAH | LUAS LAHAN | PRODUKSI | KELOMPOK |
|---|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 2 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 3 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 3 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 3 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 3 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 2 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 2 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 1 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 1 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 2 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 3 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 2 |

DENGAN KESIMPULAN DATA 3
cluster 1 : sebanyak 2 data
cluster 2 : sebanyak 5 data
cluster 3 : sebanyak 5 data
karena data 2  dan data 3 sama maka perhitungan selesai

*Redho Aidil Iqrom | 202420043*

# Tugas 5
# Advanced Database

Pelari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format ms word dan dilengkapi dengan gambar dan keterangan yang jelas untuk tiap langkah-langkahnya. Sertakan sumber sumber rujukan yang anda gunakan. Kumpulkan sebelum waktu batas akhir pengumpulan tugas.

**Jawaban :**

Pada tugas ini saya coba mengambil sample untuk mengukur performance Clustering dengan menggunakan K-Means, pengukuran performance clustering saat ini menggunakan clusterisasi sampai di dapatkan nilai stabil dan dengan rapid miner saya coba untuk bandingkan hasilnya dengan sumber data yang sama, adapun datanya sebagai berikut:

Data Komsumsi Listrik bulan Mei 2020 di Prov Lampung :

| | Kabupaten/Kota | Latitude | Longitude | Mei 2020 | | |
|---|---|---|---|---|---|---|
| | | | | Rumah Tangga | Industri | Bisnis |
| | Kabupaten Tulang Bawang Barat | -4.52569 | 105.07912 | 8647851 | 1424419 | 771600 |
| | Kabupaten Way Kanan | -4.49636 | 104.56552 | 8922355 | 904997 | 580164 |
| | Kabupaten Lampung Barat | -5.10952 | 104.1466 | 7870512 | 908299 | 1508881 |
| | Kabupaten Lampung Selatan | -5.56226 | 105.54743 | 39811088 | 28596132 | 6923204 |
| | Kabupaten Lampung Tengah | -4.8008 | 105.31311 | 43819603 | 17071293 | 5099183 |
| | Kabupaten Lampung Timur | -5.11349 | 105.68817 | 17150724 | 4651063 | 3006046 |
| | Kabupaten Lampung Utara | -4.81339 | 104.75209 | 21682207 | 3629202 | 2188612 |
| | Kabupaten Pringsewu | -5.33311 | 104.98561 | 17504659 | 701278 | 1618426 |
| | Kabupaten Tanggamus | -5.30274 | 104.56552 | 14831519 | 2175713 | 1911240 |
| | Kabupaten Tulang Bawang | -4.31765 | 105.50054 | 13723285 | 1574286 | 1730416 |
| | Kota Bandar Lampung | -5.39713 | 105.26678 | 60305690 | 7951584 | 18855425 |
| | Kota Metro | -5.11783 | 105.30726 | 17774258 | 753564 | 2565855 |

Dari data diatas saya coba untuk mengukur bagaimana performance dari clustering yang menggunakan K-Means dengan membandingkan 2 cara yaitu dengan Coding PhP dan Rapidminer.

Berbasiskan Web Stage Awal Data

Iterasi Ke-1

Proses Iterasi Selanjutnya

| | Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rumah Tangga | Industri | Bisnis | 4035984 | 396909 | 14840900 | 396909 | 4035984 | 14840900 | 14840900 | 396909 | 4035984 | | | |
| | Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 14841505 | | | 16517975 | | | 7075721 | | | 0 | 0 | 1 |
| | Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 15083214 | | | 16907244 | | | 6872410 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 13881925 | | | 15600637 | | | 7431963 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 46235745 | | | 47110171 | | | 37776216 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 44222985 | | | 46371873 | | | 33450421 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 18170250 | | | 20521510 | | | 4949129 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 21952614 | | | 24765091 | | | 7788708 | | | 0 | 0 | 1 |
| | Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 18876749 | | | 21877595 | | | 3610102 | | | 0 | 0 | 1 |
| | Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 16937645 | | | 19467785 | | | 2771059 | | | 0 | 0 | 1 |
| | Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 16343647 | | | 18855689 | | | 2819738 | | | 0 | 0 | 1 |
| | Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 56916336 | | | 60170677 | | | 48412148 | | | 0 | 0 | 1 |
| | Kota Metro | 17774258 | 753564 | 2565855 | 18426722 | | | 21527268 | | | 3300465 | | | 0 | 0 | 1 |

**Iterasi Ke-1**

| Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rumah Tangga | Industri | Bisnis | 0 | 0 | 0 | 0 | 0 | 0 | 22670312 | 5861819 | 3896587 | | | |
| Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 8798276 | | | 8798276 | | | 15036139 | | | 1 | 1 | 0 |
| Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 8986880 | | | 8986880 | | | 14985829 | | | 1 | 1 | 0 |
| Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 8065152 | | | 8065152 | | | 15788369 | | | 1 | 1 | 0 |
| Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 49503456 | | | 49503456 | | | 28632421 | | | 0 | 0 | 1 |
| Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 47303153 | | | 47303153 | | | 23966456 | | | 0 | 0 | 1 |
| Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 18022653 | | | 18022653 | | | 5720564 | | | 0 | 0 | 1 |
| Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 22092515 | | | 22092515 | | | 2979616 | | | 0 | 0 | 1 |
| Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 17593299 | | | 17593299 | | | 7648868 | | | 0 | 0 | 1 |
| Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 15111602 | | | 15111602 | | | 8886825 | | | 0 | 0 | 1 |
| Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 13921252 | | | 13921252 | | | 10155025 | | | 0 | 0 | 1 |
| Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 63683051 | | | 63683051 | | | 40553120 | | | 0 | 0 | 1 |
| Kota Metro | 17774258 | 753564 | 2565855 | 17974307 | | | 17974307 | | | 7199754 | | | 0 | 0 | 1 |

Iterasi Ke-2

**Iterasi Ke-2**

| Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rumah Tangga | Industri | Bisnis | 8480239 | 1079238 | 953548 | 8480239 | 1079238 | 953548 | 27400337 | 7456012 | 4877600 | | | |
| Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 424674 | | | 424674 | | | 20122005 | | | 1 | 1 | 0 |
| Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 604352 | | | 604352 | | | 20070365 | | | 1 | 1 | 0 |
| Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 842248 | | | 842248 | | | 20871868 | | | 1 | 1 | 0 |
| Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 42124082 | | | 42124082 | | | 24599103 | | | 0 | 0 | 1 |
| Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 39010289 | | | 39010289 | | | 19028794 | | | 0 | 0 | 1 |
| Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 9599374 | | | 9599374 | | | 10790042 | | | 1 | 1 | 0 |
| Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 13502579 | | | 13502579 | | | 7387296 | | | 0 | 0 | 1 |
| Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 9056769 | | | 9056769 | | | 12416646 | | | 1 | 1 | 0 |
| Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 6515994 | | | 6515994 | | | 13951919 | | | 1 | 1 | 0 |
| Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 5323356 | | | 5323356 | | | 15217136 | | | 1 | 1 | 0 |
| Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 55259240 | | | 55259240 | | | 35754544 | | | 0 | 0 | 1 |
| Kota Metro | 17774258 | 753564 | 2565855 | 9438452 | | | 9438452 | | | 11955265 | | | 1 | 1 | 0 |

Iterasi Ke-3

**Iterasi Ke-3**

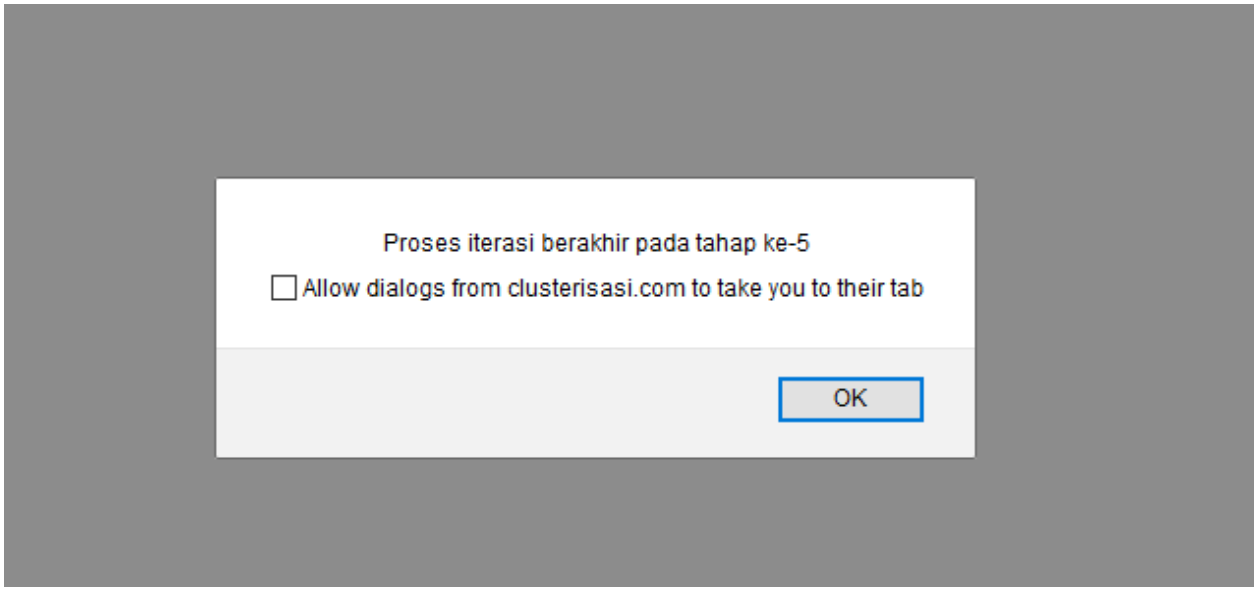
| Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rumah Tangga | Industri | Bisnis | 13303145 | 1636702 | 1711578 | 13303145 | 1636702 | 1711578 | 41404647 | 14312052 | 8266606 | | | |
| Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 4753986 | | | 4753986 | | | 35989913 | | | 1 | 1 | 0 |
| Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 4583319 | | | 4583319 | | | 35971236 | | | 1 | 1 | 0 |
| Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 5484994 | | | 5484994 | | | 36740518 | | | 1 | 1 | 0 |
| Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 38165991 | | | 38165991 | | | 14435341 | | | 0 | 0 | 1 |
| Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 34365049 | | | 34365049 | | | 4845408 | | | 0 | 0 | 1 |
| Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 5056271 | | | 5056271 | | | 26631954 | | | 1 | 1 | 0 |
| Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 8625908 | | | 8625908 | | | 23238759 | | | 1 | 1 | 0 |
| Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 4305393 | | | 4305393 | | | 28295952 | | | 1 | 1 | 0 |
| Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 1632887 | | | 1632887 | | | 29896697 | | | 1 | 1 | 0 |
| Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 425168 | | | 425168 | | | 31164567 | | | 1 | 1 | 0 |
| Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 50428448 | | | 50428448 | | | 22579372 | | | 0 | 0 | 1 |
| Kota Metro | 17774258 | 753564 | 2565855 | 4636870 | | | 4636870 | | | 27833908 | | | 1 | 1 | 0 |

Iterasi Ke-4

**Iterasi Ke-4**

| | Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rumah Tangga | Industri | Bisnis | 14234152 | 1858091 | 1764582 | 14234152 | 1858091 | 1764582 | 47978793 | 17873003 | 10292604 | | | |
| | Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 5690417 | | | 5690417 | | | 43682130 | | | 1 | 1 | 0 |
| | Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 5525072 | | | 5525072 | | | 43676654 | | | 1 | 1 | 0 |
| | Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 6439208 | | | 6439208 | | | 44425547 | | | 1 | 1 | 0 |
| | Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 37359253 | | | 37359253 | | | 13894235 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 33434412 | | | 33434412 | | | 6701733 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 4224726 | | | 4224726 | | | 34326134 | | | 1 | 1 | 0 |
| | Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 7667473 | | | 7667473 | | | 30985012 | | | 1 | 1 | 0 |
| | Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 3472145 | | | 3472145 | | | 36038623 | | | 1 | 1 | 0 |
| | Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 692271 | | | 692271 | | | 37621722 | | | 1 | 1 | 0 |
| | Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 585404 | | | 585404 | | | 38889575 | | | 1 | 1 | 0 |
| | Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 49515797 | | | 49515797 | | | 17991910 | | | 0 | 0 | 1 |
| | Kota Metro | 17774258 | 753564 | 2565855 | 3793990 | | | 3793990 | | | 35568129 | | | 1 | 1 | 0 |

Iterasi Ke-5




Pada saat iterasi ke 5 data sudah di dapatkan sampai dengan semua nilainya stabil atau sama, sehingga didapat nilai cluster yang sesuai sebagai berikut :

Iterasi ke-1

| C1 | C2 | C3 |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |

*Redho Aidil Iqrom | 202420043*

Iterasi ke-2

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |

Iterasi ke-3

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |

Iterasi ke-4

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |

Iterasi ke-5

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |

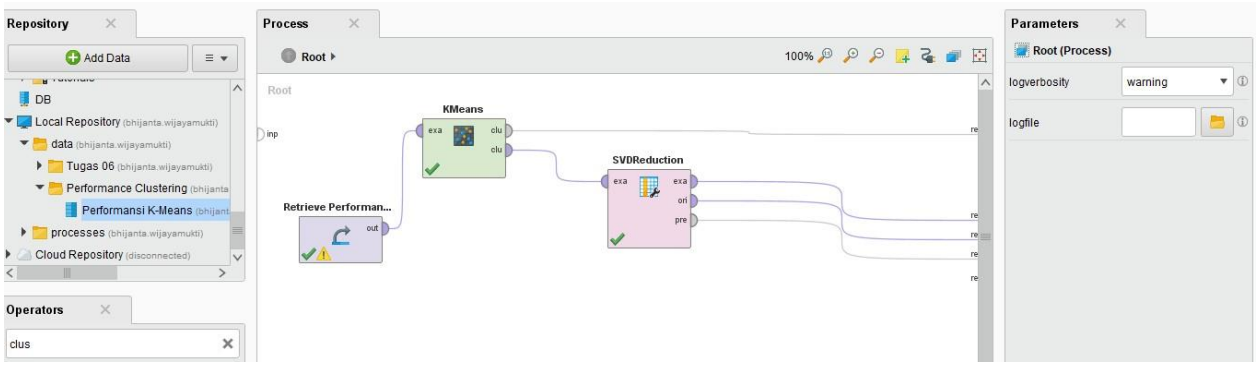
Dari Hasil Iterasi ke 5 didaptkan kesimpulan sebagai berikut :

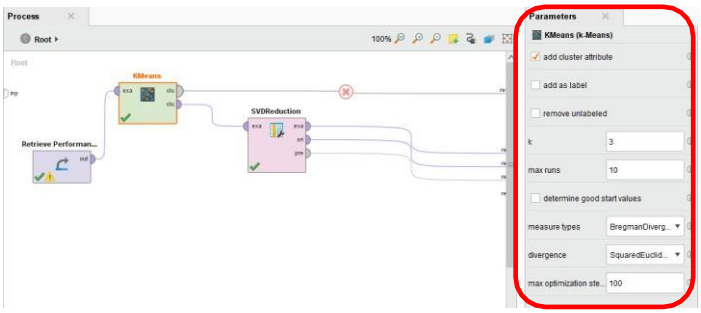
| | Kabupaten/Kota | Bulan | Zona |
|---|---|---|---|
| | Kabupaten Tulang Bawang Barat | Mei | Kuning |
| | Kabupaten Way Kanan | Mei | Kuning |
| | Kabupaten Lampung Barat | Mei | Kuning |
| | Kabupaten Lampung Selatan | Mei | Hijau |
| | Kabupaten Lampung Tengah | Mei | Hijau |
| | Kabupaten Lampung Timur | Mei | Kuning |
| | Kabupaten Lampung Utara | Mei | Kuning |
| | Kabupaten Pringsewu | Mei | Kuning |
| | Kabupaten Tanggamus | Mei | Kuning |
| | Kabupaten Tulang Bawang | Mei | Kuning |
| | Kota Bandar Lampung | Mei | Hijau |
| | Kota Metro | Mei | Kuning |

- **Berbasiskan Rapid miner**

Dengan data yang sama seperti diatas saya coba mengukur bagaimana performance dari Clustering K-Means menggunakan tool rapidminer, adapun hasilnya sebagai berikut :

Desain Mapping Rapidminer

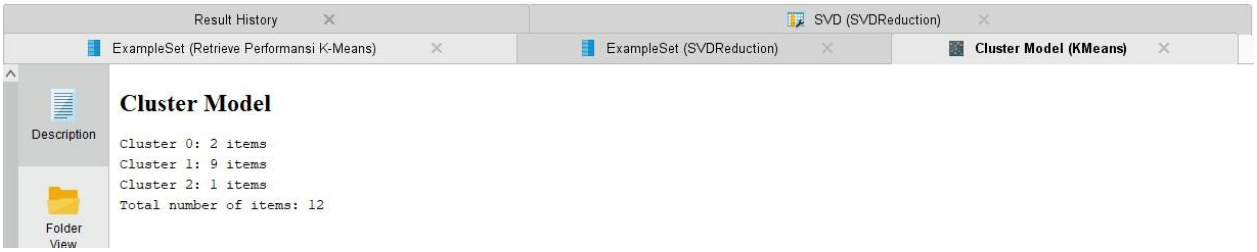


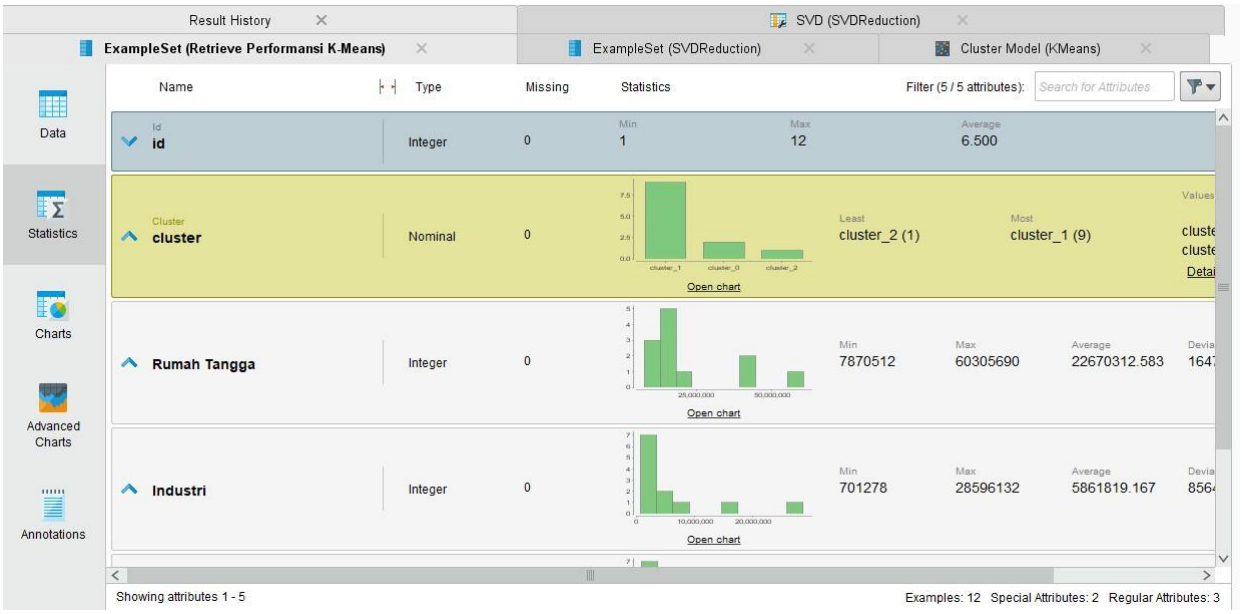



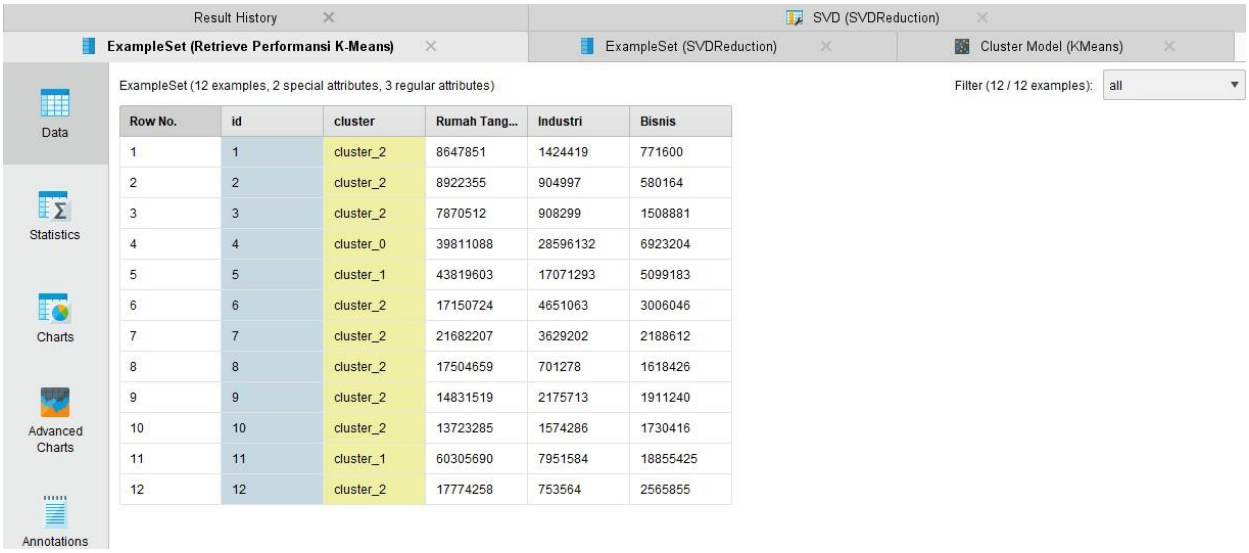

Parameter K-Means

*Redho Aidil Iqrom | 202420043*

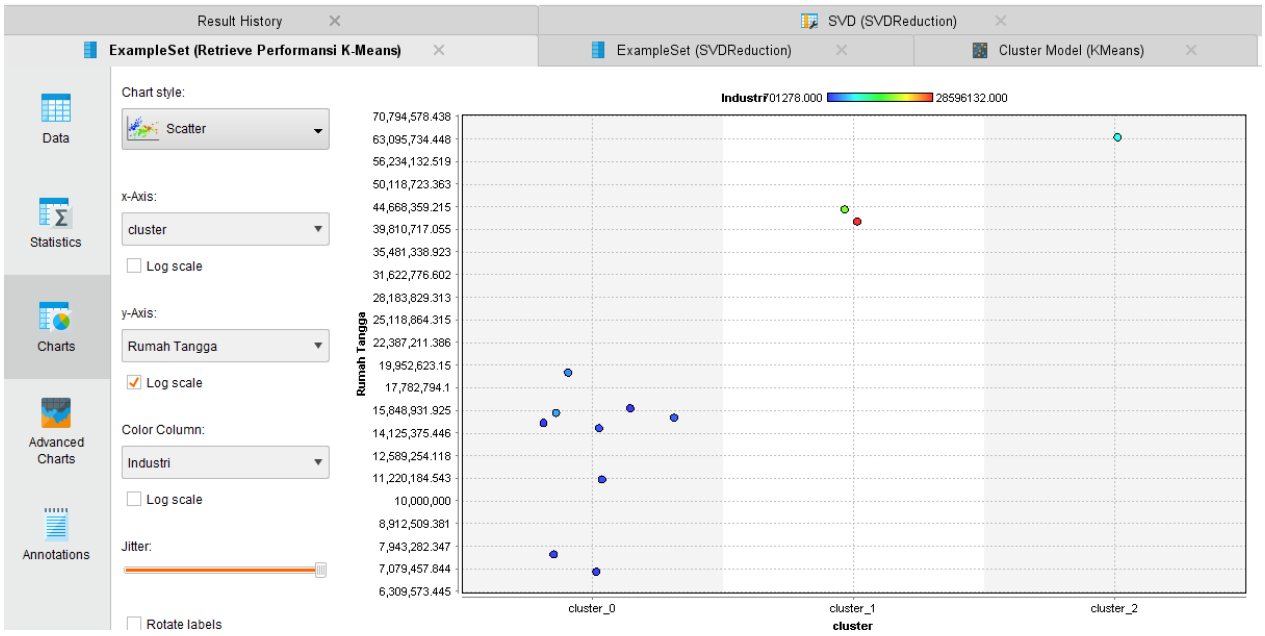
Hasil Clustering




| Row No. | id | cluster | Rumah Tang... | Industri | Bisnis |
|---|---|---|---|---|---|
| 1 | 1 | cluster_2 | 8647851 | 1424419 | 771600 |
| 2 | 2 | cluster_2 | 8922355 | 904997 | 580164 |
| 3 | 3 | cluster_2 | 7870512 | 908299 | 1508881 |
| 4 | 4 | cluster_0 | 39811088 | 28596132 | 6923204 |
| 5 | 5 | cluster_1 | 43819603 | 17071293 | 5099183 |
| 6 | 6 | cluster_2 | 17150724 | 4651063 | 3006046 |
| 7 | 7 | cluster_2 | 21682207 | 3629202 | 2188612 |
| 8 | 8 | cluster_2 | 17504659 | 701278 | 1618426 |
| 9 | 9 | cluster_2 | 14831519 | 2175713 | 1911240 |
| 10 | 10 | cluster_2 | 13723285 | 1574286 | 1730416 |
| 11 | 11 | cluster_1 | 60305690 | 7951584 | 18855425 |
| 12 | 12 | cluster_2 | 17774258 | 753564 | 2565855 |




**Kesimpulan**

K-means merupakan salah satu algoritma clustering. Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan supervised learning yang menerima masukan berupa vektor (-x-1,y1), (-x-2,y2),…,(-x-i,yi), dimana x-i merupakan data dari suatu data pelatihan dan y-i merupakan label kelas untuk x-i

Kelebihan pada algoritma k-means, yaitu :

1. Mudah untuk di implementasikan dan dijalankan.
2. Waktu yang dibutuhkan untuk menjalankan pembelajaran ini relatif cepat.
3. Mudah untuk di adaptasi.
4. Umum digunakan.
5. Secara performance dari Clustering dengan K-Means cukup dengan baik dapat menyelesaikan clustering baik secara metode Web dan Rapidminer.

Sumber :https://clusterisasi.com/k-means/admin/iterasi_kmeans_hasil

**Nama : Robby Prabowo**
**NIM : 202420001**
**TUGAS 5**

Pelajari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format ms word dan dilengkapi dengan gambar dan keterangan yang jelas.

Jawaban :
Pada tugas ini saya mengambil contoh Penerapan Metode K-Means Clustering untuk Menentukan Persediaan Stok Barang Pada Toko.

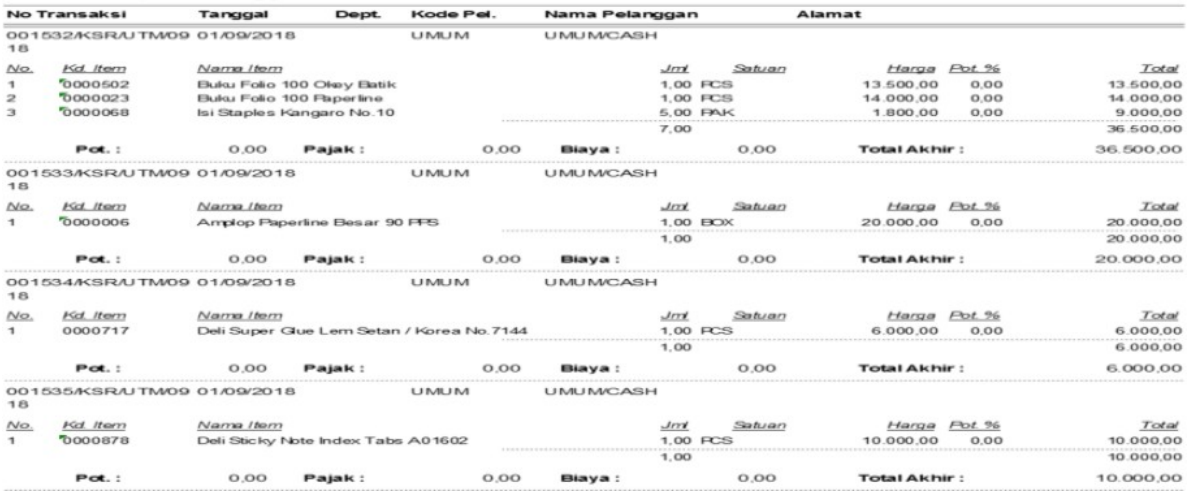
1. Data Penjualan
   Dibawah ini adalah tampilan data penjualan pada Toko;

| Kode Item | Nama Item | Jenis | Merek | Jumlah Terjual | Satuan |
|---|---|---|---|---|---|
| 0000006 | Amplop Paperline Besar 90 PPS | Amplop | Paperline | 52,00 | PAK |
| 0000007 | Amplop Paperline Kecil 104 PPS | Amplop | Paperline | 82,00 | PAK |
| 0000016 | Binder Clip Kenko No.107 | Misc ATK | Nomerk | 10,00 | BOX |
| 0000017 | Binder Clip Joyko No.111 | Misc ATK | Joyko | 8,00 | BOX |
| 0000021 | Buku Tulis Big Boss 42 | Buku | Bigboss | 70,00 | PCS |
| 0000023 | Buku Folio 100 Paperline | Buku | Paperline | 9,00 | PCS |
| ..... | | | | | |
| 0000822 | Map Dokumen Clear Holder 40 Pocket | Map / File | Nomerk | 5,00 | PCS |
| 0000825 | Buku Tulis Sidu Kotak 38 Lembar | Buku | Sidu | 50,00 | PCS |
| 0000826 | Buku Tulis Sidu 100 Lembar | Buku | Sidu | 15,00 | PCS |

2. Data Persediaan
   Dibawah ini tampilan data persediaan barang pada Toko;

| Kode Item | Nama Item | Jenis | Merek | Jumlah Stok | Satuan |
|---|---|---|---|---|---|
| 0000006 | Amplop Paperline Besar 90 PPS | Amplop | Paperline | 72,00 | PAK |
| 0000007 | Amplop Paperline Kecil 104 PPS | Amplop | Paperline | 120,00 | PAK |
| 0000013 | Bantal Stempel Besar Hero E1460 Violet | Misc ATK | Hero | 12,00 | PCS |
| 0000016 | Binder Clip Kenko No.107 | Misc ATK | Nomerk | 24,00 | BOX |
| 0000017 | Binder Clip Joyko No.111 | Misc ATK | Joyko | 12,00 | BOX |
| 0000029 | Buku Kwitansi Sedang Sidu / Paperline 40 lembar | Buku Kwintansi | Sidu | 24,00 | PCS |
| 0000037 | Buku Nota Besar 3 Ply Paperline 25 set | Buku Nota | Paperline | 60,00 | PCS |
| 0000038 | Buku Nota Kecil 1 Ply Paperline 50 lbr | Buku Nota | Paperline | 84,00 | PCS |
| ..... | | | | | |
| 0000760 | Deli Pena Ballpoint No. 6546 | Pena Pulpen | Deli | 24,00 | PCS |
| 0000764 | Bantex Ordner 1465V Blue | File | Bantex | 12,00 | PCS |
| 0000772 | Bantex Jumbo Box File Blue 4011-62 | File | Bantex | 12,00 | PCS |
| 0000773 | Pembolong Kertas Cadwell CD-30S | Misc ATK | Cadwell | 12,00 | PCS |
| 0000775 | Jangka Set Nikiko TS-688 | Alat Gambar | Nomerk | 12,00 | PCS |
| 0000776 | Pensil Warna JOYKO Kaleng CP-12RT | Alat Gambar | Nomerk | 12,00 | PCS |
| 0000777 | Lakban Bening 1" Daimaru | Lakban | Daimaru | 48,00 | PCS |

3. Data Transaksi
   Di bawah ini tampilan data transaksi pada Toko;



| No Transaksi | Tanggal | Dept. | Kode Pel. | Nama Pelanggan | Alamat |
|---|---|---|---|---|---|
| 001532/KSR/UTM/0918 | 01/09/2018 | | UMUM | UMUM/CASH | |

| No. | Kd. Item | Nama Item | Jml | Satuan | Harga | Pot. % | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0000502 | Buku Folio 100 Okey Batik | 1,00 | PCS | 13.500,00 | 0,00 | 13.500,00 |
| 2 | 0000023 | Buku Folio 100 Paperline | 1,00 | PCS | 14.000,00 | 0,00 | 14.000,00 |
| 3 | 0000068 | Isi Staples Kangaro No.10 | 5,00 | PAK | 1.800,00 | 0,00 | 9.000,00 |
| | | | 7,00 | | | | 36.500,00 |

Pot. : 0,00 Pajak : 0,00 Biaya : 0,00 Total Akhir : 36.500,00

| No Transaksi | Tanggal | Dept. | Kode Pel. | Nama Pelanggan | Alamat |
|---|---|---|---|---|---|
| 001533/KSR/UTM/0918 | 01/09/2018 | | UMUM | UMUM/CASH | |

| No. | Kd. Item | Nama Item | Jml | Satuan | Harga | Pot. % | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0000006 | Amplop Paperline Besar 90 PPS | 1,00 | BOX | 20.000,00 | 0,00 | 20.000,00 |
| | | | 1,00 | | | | 20.000,00 |

Pot. : 0,00 Pajak : 0,00 Biaya : 0,00 Total Akhir : 20.000,00

| No Transaksi | Tanggal | Dept. | Kode Pel. | Nama Pelanggan | Alamat |
|---|---|---|---|---|---|
| 001534/KSR/UTM/0918 | 01/09/2018 | | UMUM | UMUM/CASH | |

| No. | Kd. Item | Nama Item | Jml | Satuan | Harga | Pot. % | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0000717 | Deli Super Glue Lem Setan / Korea No.7144 | 1,00 | PCS | 6.000,00 | 0,00 | 6.000,00 |
| | | | 1,00 | | | | 6.000,00 |

Pot. : 0,00 Pajak : 0,00 Biaya : 0,00 Total Akhir : 6.000,00

| No Transaksi | Tanggal | Dept. | Kode Pel. | Nama Pelanggan | Alamat |
|---|---|---|---|---|---|
| 001535/KSR/UTM/0918 | 01/09/2018 | | UMUM | UMUM/CASH | |

| No. | Kd. Item | Nama Item | Jml | Satuan | Harga | Pot. % | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0000878 | Deli Sticky Note Index Tabs A01602 | 1,00 | PCS | 10.000,00 | 0,00 | 10.000,00 |
| | | | 1,00 | | | | 10.000,00 |

Pot. : 0,00 Pajak : 0,00 Biaya : 0,00 Total Akhir : 10.000,00

4. Data Cleaning dan Penggabungan Data

Hasil gabungan dari ketiga data tersebut dapat dilihat pada tampilan data gabungan;

| Kode Item | Nama Item | Jenis | Merek | Jumlah Terjua | Jumlah Stok | Volume Penjual | Satuan |
|---|---|---|---|---|---|---|---|
| 0000006 | Amplop Paperline Besar 90 PPS | Amplop | Paperline | 52,00 | 72,00 | 6,00 | PAK |
| 0000007 | Amplop Paperline Kecil 104 PPS | Amplop | Paperline | 82,00 | 120,00 | 18,00 | PAK |
| 0000023 | Buku Folio 100 Paperline | Buku | Paperline | 9,00 | 24,00 | 7,00 | PCS |
| 0000027 | Buku Kwitansi Besar Sidu 50 lembar | Buku Kwintansi | Sidu | 7,00 | 24,00 | 5,00 | PCS |
| 0000028 | Buku Kwitansi Kecil Sidu/PPL 40 lembar | Buku Kwintansi | Sidu | 13,00 | 36,00 | 8,00 | PCS |
| 0000029 | Buku Kwitansi Sedang Sidu / Paperline 40 lembar | Buku Kwintansi | Sidu | 4,00 | 24,00 | 3,00 | PCS |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 0000440 | Buku Milimeter Folio Cox | Buku | Cox | 3,00 | 12,00 | 1,00 | PCS |
| 0000441 | Peruncing Daun Fancy 033 041 | Alat Tulis | Nomerk | 5,00 | 24,00 | 5,00 | PCS |
| 0000442 | Pena Joyko Lino BP-249 | Pena Pulpen | Joyko | 47,00 | 60,00 | 13,00 | PCS |
| 0000443 | Tas Ransel Smiley Emoticon AD-200 | Tas | Nomerk | 5,00 | 6,00 | 5,00 | PCS |
| 0000444 | Paper Spear Dingli DL 85201 | Misc ATK | Nomerk | 4,00 | 12,00 | 3,00 | PCS |
| 0000446 | Jangka Besi Space Ball SP4001 | Alat Gambar | Nomerk | 4,00 | 6,00 | 4,00 | PCS |
| 0000447 | Stapler HD-45 V-tec | Stapler / Staples | V-tec | 2,00 | 12,00 | 2,00 | PCS |
| 0000448 | Paper Fastener Eco-nomical | Misc ATK | Nomerk | 5,00 | 12,00 | 2,00 | BOX |
| 0000450 | Pensil Warna 12W Panjang Disney | Alat Gambar | Nomerk | 1,00 | 12,00 | 1,00 | SET |

Pada penelitian ini data yang diambil sebagai bahan penilitian yaitu ± 1000 data item, kemudian dari data tersebut penulis kembali melakukan proses data cleaning berikutnya dengan memilih 62 data item saja secara random untuk dijadikan data sampel yang akan diolah selanjutnya, dan beberapa atribut yang berpengaruh, atribut tersebut meliputi kode item, jenis item, jumlah terjual, jumlah stok, volume penjualan, rata-rata.

Berikut hasil Data Sample;

| No | Kode Item | Jenis | Jumlah Terjual | Jumlah Stok | Volume Penjualan | Rata-Rata |
|---|---|---|---|---|---|---|
| 1 | 0000369 | Alat Gambar | 17,00 | 60,00 | 10,00 | 14,50 |
| 2 | 0001174 | Alat Gambar | 12,00 | 24,00 | 12,00 | 8,00 |
| 3 | 0000466 | Alat Tulis | 138,00 | 240,00 | 62,00 | 73,33 |
| 4 | 0000882 | Alat Tulis | 33,00 | 60,00 | 19,00 | 18,67 |
| 5 | 0000007 | Amplop | 82,00 | 120,00 | 18,00 | 36,67 |
| 6 | 0000006 | Amplop | 52,00 | 72,00 | 6,00 | 21,67 |
| | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 7 | 0000635 | Botol Minum | 6,00 | 10,00 | 4,00 | 3,33 |
| 8 | 0001404 | Botol Minum | 3,00 | 5,00 | 2,00 | 1,67 |
| 9 | 0000263 | Buku | 29,00 | 48,00 | 9,00 | 14,33 |
| 10 | 0000021 | Buku | 70,00 | 96,00 | 12,00 | 29,67 |
| 11 | 0000613 | Buku Binder | 4,00 | 5,00 | 4,00 | 2,17 |
| 12 | 0000606 | Buku Binder | 1,00 | 5,00 | 1,00 | 1,17 |
| 13 | 0001393 | Buku Diary | 6,00 | 12,00 | 6,00 | 4,00 |
| 14 | 0001395 | Buku Diary | 10,00 | 12,00 | 8,00 | 5,00 |
| 15 | 0000028 | Buku Kwintansi | 13,00 | 36,00 | 8,00 | 9,50 |
| 16 | 0000027 | Buku Kwintansi | 7,00 | 24,00 | 5,00 | 6,00 |
| 17 | 0000039 | Buku Nota | 23,00 | 60,00 | 3,00 | 14,33 |
| 18 | 0000794 | Buku Nota | 20,00 | 60,00 | 5,00 | 14,17 |
| 19 | 0001350 | Buku Sampul | 69,00 | 100,00 | 7,00 | 29,33 |
| 20 | 0000783 | Buku Sampul | 11,00 | 20,00 | 6,00 | 6,17 |
| 21 | 0000503 | File | 34,00 | 48,00 | 6,00 | 14,67 |
| 22 | 0001011 | File | 2,00 | 12,00 | 2,00 | 2,67 |
| 23 | 0000347 | Gunting | 12,00 | 24,00 | 8,00 | 7,33 |
| 24 | 00008 | Gunting | 12,0 | 24,00 | 12,00 | 8,00 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | 32 | | 0 | | | |
| 25 | 0000674 | Kalkulator | 1,00 | 4,00 | 1,00 | 1,00 |
| 26 | 0000972 | Kalkulator | 3,00 | 4,00 | 3,00 | 1,67 |
| 27 | 0000095 | Kertas | 105,00 | 200,00 | 14,00 | 53,17 |
| 28 | 0001169 | Kertas | 52,00 | 60,00 | 45,00 | 26,17 |
| 29 | 0000360 | Kertas HVS | 1,00 | 5,00 | 1,00 | 1,17 |
| 30 | 0000358 | Kertas HVS | 32,00 | 40,00 | 15,00 | 14,50 |
| 31 | 0001566 | Kotak Makan | 2,00 | 5,00 | 2,00 | 1,50 |
| 32 | 0000651 | Kotak Makan | 2,00 | 6,00 | 2,00 | 1,67 |
| 33 | 0001511 | Kotak Pensil | 5,00 | 12,00 | 5,00 | 3,67 |
| 34 | 0001512 | Kotak Pensil | 3,00 | 12,00 | 2,00 | 2,83 |
| 35 | 0000296 | Lakban | 36,00 | 60,00 | 14,00 | 18,33 |
| 36 | 0001311 | Lakban | 46,00 | 60,00 | 14,00 | 20,00 |
| 37 | 0000352 | Lem | 24,00 | 36,00 | 14,00 | 12,33 |
| 38 | 0000717 | Lem | 19,00 | 24,00 | 16,00 | 9,83 |
| 39 | 0000165 | Map / File | 52,00 | 200,00 | 3,00 | 42,50 |
| 40 | 0000785 | Map / File | 7,00 | 100,00 | 2,00 | 18,17 |
| 41 | 0000697 | Misc | 1,00 | 6,00 | 1,00 | 1,33 |
| 42 | 0000892 | Misc | 2,00 | 24,00 | 1,00 | 4,50 |
| 43 | 0001578 | Misc Ultah | 8,00 | 36,00 | 4,00 | 8,00 |
| 44 | 0000711 | Misc Ultah | 8,00 | 24,00 | 5,00 | 6,17 |
| 45 | 0000083 | Misc ATK | 9,00 | 24,00 | 9,00 | 7,00 |
| 46 | 0000364 | Misc ATK | 9,00 | 12,00 | 8,00 | 4,83 |
| 47 | 0001419 | Pena Fancy | 60,00 | 60,00 | 46,00 | 27,67 |
| 48 | 0001529 | Pena Fancy | 31,00 | 36,00 | 26,00 | 15,50 |
| 49 | 0000242 | Pena Pulpen | 203,00 | 240,00 | 27,00 | 78,33 |
| 50 | 0000505 | Pena Pulpen | 95,00 | 120,00 | 50,00 | 44,17 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 51 | 0000782 | Pensil | 44,00 | 60,00 | 28,00 | 22,00 |
| 52 | 0000619 | Pensil | 43,00 | 60,00 | 21,00 | 20,67 |
| 53 | 0001486 | Post It / Stick Note / Memo | 4,00 | 12,00 | 4,00 | 3,33 |
| 54 | 0001489 | Post It / Stick Note / Memo | 3,00 | 12,00 | 3,00 | 3,00 |
| 55 | 0000715 | Stapler / Staples | 37,00 | 48,00 | 8,00 | 15,50 |
| 56 | 0000068 | Stapler / Staples | 91,00 | 120,00 | 27,00 | 39,67 |
| 57 | 0001142 | Tas | 13,00 | 16,00 | 12,00 | 6,83 |
| 58 | 0001134 | Tas | 13,00 | 16,00 | 13,00 | 7,00 |
| 59 | 0001227 | Tip Ex | 28,00 | 36,00 | 23,00 | 14,50 |
| 60 | 0000330 | Tip Ex | 26,00 | 36,00 | 21,00 | 13,83 |
| 61 | 0001129 | Tissue | 19,00 | 24,00 | 17,00 | 10,00 |
| 62 | 0001131 | Tissue | 4,00 | 12,00 | 4,00 | 3,33 |

Dari data yang telah ada pada tabel sampel diatas, selanjutnya akan dilakukan perhitungan manual k- means clustering terlebih dahulu. Adapun beberapa langkah yang dapat dilakukan dalam perhitungan k-means clustering ini yaitu antara lain

1. Tahap Pertama

Menentukan jumlah cluster (K) terhadap data sampel yang sudah ada, maka untuk penentuan jumlah cluster (K), penulis menentukan dengan 5 cluster (K) seperti yang sudah dijelaskan sebelumnya, untuk proses perhitungan manual algoritma k-means yang akan dilakukan.

2. Tahap 2

Inisialisasi nilai centroid yaitu dengan menentukan nilai centroid awal pada data sampel yang sudah ada secara acak yaitu dipilih data dengan nomor urut ke-10 dengan kode item 0000021 dan jenis item buku, nomor urut ke-42 dengan 0000892 dan jenis item misc, nomor urut ke-34 dengan kode item 0001512 dan jenis item kotak pensil, nomor urut ke-19 dengan kode item 0001350 dan jenis item buku sampul, dan nomor urut ke-1 dengan kode item 0000369 dan jenis item alat gambar. Berikut dapat dilihat pada tabel dibawah ini nilai pada centroid awal yang ditentukan

| Cluster | Centroid Awal | | | |
|---|---|---|---|---|
| 1 | 70,00 | 96,00 | 12,00 | 29,67 |
| 2 | 2,00 | 24,00 | 1,00 | 4,50 |
| 3 | 3,00 | 12,00 | 2,00 | 2,83 |
| 4 | 69,00 | 100,00 | 7,00 | 29,33 |
| 5 | 17,00 | 60,00 | 10,00 | 14,50 |

3. Tahap Ketiga

Menghitung jarak antara centroid dengan data, yaitu menghitung jarak dari setiap data ke centroid terdekat. Centroid terdekat akan menjadi cluster yang diikuti oleh data tersebut. Untuk perhitungan jarak ke setiap centroid digunakan rumus Euclidean Distance seperti dibawah ini :

$d(xi, \mu j) = \sqrt{\sum}(xi - \mu j)2$ (4

) Dimana : d = jarak

$xi$ = data kriteria

$\mu j$ = centroid pada cluster ke-j

Jarak centroid data ke-1 pada cluster 1 adalah :

$d(xi, \mu j) = \sqrt{\sum}(xi - \mu j)2$ = √((17,00-70,00)^2+(60,00-96,00)^2+(10,00-12,00)^2+(14,50-29,67)^2)

= 65,87

Jarak centroid data ke-1 pada cluster 2 adalah :

$d(xi, \mu j) = \sqrt{\sum}(xi - \mu j)2$ = √((17,00-2,00)^2+(60,00-24,00)^2+(10,00-1,00)^2+(14,50-4,50)^2 )

= 41,26

Jarak centroid data ke-1 pada cluster 3 adalah :

$d(xi, \mu j) = \sqrt{\sum}(xi - \mu j)2$ = √((17,00-3,00)^2+(60,00-12,00)^2+(10,00-2,00)^2+(14,50-2,83)^2 )

= 51,96

Jarak centroid data ke-1 pada cluster 4 adalah :

$d(xi, \mu j) = \sqrt{\sum}(xi - \mu j)2$ = √((17,00-69,00)^2+(60,00-100,00)^2+(10,00-7,00)^2+(14,50-29,33)^2 )

= 67,33

Jarak centroid data ke-1 pada cluster 5 adalah :

$d(xi, \mu j) = \sqrt{\sum}(xi - \mu j)2$ = √((17,00-17,00)^2+(60,00-60,00)^2+(10,00-10,00)^2+(14,50-14,50)^2 )

= 0

Untuk jarak centroid selanjutnya dapat dilihat pada tabel dibawah ini, yang mana penulis memasukkan tabel dari hasil proses perhitungan keseluruhan data sampel untuk jarak centroid iterasi yang pertama.

| No | Kode Item | Jenis | Jarak Centroid | | | | | Jarak Terdekat | Iterasi yang diikuti |
|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | | |
| 1 | 0000369 | Alat Gambar | 65,87 | 41,26 | 51,96 | 67,33 | 0,00 | 0,00 | 5 |
| 2 | 0001174 | Alat Gambar | 94,96 | 15,27 | 18,75 | 97,49 | 36,98 | 15,27 | 2 |
| 3 | 0000466 | Alat Tulis | 172,53 | 271,31 | 280,68 | 171,24 | 230,67 | 171,24 | 4 |
| 4 | 0000882 | Alat Tulis | 53,24 | 52,74 | 61,19 | 56,16 | 18,82 | 18,82 | 5 |
| 5 | 0000007 | Amplop | 28,37 | 130,15 | 138,94 | 27,27 | 91,54 | 27,27 | 4 |
| 6 | 0000006 | Amplop | 31,62 | 71,58 | 79,82 | 33,66 | 37,90 | 31,62 | 1 |
| 7 | 0000635 | Botol Minum | 110,68 | 14,91 | 4,15 | 112,93 | 52,74 | 4,15 | 3 |
| 8 | 0001404 | Botol Minum | 116,85 | 19,26 | 7,10 | 119,04 | 58,73 | 7,10 | 3 |
| 9 | 0000263 | Buku | 65,03 | 38,28 | 46,40 | 67,33 | 17,00 | 17,00 | 5 |
| 10 | 0000021 | Buku | 0,00 | 102,77 | 111,20 | 6,49 | 65,87 | 0,00 | 1 |
| 11 | 0000613 | Buku Binder | 116,01 | 19,48 | 7,38 | 118,31 | 58,16 | 7,38 | 3 |
| 12 | 0000606 | Buku Binder | 118,22 | 19,32 | 7,54 | 120,33 | 59,50 | 7,54 | 3 |
| 13 | 0001393 | Buku Diary | 108,84 | 13,61 | 5,13 | 111,16 | 50,51 | 5,13 | 3 |
| 14 | 0001395 | Buku Diary | 106,21 | 16,04 | 9,47 | 108,71 | 49,47 | 9,47 | 3 |
| 15 | 0000028 | Buku Kwintansi | 85,27 | 18,41 | 27,50 | 87,33 | 24,92 | 18,41 | 2 |
| 16 | 0000027 | Buku Kwintansi | 98,80 | 6,58 | 13,38 | 100,84 | 38,64 | 6,58 | 2 |
| 17 | 0000039 | Buku Nota | 61,82 | 42,87 | 53,27 | 62,90 | 9,22 | 9,22 | 5 |
| 18 | 0000794 | Buku Nota | 63,92 | 41,59 | 52,25 | 65,08 | 5,84 | 5,84 | 5 |
| 19 | 0001350 | Buku Sampul | 6,49 | 104,49 | 113,26 | 0,00 | 67,33 | 0,00 | 4 |
| 20 | 0000783 | Buku Sampul | 99,22 | 11,17 | 12,45 | 101,50 | 41,49 | 11,17 | 2 |
| 21 | 000050 | File | 62,1 | 41,5 | 49,1 | 64,3 | 21,1 | 21,19 | 5 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3 | | 4 | 7 | 2 | 8 | 9 | | |
| 22 | 0001011 | File | 111,84 | 12,18 | 1,01 | 113,88 | 52,28 | 1,01 | 3 |
| 23 | 0000347 | Gunting | 95,20 | 12,53 | 16,77 | 97,52 | 37,10 | 12,53 | 2 |
| 24 | 0000832 | Gunting | 94,96 | 15,27 | 18,75 | 97,49 | 36,98 | 15,27 | 2 |
| 25 | 0000674 | Kalkulator | 119,03 | 20,33 | 8,51 | 121,16 | 60,46 | 8,51 | 3 |
| 26 | 0000972 | Kalkulator | 117,55 | 20,32 | 8,15 | 119,81 | 59,55 | 8,15 | 3 |
| 27 | 0000095 | Kertas | 112,24 | 210,05 | 220,06 | 109,15 | 169,87 | 109,15 | 4 |
| 28 | 0001169 | Kertas | 52,17 | 78,75 | 84,25 | 57,82 | 50,85 | 50,85 | 5 |
| 29 | 0000360 | Kertas HVS | 118,22 | 19,32 | 7,54 | 120,33 | 59,50 | 7,54 | 3 |
| 30 | 0000358 | Kertas HVS | 69,42 | 38,11 | 43,93 | 72,48 | 25,50 | 25,50 | 5 |
| 31 | 0001566 | Kotak Makan | 117,47 | 19,26 | 7,20 | 119,64 | 59,02 | 7,20 | 3 |
| 32 | 0000651 | Kotak Makan | 116,65 | 18,25 | 6,19 | 118,81 | 58,05 | 6,19 | 3 |
| 33 | 0001511 | Kotak Pensil | 109,57 | 13,03 | 3,70 | 111,82 | 50,90 | 3,70 | 3 |
| 34 | 0001512 | Kotak Pensil | 111,20 | 12,20 | 0,00 | 113,26 | 51,96 | 0,00 | 3 |
| 35 | 0000296 | Lakban | 50,84 | 53,03 | 61,46 | 53,47 | 19,79 | 19,79 | 5 |
| 36 | 0001311 | Lakban | 44,38 | 60,34 | 67,76 | 47,59 | 29,79 | 29,79 | 5 |
| 37 | 0000352 | Lem | 77,59 | 29,30 | 35,37 | 80,37 | 25,41 | 25,41 | 5 |
| 38 | 0000717 | Lem | 90,52 | 23,29 | 25,40 | 93,47 | 36,85 | 23,29 | 2 |
| 39 | 0000165 | Map / File | 106,70 | 186,88 | 198,29 | 102,36 | 147,17 | 102,36 | 4 |
| 40 | 0000785 | Map / File | 64,94 | 77,39 | 89,42 | 63,20 | 42,16 | 42,16 | 5 |
| 41 | 0000697 | Misc | 117,41 | 18,30 | 6,58 | 119,50 | 58,54 | 6,58 | 3 |
| 42 | 0000892 | Misc | 102,77 | 0,00 | 12,20 | 104,49 | 41,26 | 0,00 | 2 |
| 43 | 0001578 | Misc Ultah | 89,32 | 14,19 | 25,13 | 91,00 | 27,12 | 14,19 | 2 |
| 44 | 0000711 | Misc Ultah | 98,13 | 7,40 | 13,75 | 100,19 | 38,36 | 7,40 | 2 |
| 45 | 0000083 | Misc ATK | 97,10 | 10,92 | 15,70 | 99,39 | 37,65 | 10,92 | 2 |
| 46 | 0000364 | Misc ATK | 106,82 | 15,56 | 8,72 | 109,29 | 49,65 | 8,72 | 3 |
| 47 | 0001419 | Pena Fancy | 50,56 | 84,98 | 90,03 | 56,61 | 57,61 | 50,56 | 1 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 48 | 0001529 | Pena Fancy | 74,28 | 41,61 | 45,79 | 78,05 | 32,08 | 32,08 | 5 |
| 49 | 0000242 | Pena Pulpen | 202,53 | 305,26 | 313,54 | 200,89 | 267,13 | 200,89 | 4 |
| 50 | 0000505 | Pena Pulpen | 53,43 | 147,78 | 155,37 | 56,08 | 110,29 | 53,43 | 1 |
| 51 | 0000782 | Pensil | 47,82 | 63,99 | 70,91 | 52,15 | 33,31 | 33,31 | 5 |
| 52 | 0000619 | Pensil | 46,77 | 60,32 | 67,70 | 50,47 | 28,90 | 28,90 | 5 |
| 53 | 0001486 | Post It / Stick Note / Memo | 110,32 | 12,58 | 2,29 | 112,49 | 51,32 | 2,29 | 3 |
| 54 | 0001489 | Post It / Stick Note / Memo | 111,07 | 12,30 | 1,01 | 113,18 | 51,78 | 1,01 | 3 |
| 55 | 0000715 | Stapler / Staples | 60,08 | 44,40 | 51,46 | 62,61 | 23,43 | 23,43 | 5 |
| 56 | 0000068 | Stapler / Staples | 36,63 | 138,02 | 146,25 | 37,29 | 99,99 | 36,63 | 1 |
| 57 | 0001142 | Tas | 100,85 | 17,65 | 15,23 | 103,55 | 44,89 | 15,23 | 3 |
| 58 | 0001134 | Tas | 100,82 | 18,31 | 15,95 | 103,57 | 44,91 | 15,95 | 3 |
| 59 | 0001227 | Tip Ex | 75,60 | 37,47 | 42,17 | 79,08 | 29,43 | 29,43 | 5 |
| 60 | 0000330 | Tip Ex | 76,60 | 34,74 | 39,84 | 79,88 | 27,90 | 27,90 | 5 |
| 61 | 0001129 | Tissue | 90,54 | 23,98 | 26,01 | 93,54 | 37,00 | 23,98 | 2 |
| 62 | 0001131 | Tissue | 110,32 | 12,58 | 2,29 | 112,49 | 51,32 | 2,29 | 3 |

4. Tahap Keempat

Untuk setiap cluster membuat pilihan centroid baru

5. Tahap Kelima

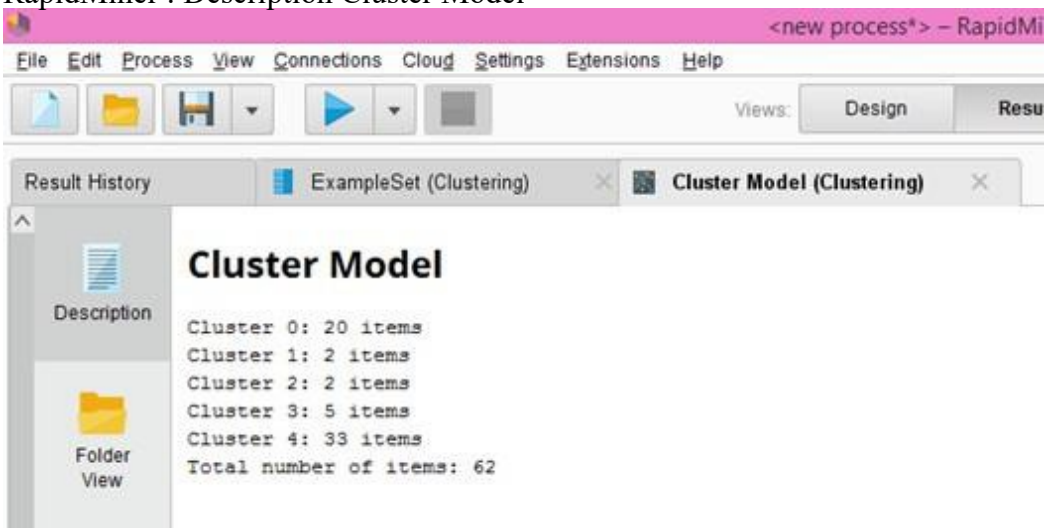
Ulangi langkah 2 dan 3 sampai nilai dari titik centroid tidak berubah lagi, maksudnya dengan melakukan perhitungan kembali (perulangan) seperti langkah ke-2 dan ke-3 yang sudah dilakukan diatas sampai benar-benar menemukan persamaan nilai dari titik centroid yang sudah tidak berubah atau bergeser lagi.

Setelah kelima tahapan selesai dilakukan, maka diperoleh hasil akhirnya yaitu untuk proses perhitungan k-means clustering secara manual yang telah selesai dilakukan pada data sampel yang ada, dan setelah perhitungan centroid yang baru berhenti pada iterasi ke-7, maka dapat diketahui untuk hasil dari 5 cluster yang telah ditentukan sebelumnya, yaitu sebagai berikut :

| No | Kode Item | Jenis | Jumlah Terjual | Jumlah Stok | Volume Penjualan | Rata-rata |
|---|---|---|---|---|---|---|
| 1 | 0000007 | Amplop | 82,00 | 120,00 | 18,00 | 36,67 |
| 2 | 0000021 | Buku | 70,00 | 96,00 | 12,00 | 29,67 |
| 3 | 0001350 | Buku Sampul | 69,00 | 100,00 | 7,00 | 29,33 |
| 4 | 0000505 | Pena Pulpen | 95,00 | 120,00 | 50,00 | 44,17 |
| 5 | 0000068 | Stapler / Staples | 91,00 | 120,00 | 27,00 | 39,67 |
| 6 | Jumlah | 5 | 407,00 | 556,00 | 114,00 | 179,50 |
| 7 | Rata-rata | - | 81,40 | 111,20 | 22,80 | 35,90 |

Berikut Hasil Pengukuran Performance terhadap clustering di atas menggunakan tools

RapidMiner : Description Cluster Model



Data Cluster 0

eSet (62 examples, 4 special attributes, 4 regular attributes) Filter (62 / 62 examples): all

| cluster | Kode Item | Jenis | Jumlah Terj... | Jumlah Stok | Volume Penj... | Rata-Rata |
|---|---|---|---|---|---|---|
| cluster_0 | 0000369 | Alat Gambar | 17 | 60 | 10 | 14.500 |
| cluster_0 | 0000882 | Alat Tulis | 33 | 60 | 19 | 18.667 |
| cluster_0 | 0000006 | Amplop | 52 | 72 | 6 | 21.667 |
| cluster_0 | 0000263 | Buku | 29 | 48 | 9 | 14.333 |
| cluster_0 | 0000039 | Buku Nota | 23 | 60 | 3 | 14.333 |
| cluster_0 | 0000794 | Buku Nota | 20 | 60 | 5 | 14.167 |
| cluster_0 | 0000503 | File | 34 | 48 | 6 | 14.667 |
| cluster_0 | 0001169 | Kertas | 52 | 60 | 45 | 26.167 |
| cluster_0 | 0000358 | Kertas HVS | 32 | 40 | 15 | 14.500 |
| cluster_0 | 0000296 | Lakban | 36 | 60 | 14 | 18.333 |
| cluster_0 | 0001311 | Lakban | 46 | 60 | 14 | 20 |
| cluster_0 | 0000352 | Lem | 24 | 36 | 14 | 12.333 |
| cluster_0 | 0000785 | Map / File | 7 | 100 | 2 | 18.167 |
| cluster_0 | 0001419 | Pena Fancy | 60 | 60 | 46 | 27.667 |
| cluster_0 | 0001529 | Pena Fancy | 31 | 36 | 26 | 15.500 |

Data *Cluster* 1,2,3, dan 4

eSet (62 examples, 4 special attributes, 4 regular attributes) Filter (62 / 62 examples): all

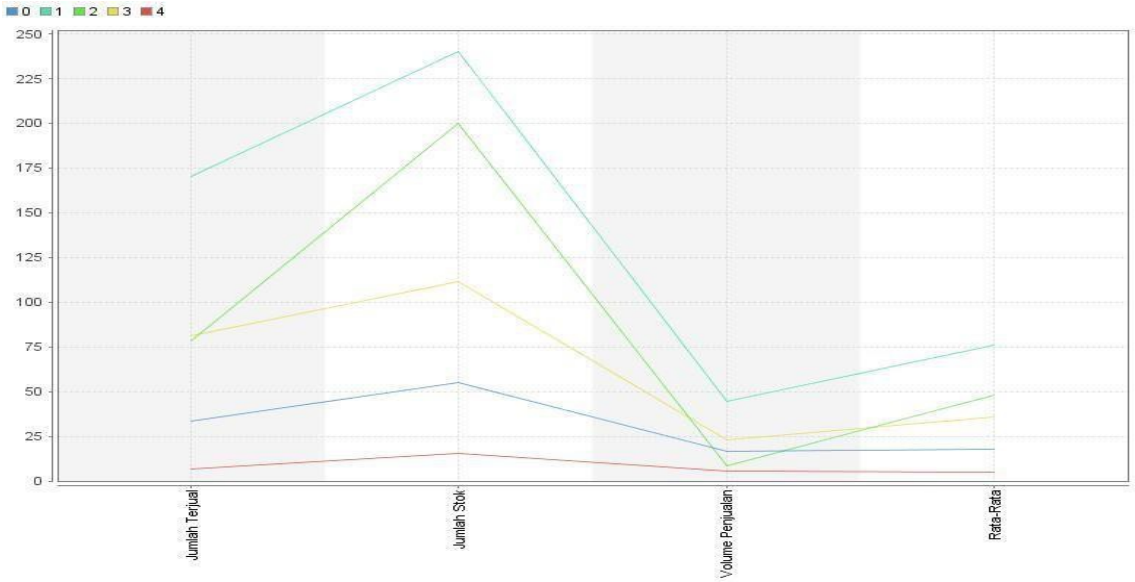
| cluster | Kode Item | Jenis | Jumlah Terj... | Jumlah Stok | Volume Penj... | Rata-Rata |
|---|---|---|---|---|---|---|
| cluster_1 | 0000466 | Alat Tulis | 138 | 240 | 62 | 73.333 |
| cluster_1 | 0000242 | Pena Pulpen | 203 | 240 | 27 | 78.333 |
| cluster_2 | 0000095 | Kertas | 105 | 200 | 14 | 53.167 |
| cluster_2 | 0000165 | Map / File | 52 | 200 | 3 | 42.500 |
| cluster_3 | 0000007 | Amplop | 82 | 120 | 18 | 36.667 |
| cluster_3 | 0000021 | Buku | 70 | 96 | 12 | 29.667 |
| cluster_3 | 0001350 | Buku Sampul | 69 | 100 | 7 | 29.333 |
| cluster_3 | 0000505 | Pena Pulpen | 95 | 120 | 50 | 44.167 |
| cluster_3 | 0000068 | Stapler / Stap... | 91 | 120 | 27 | 39.667 |
| cluster_4 | 0001174 | Alat Gambar | 12 | 24 | 12 | 8 |
| cluster_4 | 0000635 | Botol Minum | 6 | 10 | 4 | 3.333 |
| cluster_4 | 0001404 | Botol Minum | 3 | 5 | 2 | 1.667 |
| cluster_4 | 0000613 | Buku Binder | 4 | 5 | 4 | 2.167 |
| cluster_4 | 0000606 | Buku Binder | 1 | 5 | 1 | 1.167 |
| cluster_4 | 0001393 | Buku Diary | 6 | 12 | 6 | 4 |

Hasil Centroid Akhir

Description

Folder View

| Attribute | cluster_0 | cluster_1 | cluster_2 | cluster_3 | cluster_4 |
|---|---|---|---|---|---|
| Jumlah Terjual | 33.700 | 170.500 | 78.500 | 81.400 | 6.909 |
| Jumlah Stok | 55 | 240 | 200 | 111.200 | 15.242 |
| Volume Penjualan | 16.750 | 44.500 | 8.500 | 22.800 | 5.788 |
| Rata-Rata | 17.575 | 75.833 | 47.833 | 35.900 | 4.657 |




Berikut hasil Grafik menggunakan Rapid Miner :

Nama : Siti Ratu Delima

Nim : 202420025

Kelas : MTI24

TUGAS 05

Pelari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format ms word dan dilengkapi dengan gambar dan keterangan yang jelas untuk tiap langkah-langkahnya. Sertkan sumber sumber rujukan yang anda gunakan. Kumpulkan sebelum waktu batas akhir pengumpulan tugas.

Jawaban :

Pada tugas ini saya coba mengambil sample untuk mengukur performance Clustering dengan menggunakan K-Means, pengukuran performance clustering saat ini menggunakan clusterisasi sampai didapatkan nilai stabil dan dengan rapidminer saya coba untuk bandingkan hasilnya dengan sumber data yang sama, adapun datanya sebagai berikut :

Data Komsumsi Listrik bulan Mei 2020 di Prov Lampung :

| | Kabupaten/Kota | Latitude | Longitude | Mei 2020 | | |
|---|---|---|---|---|---|---|
| | | | | Rumah Tangga | Industri | Bisnis |
| | Kabupaten Tulang Bawang Barat | -4.52569 | 105.07912 | 8647851 | 1424419 | 771600 |
| | Kabupaten Way Kanan | -4.49636 | 104.56552 | 8922355 | 904997 | 580164 |
| | Kabupaten Lampung Barat | -5.10952 | 104.1466 | 7870512 | 908299 | 1508881 |
| | Kabupaten Lampung Selatan | -5.56226 | 105.54743 | 39811088 | 28596132 | 6923204 |
| | Kabupaten Lampung Tengah | -4.8008 | 105.31311 | 43819603 | 17071293 | 5099183 |
| | Kabupaten Lampung Timur | -5.11349 | 105.68817 | 17150724 | 4651063 | 3006046 |
| | Kabupaten Lampung Utara | -4.81339 | 104.75209 | 21682207 | 3629202 | 2188612 |
| | Kabupaten Pringsewu | -5.33311 | 104.98561 | 17504659 | 701278 | 1618426 |
| | Kabupaten Tanggamus | -5.30274 | 104.56552 | 14831519 | 2175713 | 1911240 |
| | Kabupaten Tulang Bawang | -4.31765 | 105.50054 | 13723285 | 1574286 | 1730416 |
| | Kota Bandar Lampung | -5.39713 | 105.26678 | 60305690 | 7951584 | 18855425 |
| | Kota Metro | -5.11783 | 105.30726 | 17774258 | 753564 | 2565855 |

Dari data diatas saya coba untuk mengukur bagaimana performance dari clustering yang menggunakan K-Means dengan membandingkan 2 cara yaitu dengan Coding PhP dan Rapidminer.

Berbasiskan Web

Stage Awal Data

Proses Iterasi Selanjutnya

| | Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rumah Tangga | Industri | Bisnis | 4035984 | 396909 | 14840900 | 396909 | 4035984 | 14840900 | 14840900 | 396909 | 4035984 | | | |
| | Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 14841505 | | | 16517975 | | | 7075721 | | | 0 | 0 | 1 |
| | Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 15083214 | | | 16907244 | | | 6872410 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 13881925 | | | 15600637 | | | 7431963 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 46235745 | | | 47110171 | | | 37776216 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 44222985 | | | 46371873 | | | 33450421 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 18170250 | | | 20521510 | | | 4949129 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 21952614 | | | 24765091 | | | 7788708 | | | 0 | 0 | 1 |
| | Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 18876749 | | | 21877595 | | | 3610102 | | | 0 | 0 | 1 |
| | Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 16937645 | | | 19467785 | | | 2771059 | | | 0 | 0 | 1 |
| | Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 16343647 | | | 18855689 | | | 2819738 | | | 0 | 0 | 1 |
| | Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 56916336 | | | 60170677 | | | 48412148 | | | 0 | 0 | 1 |
| | Kota Metro | 17774258 | 753564 | 2565855 | 18426722 | | | 21527268 | | | 3300465 | | | 0 | 0 | 1 |

Iterasi 1

**Iterasi Ke-1**

| | Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rumah Tangga | Industri | Bisnis | 0 | 0 | 0 | 0 | 0 | 0 | 22670312 | 5861819 | 3896587 | | | |
| | Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 8798276 | | | 8798276 | | | 15036139 | | | 1 | 1 | 0 |
| | Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 8986880 | | | 8986880 | | | 14985829 | | | 1 | 1 | 0 |
| | Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 8065152 | | | 8065152 | | | 15788369 | | | 1 | 1 | 0 |
| | Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 49503456 | | | 49503456 | | | 28632421 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 47303153 | | | 47303153 | | | 23966456 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 18022653 | | | 18022653 | | | 5720564 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 22092515 | | | 22092515 | | | 2979616 | | | 0 | 0 | 1 |
| | Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 17593299 | | | 17593299 | | | 7648868 | | | 0 | 0 | 1 |
| | Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 15111602 | | | 15111602 | | | 8886825 | | | 0 | 0 | 1 |
| | Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 13921252 | | | 13921252 | | | 10155025 | | | 0 | 0 | 1 |
| | Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 63683051 | | | 63683051 | | | 40553120 | | | 0 | 0 | 1 |
| | Kota Metro | 17774258 | 753564 | 2565855 | 17974307 | | | 17974307 | | | 7199754 | | | 0 | 0 | 1 |

**Iterasi Ke-2**

| | Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rumah Tangga | Industri | Bisnis | 8480239 | 1079238 | 953548 | 8480239 | 1079238 | 953548 | 27400337 | 7456012 | 4877600 | | | |
| | Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 424674 | | | 424674 | | | 20122005 | | | 1 | 1 | 0 |
| | Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 604352 | | | 604352 | | | 20070365 | | | 1 | 1 | 0 |
| | Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 842248 | | | 842248 | | | 20871868 | | | 1 | 1 | 0 |
| | Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 42124082 | | | 42124082 | | | 24599103 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 39010289 | | | 39010289 | | | 19028794 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 9599374 | | | 9599374 | | | 10790042 | | | 1 | 1 | 0 |
| | Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 13502579 | | | 13502579 | | | 7387296 | | | 0 | 0 | 1 |
| | Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 9056769 | | | 9056769 | | | 12416646 | | | 1 | 1 | 0 |
| | Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 6515994 | | | 6515994 | | | 13951919 | | | 1 | 1 | 0 |
| | Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 5323356 | | | 5323356 | | | 15217136 | | | 1 | 1 | 0 |
| | Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 55259240 | | | 55259240 | | | 35754544 | | | 0 | 0 | 1 |
| | Kota Metro | 17774258 | 753564 | 2565855 | 9438452 | | | 9438452 | | | 11955265 | | | 1 | 1 | 0 |

Iterasi 3

**Iterasi Ke-3**

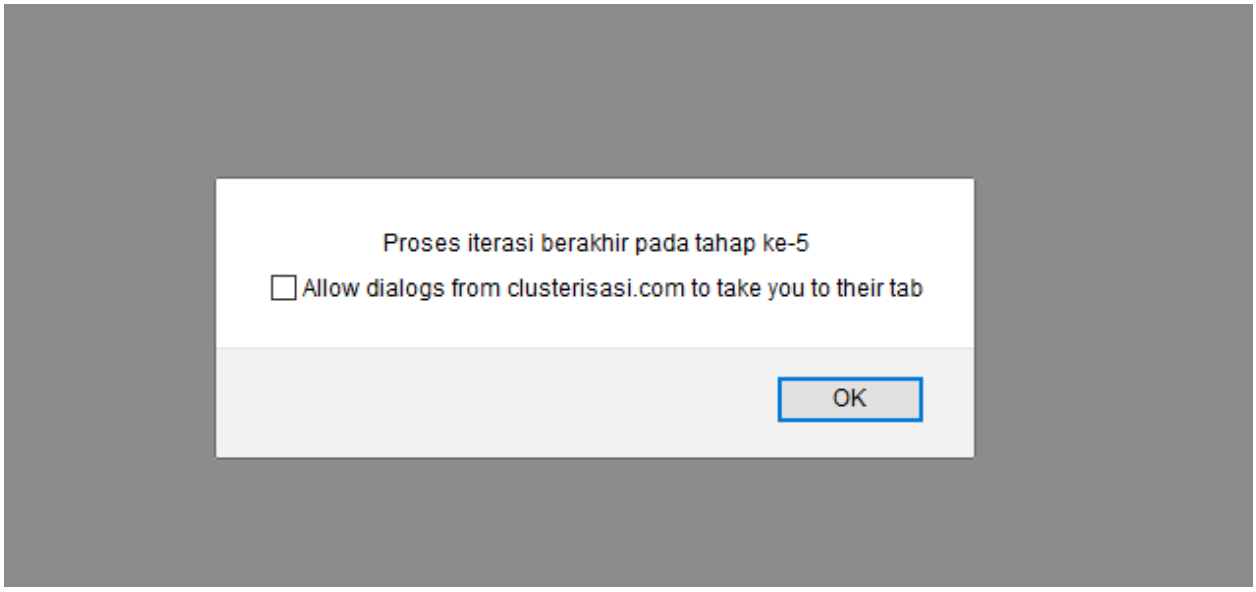
| Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rumah Tangga | Industri | Bisnis | 13303145 | 1636702 | 1711578 | 13303145 | 1636702 | 1711578 | 41404647 | 14312052 | 8266606 | | | |
| Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 4753986 | | | 4753986 | | | 35989913 | | | 1 | 1 | 0 |
| Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 4583319 | | | 4583319 | | | 35971236 | | | 1 | 1 | 0 |
| Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 5484994 | | | 5484994 | | | 36740518 | | | 1 | 1 | 0 |
| Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 38165991 | | | 38165991 | | | 14435341 | | | 0 | 0 | 1 |
| Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 34365049 | | | 34365049 | | | 4845408 | | | 0 | 0 | 1 |
| Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 5056271 | | | 5056271 | | | 26631954 | | | 1 | 1 | 0 |
| Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 8625908 | | | 8625908 | | | 23238759 | | | 1 | 1 | 0 |
| Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 4305393 | | | 4305393 | | | 28295952 | | | 1 | 1 | 0 |
| Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 1632887 | | | 1632887 | | | 29896697 | | | 1 | 1 | 0 |
| Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 425168 | | | 425168 | | | 31164567 | | | 1 | 1 | 0 |
| Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 50428448 | | | 50428448 | | | 22579372 | | | 0 | 0 | 1 |
| Kota Metro | 17774258 | 753564 | 2565855 | 4636870 | | | 4636870 | | | 27833908 | | | 1 | 1 | 0 |

Iterasi 4

**Iterasi Ke-4**

| Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rumah Tangga | Industri | Bisnis | 14234152 | 1858091 | 1764582 | 14234152 | 1858091 | 1764582 | 47978793 | 17873003 | 10292604 | | | |
| Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 5690417 | | | 5690417 | | | 43682130 | | | 1 | 1 | 0 |
| Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 5525072 | | | 5525072 | | | 43676654 | | | 1 | 1 | 0 |
| Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 6439208 | | | 6439208 | | | 44425547 | | | 1 | 1 | 0 |
| Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 37359253 | | | 37359253 | | | 13894235 | | | 0 | 0 | 1 |
| Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 33434412 | | | 33434412 | | | 6701733 | | | 0 | 0 | 1 |
| Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 4224726 | | | 4224726 | | | 34326134 | | | 1 | 1 | 0 |
| Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 7667473 | | | 7667473 | | | 30985012 | | | 1 | 1 | 0 |
| Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 3472145 | | | 3472145 | | | 36038623 | | | 1 | 1 | 0 |
| Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 692271 | | | 692271 | | | 37621722 | | | 1 | 1 | 0 |
| Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 585404 | | | 585404 | | | 38889575 | | | 1 | 1 | 0 |
| Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 49515797 | | | 49515797 | | | 17991910 | | | 0 | 0 | 1 |
| Kota Metro | 17774258 | 753564 | 2565855 | 3793990 | | | 3793990 | | | 35568129 | | | 1 | 1 | 0 |

Iterasi 5




Pada saat iterasi ke 5 data sudah didapatkan sampai dengan semua nilainya stabil atau sama, sehingga didapat nilai cluster yang sesuai sebagai berikut :

Iterasi ke-1

| C1 | C2 | C3 |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |

## Iterasi ke-2

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |

## Iterasi ke-3

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |

Iterasi ke-4

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |

Iterasi ke-5

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |

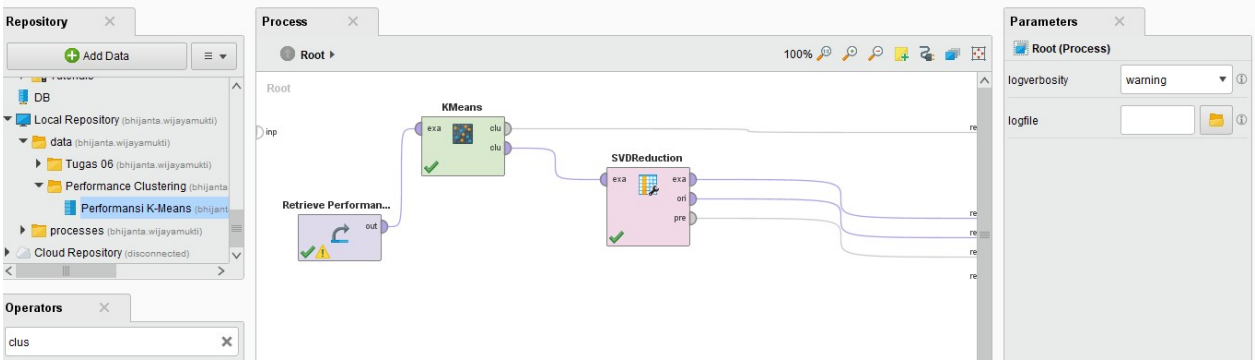
Dari Hasil Iterasi ke 5 didaptkan kesimpulan sebagai berikut :

| | Kabupaten/Kota | Bulan | Zona |
|---|---|---|---|
| | Kabupaten Tulang Bawang Barat | Mei | Kuning |
| | Kabupaten Way Kanan | Mei | Kuning |
| | Kabupaten Lampung Barat | Mei | Kuning |
| | Kabupaten Lampung Selatan | Mei | Hijau |
| | Kabupaten Lampung Tengah | Mei | Hijau |
| | Kabupaten Lampung Timur | Mei | Kuning |
| | Kabupaten Lampung Utara | Mei | Kuning |
| | Kabupaten Pringsewu | Mei | Kuning |
| | Kabupaten Tanggamus | Mei | Kuning |
| | Kabupaten Tulang Bawang | Mei | Kuning |
| | Kota Bandar Lampung | Mei | Hijau |
| | Kota Metro | Mei | Kuning |

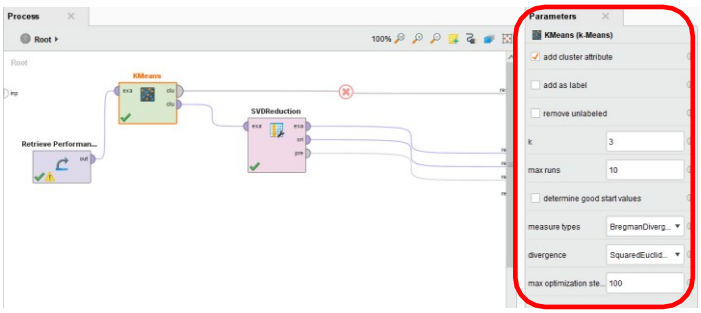
- Berbasiskan Rapidminer

  Dengan data yang sama seperti diatas saya coba mengukur bagaimana performance dari Clustering K-Means menggunakan tool rapidminer, adapun hasilnya sebagai berikut :

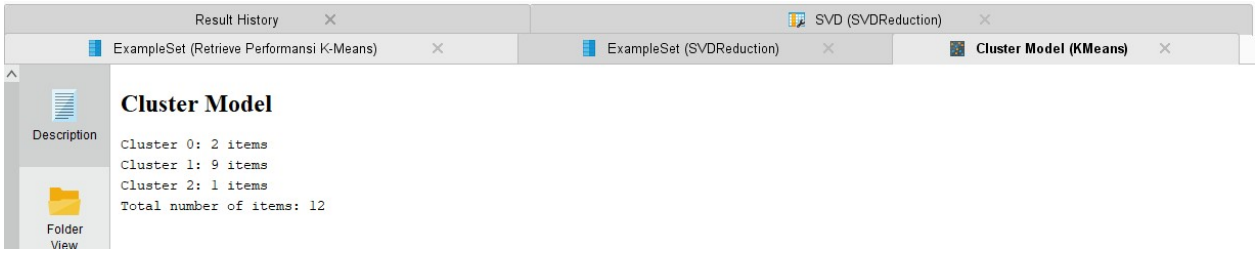
  Desain Mapping Rapidminer

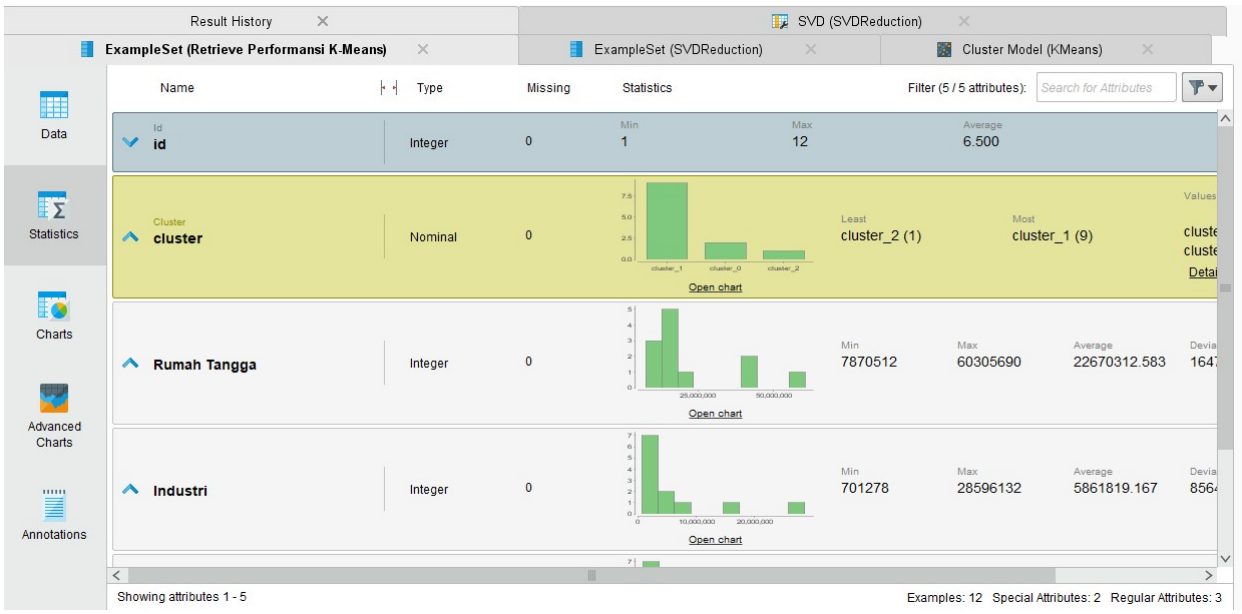


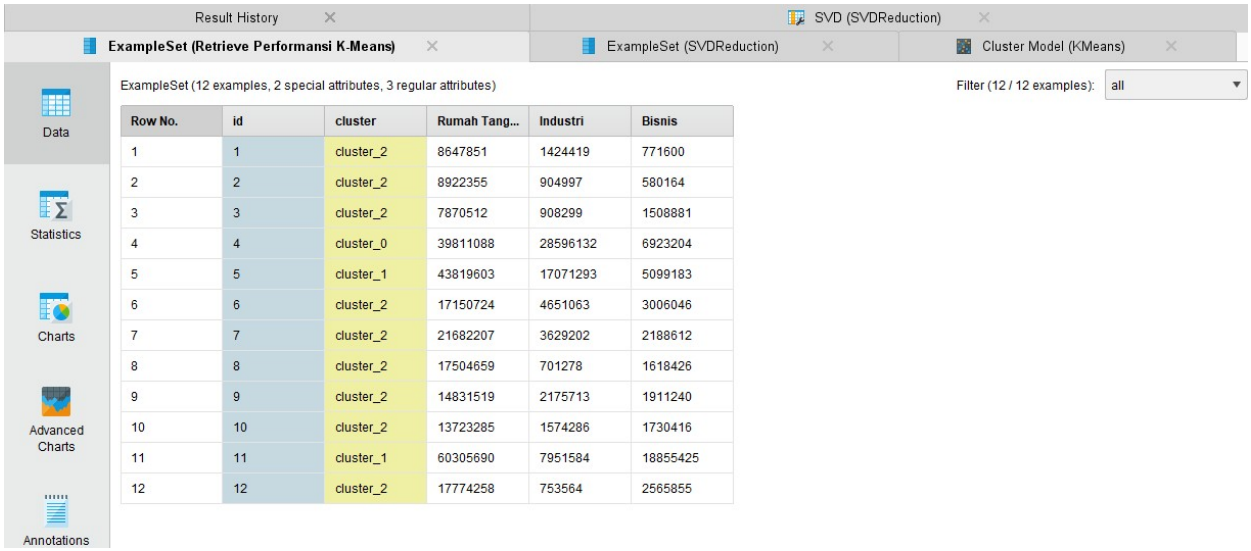

Parameter K-Means



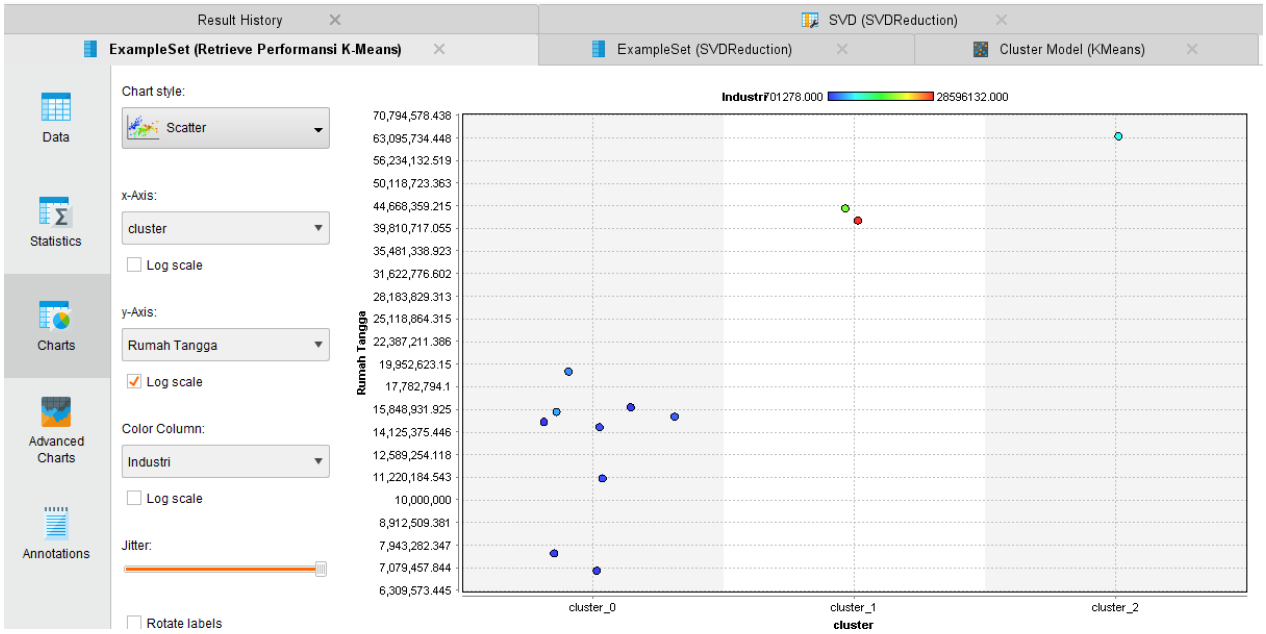
## Hasil Clustering




| Row No. | id | cluster | Rumah Tang... | Industri | Bisnis |
|---|---|---|---|---|---|
| 1 | 1 | cluster_2 | 8647851 | 1424419 | 771600 |
| 2 | 2 | cluster_2 | 8922355 | 904997 | 580164 |
| 3 | 3 | cluster_2 | 7870512 | 908299 | 1508881 |
| 4 | 4 | cluster_0 | 39811088 | 28596132 | 6923204 |
| 5 | 5 | cluster_1 | 43819603 | 17071293 | 5099183 |
| 6 | 6 | cluster_2 | 17150724 | 4651063 | 3006046 |
| 7 | 7 | cluster_2 | 21682207 | 3629202 | 2188612 |
| 8 | 8 | cluster_2 | 17504659 | 701278 | 1618426 |
| 9 | 9 | cluster_2 | 14831519 | 2175713 | 1911240 |
| 10 | 10 | cluster_2 | 13723285 | 1574286 | 1730416 |
| 11 | 11 | cluster_1 | 60305690 | 7951584 | 18855425 |
| 12 | 12 | cluster_2 | 17774258 | 753564 | 2565855 |




Kesimpulan

*Sumber :https://clusterisasi.com/k-means/admin/iterasi_kmeans_hasil*

1. K-means merupakan salah satu algoritma clustering. Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan supervised learning yang menerima masukan berupa vektor (-x-1 , y1) , (-x-2 , y2) , …, (-x-i , yi), di mana xi merupakan data dari suatu data pelatihan dan yi merupakan label kelas untuk xi
2. Kelebihan pada algoritma k-means, yaitu [2]:
    - Mudah untuk diimplementasikan dan dijalankan.
    - Waktu yang dibutuhkan untuk menjalankan pembelajaran ini relatif cepat.
    - Mudah untuk diadaptasi.
    - Umum digunakan.
3. Secara performance dari Clustering dengan K-Means cukup dengan baik dapat menyelesaikan clustering baik secara metode Web dan Rapidminer.

Jawaban Tugas 5
Nama : Surta Wijaya
NIM : 202420014

Ada beberapa cara untuk mengukur performa clustering yaitu :

- Accuracy
- Confusion matrix
- Log-loss
- Precision and Recall
- F-Scores
- Receiver operating characteristic (ROC) curve
- Area under curve (AUC) ("curve" corresponds to the ROC curve)

## Accuracy

Akurasi hanya mengukur seberapa sering pengklasifikasi membuat prediksi yang benar. Ini adalah rasio antara jumlah prediksi yang benar dan jumlah total prediksi (jumlah poin data pengujian).

```
import turicreate as tc

y    = tc.SArray(["cat", "dog", "cat", "cat"])
yhat = tc.SArray(["cat", "dog", "cat", "dog"])

print(tc.evaluation.accuracy(y, yhat))
0.75
```

Akurasi terlihat cukup mudah. Namun, tidak ada perbedaan antara kelas; jawaban yang benar untuk setiap kelas diperlakukan sama. Terkadang ini tidak cukup apabila ingin melihat berapa banyak contoh yang gagal untuk setiap kelas. Ini akan menjadi kasus jika biaya kesalahan klasifikasi berbeda, atau jika memiliki lebih banyak data pengujian dari satu kelas daripada yang lain. Misalnya, membuat panggilan bahwa pasien mengidap kanker padahal dia tidak (dikenal sebagai positif palsu) memiliki konsekuensi yang sangat berbeda daripada membuat panggilan bahwa pasien tidak menderita kanker ketika dia mengidapnya (negatif palsu).

Multiclass Averaging
Pengaturan multi-kelas merupakan perluasan dari pengaturan biner. Metrik akurasi dapat "dirata-ratakan" di semua kelas dengan berbagai cara yang memungkinkan. Beberapa di antaranya adalah:

- **mikro**: Menghitung metrik secara global dengan menghitung total berapa kali setiap kelas diprediksi dengan benar dan diprediksi secara salah.
- **makro**: Menghitung metrik untuk setiap "kelas" secara independen, dan menemukan rata-rata tak berbobotnya. Ini tidak memperhitungkan ketidakseimbangan label.
- **Tidak ada**: Menampilkan metrik yang sesuai dengan setiap kelas, yang merupakan presisi untuk setiap kelas.

```
import turicreate as tc

y    = tc.SArray(["cat", "dog", "foosa", "cat"])
yhat = tc.SArray(["cat", "dog", "cat", "dog"])
```

```
print(tc.evaluation.accuracy(y, yhat, average = "micro"))
print(tc.evaluation.accuracy(y, yhat, average = "macro"))
print(tc.evaluation.accuracy(y, yhat, average = None))
0.5
0.6666666666666666
{'foosa': None, 'dog': 0.5, 'cat': 0.5}
```

Secara umum, jika terdapat jumlah contoh yang berbeda per kelas, akurasi rata-rata per kelas akan berbeda dari akurasi rata-rata mikro. Jika kelas tidak seimbang, yaitu jumlah contoh satu kelas lebih banyak dari kelas lainnya, maka akurasi akan memberikan gambaran yang sangat terdistorsi, karena kelas dengan jumlah contoh yang lebih banyak akan mendominasi statistik. Dalam hal ini, Anda harus melihat rata-rata akurasi per kelas (rata-rata = "mikro"), serta angka akurasi per kelas individu (rata-rata = Tidak ada). Akurasi per kelas bukannya tanpa peringatannya sendiri, namun: misalnya, jika ada sangat sedikit contoh dari satu kelas, statistik pengujian untuk kelas itu tidak akan dapat diandalkan (yaitu, mereka memiliki varians yang besar), jadi tidak masuk akal secara statistik untuk jumlah rata-rata dengan derajat varian yang berbeda.
Perhatikan bahwa `average` parameter diteruskan ke `tc.evaluation.accuracy` mungkin akan segera dihentikan, dan akan merusak kode yang menggunakan API ini.

**Confusion matrix**

Confusion matrix (atau confusion table) menunjukkan rincian yang lebih rinci tentang klasifikasi yang benar dan salah untuk setiap kelas. Berikut adalah contoh bagaimana matriks kebingungan dapat dihitung.
Tabel kebingungan adalah SFrame yang terdiri dari tiga kolom:

- **target_label**: Label kebenaran dasar.
- **predict_label**: Label yang diprediksi.
- **count**: Berapa kali `target_label` diprediksi sebagai `prediksi_label`.

```
y    = tc.SArray(["cat", "dog", "foosa", "cat"])
yhat = tc.SArray(["cat", "dog", "cat", "dog"])

cf_matrix = tc.evaluation.confusion_matrix(y, yhat)

Columns:
    target_label    str
    predicted_label    str
    count    int

Rows: 4

Data:
+--------------+-----------------+-------+
| target_label | predicted_label | count |
+--------------+-----------------+-------+
|    foosa     |       cat       |   1   |
|     cat      |       dog       |   1   |
|     dog      |       dog       |   1   |
|     cat      |       cat       |   1   |
+--------------+-----------------+-------+
[4 rows x 3 columns]
```

Mencari pada matriks, dapat dengan jelas mendapatkan gambaran yang lebih baik tentang kelas mana yang paling baik diidentifikasi oleh model. Informasi ini hilang jika seseorang hanya melihat keakuratannya secara keseluruhan.

### Log-loss

log, atau kerugian logaritmik, termasuk ke detail pengklasifikasi yang lebih halus. Secara khusus, jika keluaran mentah dari pengklasifikasi adalah probabilitas numerik dan bukan label kelas, maka kerugian log dapat digunakan. Probabilitas pada dasarnya berfungsi sebagai pengukur kepercayaan. Jika label sebenarnya adalah "0" tetapi pengklasifikasi menganggapnya milik kelas "1" dengan probabilitas 0,51, maka pengklasifikasi akan membuat kesalahan. Tapi itu nyaris meleset karena probabilitasnya sangat dekat dengan batas keputusan 0,5. Kerugian log adalah pengukuran akurasi "lunak" yang menggabungkan gagasan keyakinan probabilistik ini.
Secara matematis, log-loss untuk pengklasifikasi biner terlihat seperti ini:
Di sini, adalah probabilitas bahwa titik data ke-i termasuk dalam kelas "1", seperti yang dinilai oleh pengklasifikasi. adalah label yang benar dan bisa berupa "0" atau "1". Hal yang indah tentang definisi ini adalah bahwa ia terkait erat dengan teori informasi: Secara intuitif, log-loss mengukur ketidakpastian "kebisingan tambahan" yang berasal dari penggunaan prediktor sebagai lawan dari label sebenarnya. Dengan meminimalkan cross entropy, kami memaksimalkan akurasi pengklasifikasi.
Berikut adalah contoh bagaimana ini dihitung:

```
import turicreate as tc

targets = tc.SArray([0, 1, 1, 0])
predictions = tc.SArray([0.1, 0.35, 0.7, 0.99])

log_loss = tc.evaluation.log_loss(targets, predictions)
```

```
1.5292569425208318
```

# Multi-class log-loss

Dalam pengaturan multi-kelas, log-loss memerlukan vektor probabilitas (yang berjumlah 1) untuk setiap label kelas dalam set data masukan. Dalam contoh ini, ada tiga kelas [0, 1, 2], dan vektor probabilitas sesuai dengan probabilitas prediksi untuk masing-masing dari tiga kelas (sambil mempertahankan urutan).

```
targets     = tc.SArray([ 1, 0, 2, 1])
predictions = tc.SArray([[.1, .8, 0.1],
                         [.9, .1, 0.0],
                         [.8, .1, 0.1],
                         [.3, .6, 0.1]])

log_loss = tc.evaluation.log_loss(targets, predictions)
```

```
0.785478695933018
```

Untuk klasifikasi kelas jamak, jika label target berjenis **string**, maka vektor probabilitas diasumsikan sebagai vektor probabilitas kelas yang diurutkan secara alfanumerik. Oleh karena itu, untuk vektor probabilitas [0,1, 0,2, 0,7] untuk kumpulan data dengan kelas ["kucing", "anjing", "tikus"; 0.1 untuk "kucing", 0.2 untuk "anjing", dan 0.7 untuk "tikus".

```
target    = tc.SArray([ "dog", "cat", "foosa", "dog"])
predictions = tc.SArray([[.1, .8, 0.1],
                         [.9, .1, 0.0],
                         [.8, .1, 0.1],
                         [.3, .6, 0.1]])
log_loss = tc.evaluation.log_loss(y, yhat)
```

```
1.5292569425208318
```

## Precision and Recall

Precision and Recall sebenarnya adalah dua metrik yang sering digunakan bersama. Presisi menjawab pertanyaan: *Dari item yang diprediksi oleh pengklasifikasi benar, berapa banyak yang sebenarnya benar?* Sedangkan, ingat kembali jawaban pertanyaan: *Dari semua item yang benar, berapa banyak yang ditemukan benar oleh pengklasifikasi?*
Nilai presisi mengukur kemampuan pengklasifikasi untuk tidak memberi label **negatif** contohsebagai **positif**. Nilai presisi dapat diartikan sebagai probabilitas bahwa **positif yang** prediksidibuat oleh pengklasifikasi bernilai **positif**. Nilai tersebut berada pada kisaran [0,1] dengan 0 sebagai yang terburuk, dan 1 adalah sempurna.
Nilai presisi dandapat didefinisikan sebagai:

```
targets = turicreate.SArray([0, 1, 0, 0, 0, 1, 0, 0])
predictions = turicreate.SArray([1, 0, 0, 1, 0, 1, 0, 1])

pr_score   = turicreate.evaluation.precision(targets, predictions)
rec_score  = turicreate.evaluation.recall(targets, predictions)
print(pr_score, rec_score)
```

```
0.25, 0.5
```

Presisi juga dapat ditentukan kemudian kelas target adalah tipe **string**. Untuk klasifikasi biner, jika label target berjenis **string**, maka label diurutkan secara alfanumerik dan label terbesar dianggap sebagai label "positif".

```
targets = turicreate.SArray(['cat', 'dog', 'cat', 'cat', 'cat', 'dog', 'cat', 'cat'])
predictions = turicreate.SArray(['dog', 'cat', 'cat', 'dog', 'cat', 'dog', 'cat',
'dog'])

pr_score   = turicreate.evaluation.precision(targets, predictions)
rec_score  = turicreate.evaluation.recall(targets, predictions)
print(pr_score, rec_score)
```

```
0.25, 0.5
```

## Multi-class precision-recall

nilai presisi dan recall juga dapat ditentukan dalam pengaturan multi-kelas. Di sini, metrik dapat "dirata-ratakan" di semua kelas dengan berbagai cara yang memungkinkan. Beberapa di antaranya adalah:

- **mikro**: Menghitung metrik secara global dengan menghitung total berapa kali setiap kelas diprediksi dengan benar dan diprediksi secara salah.
- **makro**: Menghitung metrik untuk setiap "kelas" secara independen, dan menemukan rata-rata tak berbobotnya. Ini tidak memperhitungkan ketidakseimbangan label.
- **Tidak ada**: Kembalikan nilai akurasi untuk setiap kelas. Sesuai dengan setiap kelas.

```
targets = turicreate.SArray(['cat', 'dog', 'cat', 'cat', 'cat', 'dog', 'cat',
'foosa'])
predictions = turicreate.SArray(['dog', 'cat', 'cat', 'dog', 'cat', 'dog', 'cat',
'foosa'])

macro_pr = turicreate.evaluation.precision(targets, predictions, average='macro')
micro_pr = turicreate.evaluation.precision(targets, predictions, average='micro')
per_class_pr = turicreate.evaluation.precision(targets, predictions, average=None)

print(macro_pr, micro_pr)
print(per_class_pr)

0.694444444444 0.625
{'foosa': 1.0, 'dog': 0.3333333333333333, 'cat': 0.75}
```

**Catatan:** Nilai presisi, recall, dan akurasi rata-rata mikro secara matematis setara.
**Undefined Precision-Recall** Nilai presisi (atau recall) tidak ditentukan ketika jumlah positif benar + positif palsu (positif benar + negatif palsu) adalah nol. Dengan kata lain, penyebut dari masing-masing persamaan adalah 0, metrik tidak ditentukan. Dalam pengaturan tersebut, kami mengembalikan nilai `Tidak Ada`. Dalam pengaturan multi-kelas,pembuatan `Tidak` Ada dilewati selamarata-rata.

## F-nilai (F1, F-beta)

F1-score adalah metrik tunggal yang menggabungkan presisi dan recall melalui rata-rata harmonisnya:
Nilai terletak pada kisaran [0,1] dengan 1 ideal dan 0 untuk terburuk. Berbeda dengan rata-rata aritmatika, nilai rata-rata harmonik cenderung lebih kecil dari kedua elemen. Oleh karena itu, nilai F1 akan kecil jika presisi atau recall kecil.
Nilai F1 (terkadang dikenal sebagai nilai F-beta seimbang), adalah kasus khusus dari metrik yang dikenal sebagai nilai F-Beta, yang *mengukur keefektifan pengambilan sehubungan dengan pengguna yang melampirkan waktu sebagai hal yang penting untuk diingat untuk presisi*.

```
targets = turicreate.SArray([0, 1, 0, 0, 0, 1, 0, 0])
predictions = turicreate.SArray([1, 0, 0, 1, 0, 1, 0, 1])

f1    = turicreate.evaluation.f1_score(targets, predictions)
```

```
fbeta = turicreate.evaluation.fbeta_score(targets, predictions, beta = 2.0)
print(f1, fbeta)
```

```
0.333333333333 0.416666666667
```

Seperti metrik lainnya, nilai F1 (atau nilai F-beta) juga dapat didefinisikan ketika kelas target berjenis **string**. Untuk klasifikasi biner, jika label target berjenis **string**, maka label diurutkan secara alfanumerik dan label terbesar dianggap sebagai label "positif".

```
targets = turicreate.SArray(['cat', 'dog', 'cat', 'cat', 'cat', 'dog', 'cat', 'cat'])
predictions = turicreate.SArray(['dog', 'cat', 'cat', 'dog', 'cat', 'dog', 'cat',
'dog'])
```

```
f1    = turicreate.evaluation.f1_score(targets, predictions)
fbeta = turicreate.evaluation.fbeta_score(targets, predictions, beta = 2.0)
print(f1, fbeta)
```

```
0.333333333333 0.416666666667
```

## Multi-class F-scores

F-scores juga dapat ditentukan dalam pengaturan multi-kelas. Di sini, metrik dapat "dirata-ratakan" di semua kelas dengan berbagai cara yang memungkinkan. Beberapa di antaranya adalah:

- **mikro**: Menghitung metrik secara global dengan menghitung total berapa kali setiap kelas diprediksi dengan benar dan diprediksi secara salah.
- **makro**: Menghitung metrik untuk setiap "kelas" secara independen, dan menemukan rata-rata tak berbobotnya. Ini tidak memperhitungkan ketidakseimbangan label.
- **Tidak ada**: Kembalikan nilai akurasi untuk setiap kelas. Sesuai dengan setiap kelas.

**Catatan:** Nilai presisi, perolehan, dan akurasi rata-rata mikro secara matematis setara.

## Receiver Operating Characteristic (ROC Curve)

Dalam statistik, karakteristik operasi penerima (ROC), atau kurva KOP, adalah plot grafis yang menggambarkan kinerja sistem pengklasifikasi biner karenaprediksinya **ambang** bervariasi. Kurva KOP memberikan detail bernuansa tentang perilaku pengklasifikasi. Kurva dibuat dengan memplot rasio **positif benar** (TPR) terhadap rasio **positif palsu** (FPR) di berbagai pengaturan ambang.
Nama yang terdengar eksotis ini berasal dari tahun 1950-an dari analisis sinyal radio, dan dipopulerkan oleh makalah tahun 1978 oleh Charles Metz yang berjudul "Prinsip Dasar Analisis ROC". Kurva KOP menunjukkan sensitivitas pengklasifikasi dengan memplot rasio positif benar ke rasio positif palsu. Dengan kata lain, ini menunjukkan kepada Anda berapa banyak klasifikasi positif benar yang dapat diperoleh saat Anda mengizinkan lebih banyak dan lebih banyak lagi positif palsu. Pengklasifikasi sempurna yang tidak membuat kesalahan akan mencapai rasio positif benar 100% dengan segera, tanpa menimbulkan positif palsu. Ini hampir tidak pernah terjadi dalam praktiknya.
Kurva KOP yang baik memiliki banyak ruang di bawahnya (karena rasio positif sebenarnya meningkat dengan sangat cepat hingga 100%). Kurva KOP yang buruk mencakup sangat sedikit area.

```
targets = turicreate.SArray([0, 1, 1, 0])
predictions = turicreate.SArray([0.1, 0.35, 0.7, 0.99])
```

```
roc_curve = turicreate.evaluation.roc_curve(targets, predictions)
```

```
Data:
+-----------+-----+-----+---+---+
| threshold | fpr | tpr | p | n |
+-----------+-----+-----+---+---+
|    0.0    | 1.0 | 1.0 | 2 | 2 |
|   1e-05   | 1.0 | 1.0 | 2 | 2 |
|   2e-05   | 1.0 | 1.0 | 2 | 2 |
|   3e-05   | 1.0 | 1.0 | 2 | 2 |
|   4e-05   | 1.0 | 1.0 | 2 | 2 |
|   5e-05   | 1.0 | 1.0 | 2 | 2 |
|   6e-05   | 1.0 | 1.0 | 2 | 2 |
|   7e-05   | 1.0 | 1.0 | 2 | 2 |
|   8e-05   | 1.0 | 1.0 | 2 | 2 |
|   9e-05   | 1.0 | 1.0 | 2 | 2 |
+-----------+-----+-----+---+---+
[100001 rows x 5 columns]
```

Hasil dari **kurva roc** adalah kelipatan- kolom **SFrame** dengan kolom-kolom berikut:

- **tpr**: True positive rate, jumlah true positive dibagi jumlah positif.
- **fpr**: Rasio positif palsu, jumlah positif palsu dibagi dengan jumlah negatif.
- **p**: Jumlah total nilai positif.
- **n**: Jumlah total nilai negatif.
- **classKelas**:referensi untuk kurva KOP ini (untuk klasifikasi kelas jamak).

**Catatan**: Kurva KOP dihitung menggunakan histogram binned dan karenanya selalu berisi 100 ribu baris. Histogram binned memberikan kurva yang akurat hingga desimal ke-5.

```
targets = turicreate.SArray(["cat", "dog", "cat", "dog"])
predictions = turicreate.SArray([0.1, 0.35, 0.7, 0.99])

roc_curve = turicreate.evaluation.roc_curve(targets, predictions)
```

```
Data:
+-----------+-----+-----+---+---+
| threshold | fpr | tpr | p | n |
+-----------+-----+-----+---+---+
|    0.0    | 1.0 | 1.0 | 2 | 2 |
|   1e-05   | 1.0 | 1.0 | 2 | 2 |
|   2e-05   | 1.0 | 1.0 | 2 | 2 |
|   3e-05   | 1.0 | 1.0 | 2 | 2 |
|   4e-05   | 1.0 | 1.0 | 2 | 2 |
|   5e-05   | 1.0 | 1.0 | 2 | 2 |
|   6e-05   | 1.0 | 1.0 | 2 | 2 |
|   7e-05   | 1.0 | 1.0 | 2 | 2 |
|   8e-05   | 1.0 | 1.0 | 2 | 2 |
|   9e-05   | 1.0 | 1.0 | 2 | 2 |
+-----------+-----+-----+---+---+
[100001 rows x 5 columns]
```

Untuk klasifikasi biner, saat label target adalah ketik **string**, kemudian label diurutkan secara alfanumerik dan label terbesar dipilih sebagai **kelas positif**. Kurva KOP juga dapat ditentukan dalam pengaturan kelas jamak dengan mengembalikan kurva tunggal untuk setiap kelas.

## Area Under the Curve (AUC)

AUC adalah singkatan dari Area Under the Curve. Di sini, kurva tersebut adalah kurva KOP. Seperti disebutkan di atas, kurva KOP yang baik memiliki banyak ruang di bawahnya (karena rasio positif benar naik hingga 100% dengan sangat cepat). Kurva KOP yang buruk mencakup sangat sedikit area.

```
targets = turicreate.SArray([0, 1, 1, 0])
predictions = turicreate.SArray([0.1, 0.35, 0.7, 0.99])

auc = turicreate.evaluation.auc(targets, predictions)
print(auc)
```

```
0.5
```

**Catatan**: Nilai AUC dihitung menggunakan binned histogram dan karenanya selalu berisi 100 ribu baris. Histogram binned memberikan kurva yang akurat hingga desimal ke-5.

Nilai AUC juga dapat ditentukan ketika kelas target berjenis **string**. Untuk klasifikasi biner, jika label target berjenis **string**, maka label diurutkan secara alfanumerik dan label terbesar dianggap sebagai label "positif".

```
targets = turicreate.SArray(["cat", "dog", "cat", "dog"])
predictions = turicreate.SArray([0.1, 0.35, 0.7, 0.99])

auc = turicreate.evaluation.auc(targets, predictions)
print(auc)
```

```
0.5
```

## Multi-class area under curve

Nilai AUC juga dapat ditentukan dalam pengaturan multi-kelas. Di sini, metrik dapat "dirata-ratakan" di semua kelas dengan berbagai cara yang memungkinkan. Beberapa di antaranya adalah:

- **makro**: Hitung metrik untuk setiap "kelas" secara independen, dan temukan rata-rata tak berbobotnya. Ini tidak memperhitungkan ketidakseimbangan label.
- **Tidak ada**: Menampilkan metrik yang sesuai untuk setiap kelas.

## Calibration curve

kalibrasi Kurva yaitu kalibrasi yang membandingkan probabilitas sebenarnya dari suatu peristiwa dengan probabilitas yang diprediksi. Bagan ini menunjukkan apakah model yang dibuat memiliki prediksi probabilitas yang "dikalibrasi" dengan baik. Semakin dekat kurva kalibrasi ke kurva "ideal", semakin yakin seseorang dapat menafsirkan prediksi probabilitas sebagai nilai keyakinan untuk memprediksi churn.

Grafik dibuat dengan membagi ruang probabilitas, dari 0,0 menjadi 1,0, menjadi sejumlah langkah diskrit yang tetap. Misalnya, untuk ukuran langkah 0,1, ruang probabilitas akan dibagi menjadi 10 kelompok: [0,0, 0,1), [0,1, 0,2) ... [0,9, 1,0]. Dengan adanya serangkaian peristiwa yang probabilitas yang diprediksi telah dihitung, setiap prediksi probabilitas dibulatkan ke keranjang yang sesuai. Agregasi kemudian dilakukan untuk menghitung jumlah total titik data dalam setiap rentang probabilitas bersama dengan pecahan dari titik data tersebut yang merupakan peristiwa yang benar-benar positif. Kurva kalibrasi memplot pecahan peristiwa yang benar-benar positif untuk setiap kelompok probabilitas. Model yang dikalibrasi sempurna, kurva akan menjadi garis dari (0,0, 0,0) hingga (1,0, 1,0).
Pertimbangkan contoh sederhana dari masalah klasifikasi biner di mana model dibuat untuk memprediksi peristiwa klik atau tanpa klik. Jika sebuah peristiwa "diklik", dan memiliki probabilitas prediksi 0,345, menggunakan binning 0,1, "jumlah peristiwa" untuk bin [0,3, 0,4) akan ditingkatkan satu, dan "jumlah peristiwa sebenarnya" akan ditingkatkan satu. Jika peristiwa dengan probabilitas 0,231 "tidak diklik", "jumlah peristiwa" untuk bin [0,2, 0,3) akan bertambah satu, sedangkan "jumlah peristiwa sebenarnya" tidak akan ditingkatkan. Untuk setiap bin, "jumlah benar" di atas "jumlah total peristiwa" menjadi probabilitas yang diamati (nilai sumbu y). Probabilitas prediksi rata-rata untuk bin menjadi nilai sumbu x.

## Kurva Precision-Recall Kurva

presisi-recall menggambarkan trade-off yang melekat antara presisi dan recall. Lebih mudah untuk memahami dan menafsirkan kurva ini dengan memahami setiap komponen dari kurva ini pada "ambang" yang berbeda.
Pertimbangkan contoh sederhana dari masalah klasifikasi biner di mana model dibuat untuk memprediksi peristiwa klik atau tanpa klik. Jika peristiwa "diklik", Melihat matriks kebingungan, prediksi model dapat dibagi menjadi empat kategori:

- Peristiwa yang diprediksi menjadi "klik" dan sebenarnya adalah peristiwa "klik" (True Positive, TP).
- Peristiwa yang diperkirakan sebagai peristiwa "bukan-klik" tetapi bukan peristiwa "klik" (False Positive, FP).
- Peristiwa yang diperkirakan sebagai peristiwa "bukan-klik" dan sebenarnya merupakan peristiwa "bukan-klik" (True Negative, TN).
- Peristiwa yang diperkirakan menjadi "klik" tetapi sebenarnya merupakan peristiwa "bukan-klik" (False Negative, FN).

Seperti yang didefinisikan sebelumnya, presisi adalah sebagian kecil dari peristiwa klik yang diprediksi yang sebenarnya merupakan peristiwa klik, sedangkan ingatan adalah sebagian kecil peristiwa klik yang diprediksi dengan benar oleh model. Definisi presisi dan perolehan tidak selalu merupakan peristiwa diskrit dan dapat bergantung pada probabilitas prediksi. Seseorang dapat menentukan "ambang" probabilitas yang di atasnya prediksi model dapat dianggap sebagai prediksi "klik".
Misalnya, "ambang" 0,5 menyiratkan bahwa semua prediksi probabilitas di atas 0,5 dianggap sebagai peristiwa "klik" dan semua prediksi di bawah dianggap sebagai peristiwa "bukan-klik". Dengan memvariasikan ambang antara 0,0 dan 1,0, kami dapat menghitung angka presisi dan perolehan yang berbeda. Pada ambang 0,0 semua prediksi model adalah "klik". Pada titik ini, model menunjukkan daya ingat yang sempurna tetapi memiliki kemungkinan presisi yang paling buruk. Pada ambang 1,0, model memprediksi setiap peristiwa sebagai "bukan-klik"; ini adalah presisi prefek (model tidak pernah salah dalam memprediksi klik!) tetapi kemungkinan ingatan terburuk.

Ini dengan memplot presisi dan perolehan di berbagai ambang, kurva presisi-recall menggambarkan trade-off dari presisi dan perolehan untuk model, pada berbagai ambang. Dua model dapat dibandingkan dengan mudah dengan memplot kurva presisi-recall. Kurva yang lebih ke sisi kanan atas plot mewakili model yang lebih baik.

Dalam tugas regresi, model belajar memprediksi skor numerik. Contohnya adalah memprediksi harga saham di masa depan berdasarkan riwayat harga masa lalu dan informasi lain tentang perusahaan dan pasar.
Bagian ini akan membahas dua cara untuk mengukur kinerja regresi:

- Root-Mean-Squared-Error
- Max-Error

## Root-Mean-Squared-Error (RMSE)

Metrik yang paling umum digunakan untuk tugas-tugas regresi adalah RMSE (Root Mean Square Error). Ini didefinisikan sebagai akar kuadrat dari jarak kuadrat rata-rata antara skor aktual dan skor prediksi:
Di sini, menunjukkan skor sebenarnya untuk titik data ke-i, dan
menunjukkan nilai prediksi. Salah satu cara intuitif untuk memahami rumus ini adalah jarak Euclidean antara vektor skor sebenarnya dan vektor skor prediksi, dirata-ratakan dengan, di mana adalah jumlah titik data.

```
import turicreate as tc

y    = tc.SArray([3.1, 2.4, 7.6, 1.9])
yhat = tc.SArray([4.1, 2.3, 7.4, 1.7])

print(tc.evaluation.rmse(y, yhat))
```

```
0.522015325446
```

## Max-Error

Meskipun RMSE adalah metrik yang paling umum, mungkin sulit untuk menafsirkannya. Salah satu alternatifnya adalah dengan melihat kuantitas distribusi dari persentase kesalahan absolut. Metrik Max-Error adalah **kasus terburuk** kesalahanantara nilai prediksi dan nilai sebenarnya.

```
import turicreate as tc

y    = tc.SArray([3.1, 2.4, 7.6, 1.9])
yhat = tc.SArray([4.1, 2.3, 7.4, 1.7])

print(tc.evaluation.max_error(y, yhat))
```

```
1.0
```

https://apple.github.io/turicreate/docs/userguide/evaluation/

TUGAS 5

**Nama : Vero Faloris**
**Nim : 202420032**
**Kelas : MTI 23 Reguler A**
**Mk : Advanced Database**

**LINK RUJUKAN :**

**https://youtu.be/0IovbG6E0Go**
**https://www.youtube.com/watch?v=F5NHk6c94-8**

tugas 5 : Pelajari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format ms word dan dilengkapi dengan gambar dan keterangan yang jelas untuk tiap langkah-langkahnya. Sertkan sumber sumber rujukan yang anda gunakan. Kumpulkan sebelum waktu batas akhir pengumpulan tugas.

SAMPEL DATA YANG AKAN DIKELOMPOKKAN

| NO | DAERAH | LUAS LAHAN | PRODUKSI |
|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 |

Langkah Langkah Perhitungan
1. Menentukan Jumlah Cluster data
2. Tentukan titik Pusat Cluster
3. menentukan jarak obyek dengan centroid
4. pengelompokan obyek
5. jika kelompok data hasil perhitungan baru sama dengan hasil perhitungan kelompok data baru maka perhitungan selesai

Langkah Penyelesaian

**1. Tentukan cluster atau kategori pusat dengan asumsi (Max, Mid, Min) namun dalam konsepnya dipilih secara acak**

| | | |
|---|---|---|
| Cluster I | 162 | 964 |
| Cluster II | 72 | 387 |
| Cluster III | 31 | 182 |

**2. Penentuan Jarak Terdekat**

**CLUSTER = SQRT(((C8-$B$3)^2)+(D8-$C$3)^2)**

| NO | DAERAH | LUAS LAHAN | PRODUKSI | C1 | C2 | C3 | JARAK TERPENDEK |
|---|---|---|---|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 570 | 16 | 223 | 16 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 793 | 209 | 0 | 0 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 714 | 130 | 79 | 79 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 684 | 100 | 109 | 100 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 692 | 108 | 101 | 101 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 509 | 77 | 284 | 77 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 584 | 0 | 209 | 0 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 18 | 566 | 775 | 18 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 263 | 322 | 530 | 263 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 643 | 59 | 150 | 59 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 684 | 100 | 109 | 100 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 661 | 77 | 132 | 77 |

3. Pengelompokan Objek

| NO | DAERAH | LUAS LAHAN | PRODUKSI | KELOMPOK |
|---|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 2 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 3 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 3 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 2 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 3 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 2 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 2 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 1 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 1 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 2 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 2 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 2 |

DENGAN KESIMPULAN DATA 1 VERO

cluster 1 : sebanyak 2 data

cluster 2 : sebanyak 7 adata

cluster 3 : sebanyak 3 data

| | | | | | cluster baru | | |
|---|---|---|---|---|---|---|---|
| NO | DAERAH | LUAS LAHAN | PRODUKSI | KELOMPOK | C1 | C2 | C3 |
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 2 | 137,5 | 60,14286 | 43,66666667 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 3 | 826 | 353,1429 | 240,6666667 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 3 | | | |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 2 | | | |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 3 | | | |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 2 | | | |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 2 | | | |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 1 | | | |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 1 | | | |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 2 | | | |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 2 | | | |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 2 | | | |

penentuan cluster baru

| cluster | luas lahan | produksi |
|---|---|---|
| I | 137,5 | 826 |
| II | 60,14286 | 353,1429 |
| III | 43,66667 | 240,6667 |

| NO | DAERAH | LUAS LAHAN | PRODUKSI | C1 | C2 | C3 | JARAK TERPENDEK |
|---|---|---|---|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 430 | 49 | 163 | 49 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 653 | 174 | 60 | 60 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 574 | 95 | 19 | 19 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 544 | 65 | 49 | 49 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 552 | 73 | 41 | 41 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 369 | 111 | 224 | 111 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 444 | 36 | 149 | 36 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 122 | 602 | 715 | 122 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 122 | 357 | 470 | 122 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 503 | 24 | 90 | 24 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 543 | 64 | 50 | 50 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 521 | 42 | 72 | 42 |

| NO | DAERAH | LUAS LAHAN | PRODUKSI | KELOMPOK |
|---|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 2 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 3 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 3 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 3 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 3 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 2 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 2 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 1 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 1 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 2 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 3 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 2 |

DENGAN KESIMPULAN DATA 2 VERO
cluster 1 : sebanyak 2 data
cluster 2 : sebanyak 5 adata
cluster 3 : sebanyak 5 data

| | | | | | cluster baru | | |
|---|---|---|---|---|---|---|---|
| NO | DAERAH | LUAS LAHAN | PRODUKSI | KELOMPOK | C1 | C2 | C3 |
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 2 | 137,5 | 64,6 | 45,8 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 3 | 826 | 378,6 | 260,2 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 3 | | | |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 3 | | | |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 3 | | | |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 2 | | | |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 2 | | | |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 1 | | | |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 1 | | | |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 2 | | | |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 3 | | | |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 2 | | | |

penentuan cluster baru

| cluster | luas lahan | produksi |
|---|---|---|
| I | 137,5 | 826 |
| II | 64,6 | 378,6 |
| III | 45,8 | 260,2 |

| NO | DAERAH | LUAS LAHAN | PRODUKSI | C1 | C2 | C3 | JARAK TERPENDEK |
|---|---|---|---|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 4688 | 25 | 550 | 25 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 10698 | 932 | 141 | 141 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 7265 | 124 | 9 | 9 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 7119 | 124 | 46 | 46 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 6937 | 87 | 48 | 48 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 4894 | 86 | 572 | 86 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 3851 | 63 | 813 | 63 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 720 | 10054 | 14188 | 720 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 480 | 2670 | 4962 | 480 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 5355 | 37 | 300 | 37 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 7474 | 187 | 35 | 35 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 5647 | 36 | 225 | 36 |

| NO | DAERAH | LUAS LAHAN | PRODUKSI | KELOMPOK |
|---|---|---|---|---|
| 1 | Lubuklinggau Barat 1 | 66 | 402 | 2 |
| 2 | Lubuklinggau Barat 2 | 31 | 182 | 3 |
| 3 | Lubuklinggau Barat 3 | 49 | 259 | 3 |
| 4 | Lubuklinggau Timur 1 | 50 | 289 | 3 |
| 5 | Lubuklinggau Timur 2 | 51 | 281 | 3 |
| 6 | Lubuklinggau Timur 3 | 65 | 464 | 2 |
| 7 | Lubuklinggau selatan 1 | 72 | 387 | 2 |
| 8 | Lubuklinggau selatan 2 | 162 | 946 | 1 |
| 9 | Lubuklinggau selatan 3 | 113 | 706 | 1 |
| 10 | Lubuklinggau Utara 1 | 61 | 329 | 2 |
| 11 | Lubuklinggau Utara 2 | 48 | 290 | 3 |
| 12 | Lubuklinggau Utara 3 | 59 | 311 | 2 |

DENGAN KESIMPULAN DATA 3 VERO

cluster 1 : sebanyak 2 data

cluster 2 : sebanyak 5 adata

cluster 3 : sebanyak 5 data

karena data 2 vero dan data 3 vero sama maka perhitungan selesai

**NAMA : WIDIA ASTUTI**

**NIM : 202420021**

**MATA KULIAH : ADVANCED DATABASE**

**Tugas 5**

Pelajari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format ms word dan dilengkapi dengan gambar dan keterangan yang jelas untuk tiap langkah-langkahnya. Sertakan sumber sumber rujukan yang anda gunakan. Kumpulkan sebelum waktu batas akhir pengumpulan tugas.

**Jawab :**

**Clustering** merupakan teknik mengelompokan data secara otomatis tanpa perlu diketahui label kelasnya, tetapi dapat juga digunakan untuk mengelompokan himpunan data dengan kelas label yang diketahui

Metode K Means clustering digunakan dalam data masing-masing untuk mengelompokkan data –data ke dalam cluster /beberapa kelompok berdasarkan suatu kemirirpan variable atau atribut data. Metode ini akan mempresentasikan setiap kelompok dengan suatu pusat kelompok atau centroid, serta menghitung jarak setiap data ke setiap centroid terdekat. Sehingga centroid ini diangap sebagai nilai pusat baru untuk masing-masing cluster.

Tahapan Cluster :

1. Menentukan jumlah cluster data atau nilai K
2. Pilih K data dari data set x sebagai centroid
3. Menghitung jarak objek ke centroid dengan cara mengalokasikan semua data ke centroid terdekat
4. Hitunglah kembali centroid C berdasarkan data yang mengikuti cluster masing-masing.
5. Ulangi langkah 3 dan 4 hingga kondisi konvergen tercapai
6. Jika kelompok data hasil perhitungan baru sama dengan hasil perhitungan kelompok data baru maka perhitungan selesai.

**Tutorial Contoh soal :**

Diberikan data sample berupa 6 data, dimana data tsb memliki dua dimensi (x dan y) agar mudah divisualisasikan dalam koordinat kartesius. Dibawah ini merupakan tabel data sample

| Data ke-i | x | y |
|---|---|---|
| 1 | 3 | 2 |
| 2 | 2 | 3 |
| 3 | 3 | 5 |
| 4 | 5 | 7 |
| 5 | 2 | 5 |
| 6 | 4 | 2 |

Langkah-langkah tahapan :

1. Berdasarkan data diatas, proses dilakukan dengan mengelompokan menjadi 2 cluster (k=2)
2. Metentukan titik centroid sebanyak k berdasarkan titik-titik tertentu data set, dapat dipilih secara random ataupun ditentukan langsung. Maka kita pilih data set ke 2 dan 4 sebagai centroid awal.

| Centroid | x | y |
|---|---|---|
| A | 1 | 2 |
| B | 5 | 6 |

3. Pengukuran jarak pada setiap data terhadap titik centroid menggunakan perhitungan jarak euclideon

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

Maka perhitungannya :

- Jarak titik 1 ke centroid A

$d_{(x,y)} = \sqrt{(1-3)^2 + (2-2)^2}$

$= \sqrt{4} = 2$

- jarak titik 1 ke centroid B

$d_{(x,y)} = \sqrt{(5-3)^2 + (6-2)^2}$

$= \sqrt{20} = 4{,}4721$

Dan dihitung seterusnya hingga data paling akhir (data ke 6), sehingga diperoleh rekap hasil perhitungan jarak terdekat ke setiap cluster sebagai berikut :

| Data ke-i | x | y | C1 | C2 |
|---|---|---|---|---|
| 1 | 3 | 2 | 2 | 4,472136 |
| 2 | 2 | 3 | 1,414214 | 4,242641 |
| 3 | 3 | 5 | 3,605551 | 2,236068 |
| 4 | 5 | 7 | 6,403124 | 1 |
| 5 | 2 | 5 | 3,605551 | 3,162278 |
| 6 | 4 | 2 | 3 | 4,123106 |

4. Menghitung jarak objek ke centroid dengan cara mengalokasikan semua data ke centroid terdekat

| Data ke-i | x | y | C1 | C2 | Cluster |
|---|---|---|---|---|---|
| 1 | 3 | 2 | 2 | 4,472136 | C1 |
| 2 | 2 | 3 | 1,414214 | 4,242641 | C1 |
| 3 | 3 | 5 | 3,605551 | 2,236068 | C2 |
| 4 | 5 | 7 | 6,403124 | 1 | C2 |
| 5 | 2 | 5 | 3,605551 | 3,162278 | C2 |
| 6 | 4 | 2 | 3 | 4,123106 | C1 |

Setelah cluster(kelompok) untuk setiap record ditentukan seperti pada tabel di atas, tugas kita sekarang adalah memperbaharui nilai titik pusat cluster, yang mana sebelumnya titik pusat cluster kita adalah M1{1,1} dan M2{2,1}. Untuk meng-update nilai titik pusat cluster, kita dapat menggunakan persamaan cluster center sebagai berikut :

$$ClusterCenter = \sum \frac{a_i}{n}$$

**Cluster 1**

Cluster center : $M_{1\,(x)} = \frac{X1+X2+X6}{3} = \frac{3+2+4}{3} = 3$

$M_{1(y)} = \frac{y1+y2+y6}{3} = \frac{2+3+2}{3} = 2{,}3333$

**Cluster 2**

Cluster center : $M_{1\,(x)} = \frac{X3+X4+X5}{3} = \frac{3+5+2}{3} = 3{,}3333$

$M_{1(y)} = \frac{y3+y4+y5}{3} = \frac{3+5+7}{3} = 5{,}6667$

Sehingga didapat update pusat cluster terbaru :

pusat cluster terbaru

| | x | y |
|---|---|---|
| M1 | 3 | 2,3333 |
| M2 | 3,3333 | 5,6667 |

5. Selanjutnya kita akan mengulangi langkah ke-3 dan ke-4, guna menentukan kembali kelompok tiap data. Untuk menguji apakah terjadi perpindahan kelompok pada data tersebut. setelah di hitung kembali didapatlah hasil nilai terdekat tiap kelompok sebagai berikut.

| Data ke-i | x | y | C1 | C2 | Cluster |
|---|---|---|---|---|---|
| 1 | 3 | 2 | 0,3333 | 3,681817 | C1 |
| 2 | 2 | 3 | 1,201869 | 2,981439 | C1 |
| 3 | 3 | 5 | 2,6667 | 0,745371 | C2 |
| 4 | 5 | 7 | 5,077213 | 2,13438 | C2 |
| 5 | 2 | 5 | 3,33336 | 1,490697 | C2 |
| 6 | 4 | 2 | 1,054082 | 3,726819 | C1 |

6. Data hasil perhitungan baru sama dengan hasil perhitungan kelompok data baru maka perhitungan selesai atau kondisi konvergen tercapai.

Sumber :

https://www.youtube.com/watch?v=U6hRPOm82vg

https://www.youtube.com/watch?v=5o_JQ6AHA0Y&t=365s

# TUGAS 5
# ADVANCED DATABASE


Di Susun Oleh :
AAN NOVRIANTO
NIM : 202420010

Dosen Pengasuh :
Tri Basuki Kurniawan , S.Kom., M.Eng. Ph.D

**Program Pasca Sarjana**

**Universitas Bina Darma Palembang**

2020/2021

Clustering Pada Data Pengemudi Go-Ride

Sebagai contoh, saya akan menunjukkan cara kerja algoritma K-means dengan dataset sampel dari data pengemudi (driver) pada aplikasi Go-Ride. Disini saya hanya akan menggunakan dua fitur driver pada aplikasi Go-Ride untuk dijadikan variabel : jarak rata-rata per hari dan kecepatan dalam mengemudi per-hari . .

## Langkah 1 : Import Libraries

Pada kasus ini , menggunakan library Scikit-learn dan beberapa data acak untuk mengilustrasikan penjelasan sederhana pengelompokan K-means.

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as pltfrom sklearn.cluster

import KMeans
from sklearn.preprocessing import MinMaxScaler
```

Keterangan

>> Panda untuk membaca dan menulis spreadsheet

>> Numpy untuk melakukan perhitungan yang efisien

>>Matplotlib untuk visualisasi data

## Langkah 2 : Menginput Data

Data yang digunakan yakni dataset sampel data pengemudi pada aplikasi Go-Ride.

Dibawah ini adalah script untuk :

>>menginput data pada (baris 1)

>>membaca data pada (baris 2)

```
--- Membaca Data ---

driver=pd.read_csv("go_ride_ride.csv")
driver.head()
```

| | id | id_android | speed | time | distance | rating | rating_bus | rating_weather | car_or_bus | linha |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 19.210586 | 0.138049 | 2.652 | 3 | 0 | 0 | 1 | NaN |
| **1** | 2 | 0 | 30.848229 | 0.171485 | 5.290 | 3 | 0 | 0 | 1 | NaN |
| **2** | 3 | 1 | 13.560101 | 0.067699 | 0.918 | 3 | 0 | 0 | 2 | NaN |
| **3** | 4 | 1 | 19.766679 | 0.389544 | 7.700 | 3 | 0 | 0 | 2 | NaN |
| **4** | 8 | 0 | 25.807401 | 0.154801 | 3.995 | 2 | 0 | 0 | 1 | NaN |

Dari data yang di miliki terdapat 10 variabel pada data set yang ada . Ada beberapa variabel yang tidak dibutuhkan sehingga harus dihapuskan

## Langkah 3 : Menghilangkan kolom yang tidak diperlukan

Untuk contoh ini, saya telah menghilangkan beberapa kolom yang tidak diperlukan . Sehingga hanya menyisahkan 4 kolom yakni id , **id_android,speed dan distance** seperti ditunjukkan pada gambar dibawah :

```
--- Menghilangkan Kolom Yang Tidak Perlu ---

driver=driver.drop(["linha","car_or_bike","rating_weather",
"rating_bike","rating","time"],axis=1)
driver.head()
```

| | id | id_android | speed | distance |
|---|---|---|---|---|
| **0** | 1 | 0 | 19.210586 | 2.652 |
| **1** | 2 | 0 | 30.848229 | 5.290 |
| **2** | 3 | 1 | 13.560101 | 0.918 |
| **3** | 4 | 1 | 19.766679 | 7.700 |
| **4** | 8 | 0 | 25.807401 | 3.995 |

Pada gambar diatas menunjukkan dataset untuk 2652 driver . Selanjutnya menentukan variabel yang diklusterkan , disini menggunakan variabel jarak pada sumbu X dan variabel kecepatan pada sumbu Y .

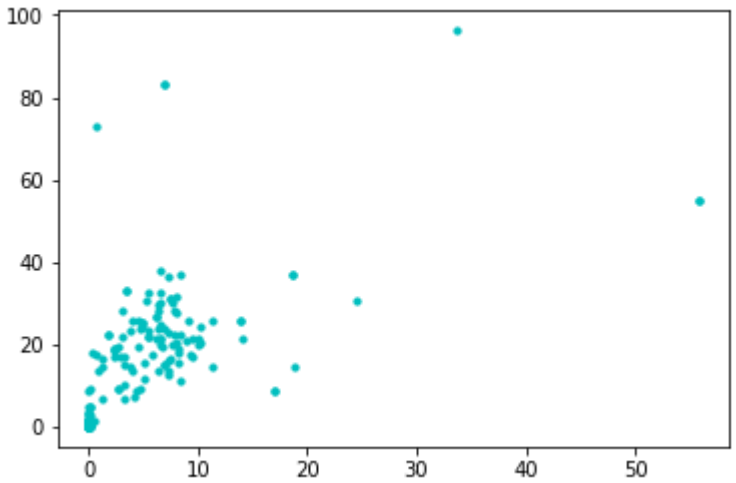
```
-- Menentukan variabel yang akan di klusterkan ---

driver_x=driver.iloc[:,1:3]
driver_x.head()
```

| | speed | distance |
|---|---|---|
| 0 | 19.210586 | 2.652 |
| 1 | 30.848229 | 5.290 |
| 2 | 13.560101 | 0.918 |
| 3 | 19.766679 | 7.700 |
| 4 | 25.807401 | 3.995 |

Dataset diatas divisualisasikan persebaran datanya sebagai berikut :

```
--- Memvisualkan persebaran data ---
plt.scatter(driver.distance, driver.speed, s =10, c = "c", marker = "o", alpha = 1)
plt.show()
```



## Langkah 4 : Menentukan nilai K

Sebelum menetukan nilai K , saya harus mengubah variabel data frame menjadi array terlebih dahulu :

```
--- Mengubah Variabel Data Frame Menjadi Array ---
x_array=np.array(driver_x)
print(x_array)
```

```
[[1.92105856e+01 2.65200000e+00]
 [3.08482291e+01 5.29000000e+00]
 [1.35601009e+01 9.18000000e-01]
 [1.97666790e+01 7.70000000e+00]
 [2.58074009e+01 3.99500000e+00]
 [1.34691332e+00 9.00000000e-03]
 [3.68507874e+01 8.40200000e+00]
 [1.74051313e+01 6.75000000e-01]
 [1.53954361e+01 8.11100000e+00]
 [8.90272944e+00 2.70000000e-02]
 [1.50413480e+01 3.27700000e+00]
 [1.44400981e+01 3.87200000e+00]
 [1.63567325e+01 1.26000000e+00]
 [1.75427999e+01 5.85700000e+00]
 [9.45181557e+00 2.61600000e+00]
 [9.45181557e+00 2.61600000e+00]
 [1.62635039e+01 7.33400000e+00]
 [2.12235944e+01 6.14900000e+00]
 [1.94236545e+01 4.59500000e+00]
```

Kemudian saya harus menstandarkan kembali ukuran variabel . Agar data dapat kembali seperti jenis data diawal sebelum di array kan

```
--- Menstandarkan Ukuran Variabel ---
scaler = MinMaxScaler()
x_scaled = scaler.fit_transform(x_array)
x_scaled
```

```
array([[1.99600362e-01, 4.75353691e-02],
       [3.20578504e-01, 9.48376338e-02],
       [1.40861225e-01, 1.64428267e-02],
       [2.05381184e-01, 1.38051606e-01],
       [2.68176999e-01, 7.16168481e-02],
       [1.39000630e-02, 1.43448869e-04],
       [3.82977592e-01, 1.50639244e-01],
       [1.80831914e-01, 1.20855673e-02],
       [1.59940297e-01, 1.45421291e-01],
       [9.24459124e-02, 4.66208826e-04],
       [1.56259404e-01, 5.87423120e-02],
       [1.50009162e-01, 6.94113217e-02],
       [1.69933373e-01, 2.25752658e-02],
       [1.82263037e-01, 1.05004572e-01],
       [9.81538910e-02, 4.68898492e-02],
       [9.81538910e-02, 4.68898492e-02],
       [1.68964223e-01, 1.31488820e-01],
       [2.20526426e-01, 1.10240456e-01],
       [2.01815302e-01, 8.23755133e-02],
```

Dalam hal tersebut, nilai K (n_clusters) atau nilai arbitrer bebas ditentukan. Untuk ini , saya akan mengkonfigurasi dan menentukan nilai K sebesar 3 cluster. Selain itu saya juga menetukan kluster dari data yang telah di standarkan .

```
--- Menentukan dan mengkonfigurasi fungsi kmeans ---
kmeans = KMeans(n_clusters = 3, random_state=123)---

Menentukan kluster dari data ---
kmeans.fit(x_scaled)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=2, n_init=10, n_jobs=1, precompute_distances='auto',
    random_state=123, tol=0.0001, verbose=0)
```

## Langkah 5 : Memvisualisasikan Kluster

Untuk memvisualisasikan kluster, harus menampilkan centroid terlebih dahulu seperti gambar dibawah ini :

```
--- Menampilkan pusat cluster ---
print(kmeans.cluster_centers_)
```

```
[[0.23246003 0.12863014]
 [0.02863662 0.00851813]
 [0.77224472 0.47782221]]
```

Selanjutnya menampilkan hasil cluster dan Menambahkan kolom data frame driver .

```
--- Menampilkan Hasil Kluster ---
print(kmeans.labels_)
-- Menambahkan Kolom "kluster" Dalam Data Frame Driver --
driver["kluster"] = kmeans.labels_
```

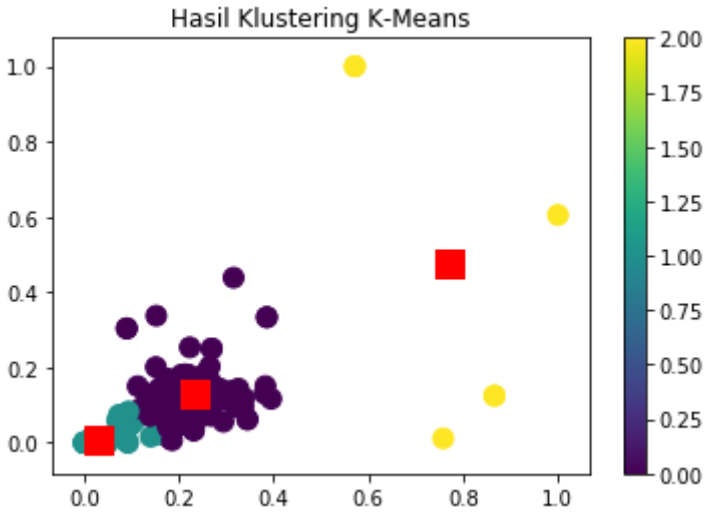
```
[0 0 1 0 0 1 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 1 2 2 0 1 1 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 2 1 2 2 2 1 1 1 1 1 1 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 1 1
 0 0 0 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0
 0 1 0 1 1 1 1 1 1 0 1 0 1 1 0 0 1 0 0 0 1 0 0 0 0 0 1 1 0 0 0 1 1 1 1 0 0
 1 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1]
```

Kemudian yang terakhir memvisualisasikan hasil cluster seperti gambar dibawah ini :

```
--- Memvisualkan hasil kluster ---
output = plt.scatter(x_scaled[:,0], x_scaled[:,1], s = 100, c = driver.kluster,
marker = "o", alpha = 1, )

centers = kmeans.cluster_centers_
plt.scatter(centers[:,0], centers[:,1], c='red', s=200, alpha=1 , marker="s");

plt.title("Hasil Klustering K-Means")
plt.colorbar (output)
plt.show()
```




Dari gambar diatas dapat kita lihat bahwa hasil dari data pengemudi Go-Track telah ter cluster menjadi 3.

NAMA : AHMAD ALI MA'MUN
NIM : 202420037

PENGELOMPOKKAN PERFORMA AKADEMIK MAHASISWA BERDASARKAN INDEKS PRESTASI MENGGUNAKAN K-MEANS CLUSTERING

Penelitian ini bertujuan untuk mengembangkan suatu perangkat lunak yang dapat mengelompokkan mahasiswa berdasarkan performa akademiknya. Keberhasilan

1. Studi literatur mengenai K-Means *Clustering*,
2. Menganalisis penelitian yang telah ada sebelumnya,
3. Mengumpulkan data untuk penelitian dengan cara mengumpulkan IP yang didapatkan dari bagian administrasi fakultas,
4. Mengubah format data yang telah dikumpulkan sehingga dapat memenuhi kebutuhanpenelitian,
5. Melakukan pengembangan perangkat lunak dengan metode Rational Unified Process (RUP),
6. Melakukan analisis dan pembahasan hasil eksperimen masukan serta penggunaannya pada perangkat lunak,
7. Membuat kesimpulan terhadap hasil penelitian.

### *3.2 Pengelompokkan Performa Akademik Mahasiswa dengan K-Means*

Metode K-Means *Clustering* memiliki enam tahapan. Pada tahap pertama yaitu menentukan jumlah *cluster* (k) awal untuk menjadi acuan proses *clustering*. Banyaknya *cluster* (k) harus lebih kecil dari pada banyaknya data (k < n). Pada penelitian ini jumlah *cluster* adalah sebanyak 4 *cluster* (k = 4).

Data pada Tabel 1 adalah contoh data yang digunakan.Data yang digunakan pada penelitian ini adalah data performa akademik mahasiswa Fakultas Ilmu Komputer Universitas Sriwijaya.Data yang diambil berupa Indeks Prestasi mahasiswa dari semester 1 sampai dengan 4.

Tabel 1. Contoh Data Akademik Mahasiswa

| No | Jurusan | NIM | Angkatan | IP1 | IP2 | IP3 | IP4 |
|---|---|---|---|---|---|---|---|
| 1 | SK | 09091001004 | 2009 | 3,21 | 3,5 | 2,96 | 3,32 |
| 2 | SK | 09091001005 | 2009 | 3,26 | 3,67 | 3,21 | 3,17 |
| 3 | SK | 09091001006 | 2009 | 3,11 | 3,62 | 3,04 | 3,05 |
| 4 | IF(BILINGUAL) | 09101402003 | 2010 | 3,63 | 3,42 | 3,58 | 3,48 |
| 5 | IF | 09101002068 | 2010 | 3,84 | 3,7 | 3,35 | 3,26 |
| 6 | SI | 09091003034 | 2009 | 3,47 | 3,71 | 3,62 | 2,89 |
| 7 | SI | 09111003036 | 2011 | 3,29 | 3,43 | 3,33 | 3,46 |
| 8 | SI (BILINGUAL) | 09111403015 | 2011 | 3,52 | 3,17 | 3,08 | 2,88 |
| 9 | IF | 09101002076 | 2010 | 3,11 | 3,05 | 3,2 | 2,37 |
| 10 | SI (BILINGUAL) | 09111403040 | 2011 | 2,95 | 3,1 | 2,48 | 2,28 |

pengelompokan data mahasiswa dihitung berdasarkan nilai standard eror dari proses pengelompokkan yang dilakukan secara berulang.

Langkah-langkah yang dilakukan pada penelitian ini adalah:

Berdasarkan data tersebut, *centroid* awal diambil secara acak di tiap *cluster* IP. Pembangkitan nilai acak pada *centroid* awal menggunakan metode pseudo random.Contoh nilai *centroid* acak yang diambil dapat dilihat pada Tabel 2.

Tabel 2.*Centroid* Awal pada pengulangan ke - 0

| C | a | b | c | d |
|---|---|---|---|---|
| c1 | 3,04 | 3,11 | 3,27 | 2,43 |
| c2 | 3,07 | 3,15 | 3,09 | 2,39 |
| c3 | 3,17 | 3,17 | 2,68 | 2,83 |
| c4 | 3,1 | 3,12 | 3,6 | 2,87 |

Keterangan:
c1 :*cluster* 1 a : nilai ip1 *cluster*
c2 :*cluster* 2 b : nilai ip2 *cluster*
c3 :*cluster* 3 c : nilai ip3 *cluster*
c4 :*cluster* 4 d : nilai ip4 *cluster*

Untuk pengulangan berikutnya (pengulangan ke–1 sampai selesai), *centroid* baru ditentukan dengan menghitung rata – rata data pada sertiap *cluster*. Jika ditemukan nilai *centroid* baru yang berbeda dengan *centroid* sebelumnya, maka proses dilanjutkan ke langkah berikutnya. Namun, jika nilai *centroid* yang baru dihitung sama dengan *centroid* sebelumnya maka proses *clustering* berhenti.

Proses selanjutnya adalah menghitung jarak masing – masing data terhadap *centroid* dengan menggunakan persamaan (1), jarak data 1 dengan tiap *cluster* adalah :

$$d(x_1, c_1) = \sqrt{(IP_1 - C1_a)^2 + (IP_2 - C1_b)^2 + (IP_3 - C1_c)^2 + (IP_4 - C1_d)^2}$$
$$= \sqrt{(3{,}21 - 3{,}04)^2 + (3{,}5 - 3{,}11)^2 + (2{,}96 - 3{,}27)^2 + (3{,}32 - 2{,}43)^2}$$
$$= 1{,}0340$$

$$d(x_1, c_2) = \sqrt{(IP_1 - C2_a)^2 + (IP_2 - C2_b)^2 + (IP_3 - C2_c)^2 + (IP_4 - C2_d)^2}$$
$$= \sqrt{(3{,}21 - 3{,}07)^2 + (3{,}5 - 3{,}15)^2 + (2{,}96 - 3{,}09)^2 + (3{,}32 - 2{,}39)^2}$$
$$= 1{,}0118$$

$$d(x_1, c_3) = \sqrt{(IP_1 - C3_a)^2 + (IP_2 - C3_b)^2 + (IP_3 - C3_c)^2 + (IP_4 - C3_d)^2}$$
$$= \sqrt{(3{,}21 - 3{,}17)^2 + (3{,}5 - 3{,}17)^2 + (2{,}96 - 2{,}68)^2 + (3{,}32 - 2{,}83)^2}$$
$$= 0{,}6549$$

$$d(x_1, c_4) = \sqrt{(IP_1 - C4_a)^2 + (IP_2 - C4_b)^2 + (IP_3 - C4_c)^2 + (IP_4 - C4_d)^2}$$
$$= \sqrt{(3{,}21 - 3{,}1)^2 + (3{,}5 - 3{,}12)^2 + (2{,}96 - 3{,}6)^2 + (3{,}32 - 2{,}87)^2}$$
$$= 0{,}8766$$

Berdasarkan perhitungan di atas data 1 masuk ke dalam *cluster* 3, karena data 1 memiliki jarak terpendek terhadap *centroid* pada *cluster* 3. Untuk hasil perhitungan selanjutnya dapat dilihat pada Tabel 3.

Tabel 3 Hasil Perhitungan Jarak dan Pengelompokkan Data

| Data | dc1 | dc2 | dc3 | dc4 |
|---|---|---|---|---|
| 1 | 1,034021 | 1,011879 | 0,654981 | 0,876698 |
| 2 | 0,955615 | 0,964002 | 0,809074 | 0,755116 |
| 3 | 0,838033 | 0,812773 | 0,619758 | 0,772075 |
| 4 | 1,281718 | 1,347108 | 1,227436 | 0,862206 |
| 5 | 1,297459 | 1,31145 | 1,167733 | 1,048141 |
| 6 | 0,93755 | 1,002247 | 1,12641 | 0,696994 |
| 7 | 1,108783 | 1,152953 | 0,949421 | 0,743774 |
| 8 | 0,687459 | 0,665658 | 0,533854 | 0,670373 |
| 9 | 0,130384 | 0,155242 | 0,707107 | 0,644205 |
| 10 | 0,809197 | 0,633325 | 0,629126 | 1,274912 |

Pada Tabel 3 kolom data yang diwarnai menunjukan bahwa data tersebut memiliki jarak terpendek terhadap masing-masing *cluster*, sehingga data tersebut dikelompokan sesuai warna *cluster*-nya. Pada hasil perhitungan di atas diperoleh 1 data untuk *cluster* 1, 4 data untuk *cluster* 3, 5 data untuk *cluster* 4 dan tidak ada data yang masuk ke *cluster* 2.

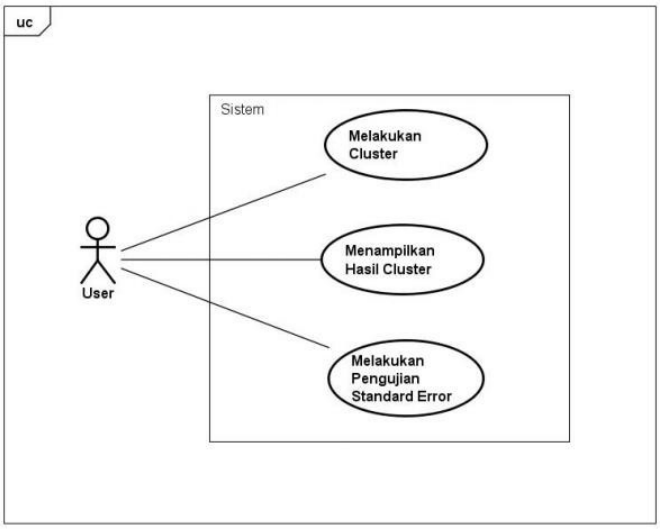
Setelah beberapa kali perulangan, proses pengelompokkan data menunjukkan data terletak pada *cluster* yang sama seperti di perulangan sebelumnya. Jika *cluster* data tetap, maka nilai *centroid* juga tetap maka perulangan berhenti. Pada perulangan ke empat diperoleh hasil akhir yaitu 1 data 6 untuk *cluster* 1, 1 data untuk *cluster*
2, 4 data untuk *cluster* 3 dan 4 data untuk *cluster* 4. Hasil akhir pada perulangan ke empat dapat dilihat pada Tabel 4.

Tabel 4. Hasil Akhir Pengelompokkan

| No | dc1 | dc2 | dc3 | dc4 |
|---|---|---|---|---|
| 1 | 1,082867 | 1,240806 | 0,253377 | 0,6179 |
| 2 | 1,02323 | 1,321363 | 0,238747 | 0,416893 |
| 3 | 0,90161 | 1,096586 | 0,214243 | 0,656125 |
| 4 | 1,33559 | 1,792986 | 0,734166 | 0,286705 |
| 5 | 1,330413 | 1,69393 | 0,68775 | 0,343802 |
| 6 | 1,005982 | 1,521249 | 0,659545 | 0,442493 |
| 7 | 1,1755 | 1,52951 | 0,448553 | 0,374433 |
| 8 | 0,676018 | 1,024597 | 0,461952 | 0,676166 |
| 9 | 0 | 0,744715 | 0,876926 | 1,156114 |
| 10 | 0,744715 | 0 | 1,129159 | 1,591163 |

*Perangkat Lunak Pengelompokkan Performa Akademik Mahasiswa*

Berdasarkan tujuan dari penelitian untuk membangun perangkat lunak, telah dilaksanakan proses analisis kebutuhan yang menghasilkan diagram use case yang dapat dilihat pada gambar 1.

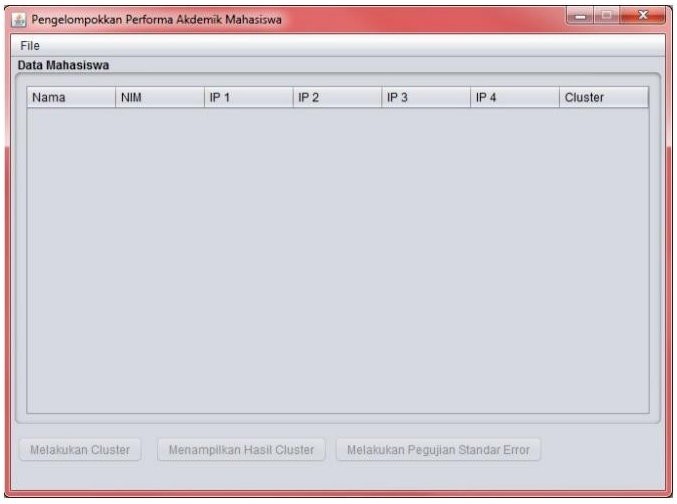



Gambar 1. Use Case Perangkat Lunak Pengelompokkan Performa Akademik Mahasiswa

Terdapat tiga use case utama, yaitu melakukan *cluster*, menampilkan hasil *cluster*, dan melakukan pengujian standardd error. Use case melakukan *cluster* adalah proses yang digunakan saat pengguna akan mengelompokkan data akademik mahasiswa. Data dengan format TXT diunggah pada perangkat lunak, dan kemudian dikelompokkan dengan menggunakan algoritma K-Means Clustering.

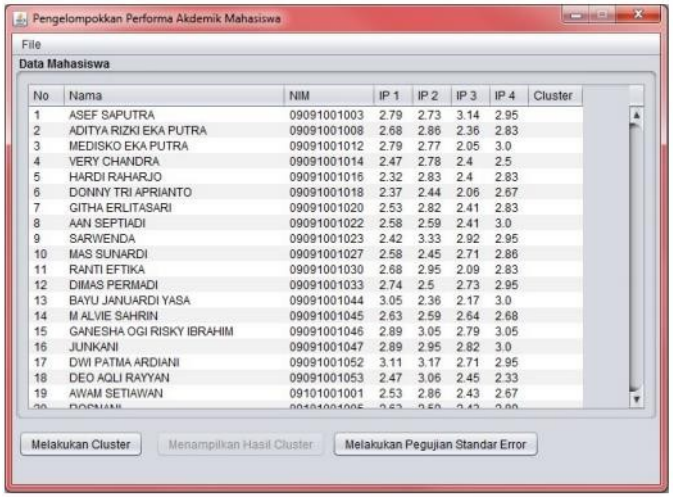
Selanjutnya, use case kedua, menampilkan hasil *cluster*, digunakan untuk menampilkan hasil pengelompokkan dalam bentuk grafis.Hal ini dilakukan untuk mempermudah pengguna dalam melihat hasil pengelompokkan. Pengguna diharapkan dapat mengambil keputusan dengan lebih mudah jika dapat melihat data secara grafik dan secara detil pada saat yang bersamaan. Use case terakhir adalah Melakukan Pengujian Standardd Error. Use case ini digunakan untuk mencari standard deviasi dan standardd error dari proses pengelompokkan. Jika fungsi ini digunakan oleh pengguna, akan dilakukan proses pengelompokkan sebanyak n kali yang kemudian dihitung nilai standardd errornya.

Berdasarkan use case yang telah diidentifikasi, telah dibangun perangkat lunak Pengelompokkan Performa Akademik Mahasiswa. Saat perangkat lunak dibuka, maka pengguna akan mendapatkan tampilan berupa antar muka awal dari perangkat lunak. Antarmuka ini dapat dilihat pada gambar 2.




Gambar 2. Antarmuka Awal Perangkat Lunak Pengelompokkan Performa Akademik Mahasiswa

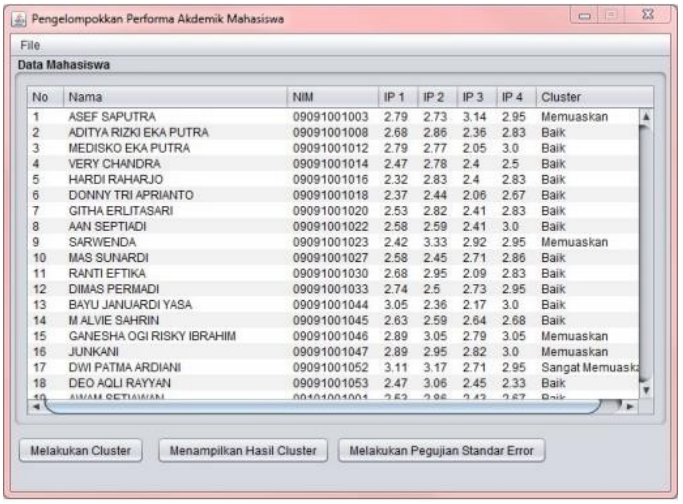
Pada antarmuka ini, hanya menu File yang aktif dan dapat diakses.Menu ini digunakan untuk mengunggah berkas berisi Nama, NIM, dan IP semester 1-4 dari sekelompok mahasiswa. Jika berkas telah diunggah, maka data-data tersebut akan ditampilkan pada panel yang berada di tengah, seperti pada Gambar 3.




Gambar 3. Tampilan Data yang Telah Diunggah pada Sistem

Setelah data diunggah, maka tombol melakukan *cluster* dan Melakukan Pengujian Standard Error menjadi aktif. Pengguna dapat menekan tombol melakukan *cluster* untuk mengelompokkan data menggunakan algoritma K-Means sebanyak 1 kali, sedangkan tombol pengujian akan mengulang pengelompokkan sebanyak n kali sebagai bagian dan proses pengukuran akurasi algoritma K-Means *Clustering*.

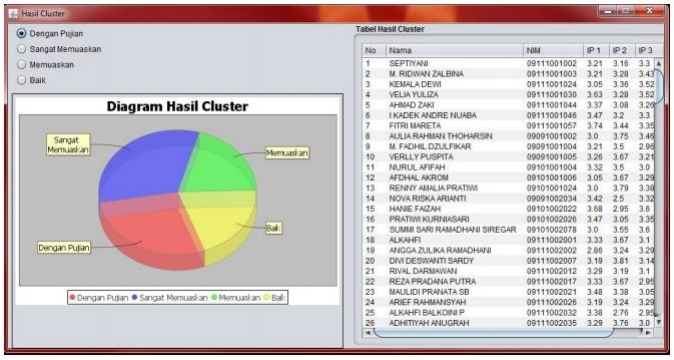
Jika pengguna menekan tombol Melakukan *cluster*, maka proses *clustering* seperti pada Use Case Melakukan *cluster*akan dijalankan. Diakhir proses, *cluster* dari masing-masing mahasiswa akan ditampilkan pada panel, seperti pada Gambar 4.




Gambar 4. Hasil Pengelompokkan Performa Akademik Mahasiswa.

Pada perangkat lunak yang dibangun, *centroid* pada masing-masing *cluster* terlebih dahulu diurutkan, dan diberi label "dengan pujian", "sangat memuaskan", "memuaskan", dan "baik". Predikat ini diberikan berdasarkan Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Tentang Standard Nasional Pendidikan Tinggi [6].

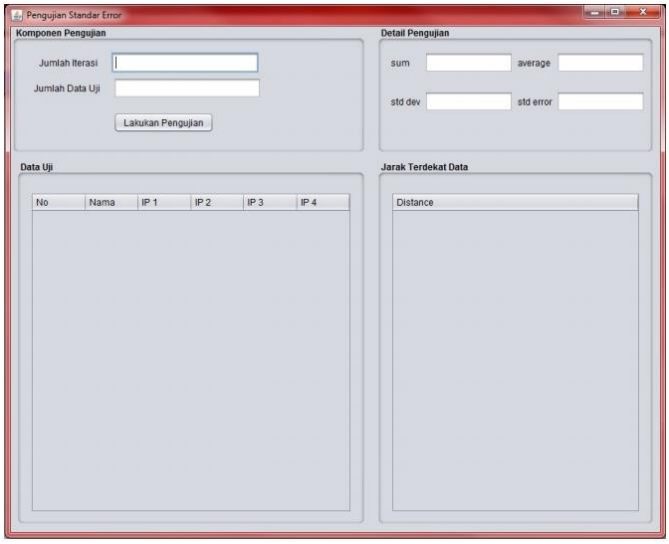
Selanjutnya, pengguna dapat memilih Menampilkan hasil *cluster*. Tombol ini adalah representasi dari Use Case "Menampilkan Hasil *Cluster*", yang akan menampilkan hasil dalam bentuk *pie-chart*. Tampilan ini dapat dilihat pada gambar 5.




Gambar 5.Hasil pengelompokkan dalam bentuk grafik *pie-chart*.

Grafik persentase hasil pengelompokkan ditampilkan bersebelahan dengan hasil *clustering* dalam bentuk tabel.Hal ini dimaksudkan untuk mempermudah pengambilan keputusan mengenai mahasiswa dengan performa akademik tertentu.

Tombol lainnya yang dapat diakses adalah tombol Melakukan Pengujian Standard Eror yang merupakan implementasi dari use case "Melakukan Pengujian Standard Error. Tampilan dari fitur ini dapat dilihat pada Gambar 6.




Gambar 6. Tampilan Pengujian Standard Error

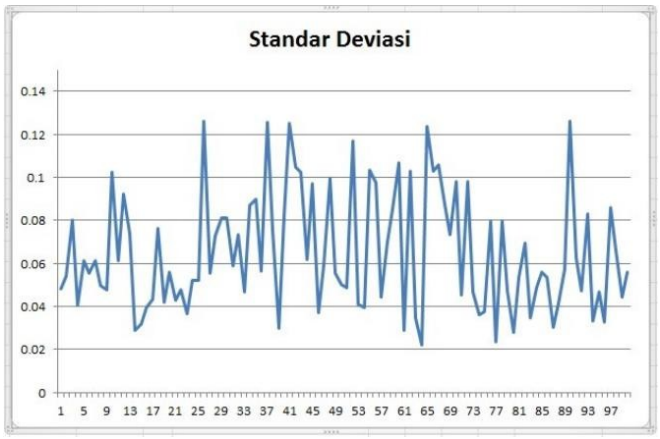
Pada fitur ini, pengguna dapat memilih jumlah iterasi dan banyak data uji yang ingin digunakan dari data yang telah diunggah sebelumnya. Jika Tombol lakukan pengujian ditekan, maka perangkat lunak akan melakukan *clustering* dari data uji yang dipilih sebanyak jumlah iterasi. Perangkat lunak kemudian menampilkan nilai sum, rata-rata, standard deviasi, dan standard error dari pengujian.

*Pengujian Performa K-Means Clustering*

Pada penelitian ini, performa dari K-Means *clustering* dinilai dengan menghitung standard deviasi dan standard error dari proses pengelompokkan yang dilakukan berkali-kali. Standard deviasi adalah ukuran yang digunakan untuk mengukur sebaran data. Sebuah standard deviasi yang rendah menunjukkan bahwa titik data cenderung dekat dengan mean dari himpunan, sementara standard deviasi yang tinggi menunjukkan bahwa titik data tersebar lebih besar dari rentang nilai. Standardd error adalah standard deviasi dari distribusi sampling statistik yang mengukur akurasi sampel yang mewakili populasi.

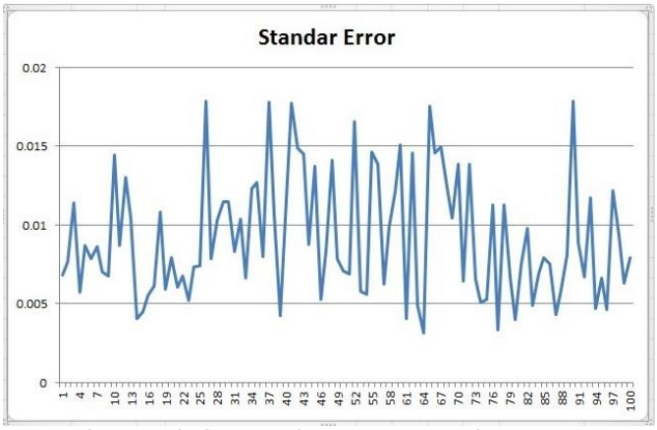
Sebanyak 100 sampel data mahasiswa semester 5 fakultas ilmu komputer Universitas Sriwijaya telah dikumpulkan sebagai data uji yang akan dikelompokkan pada penelitian ini. Data yang digunakan sebagai parameter performa akademik adalah data Indeks Prestasi Mahasiswa semester
1 sampai dengan semester 4.Data tersebut kemudian dikelompokkan sebanyak 50 kali.

Grafik standard deviasi dari 100 data dapat dilihat paga Gambar 7.




Gambar 7. Standard Deviasi Hasil Pengujian K-Means *Clustering*

Berdasarkan pengujian, standard deviasi maksimum dari 100 data adalah sebesar 0.012616 dan minimum sebesar 0.022089.Sedangkan, rata-rata standard deviasi dari masing-masing data adalah sebesar 0.064926.Grafik nilai standard error dari 100 data dapat dilihat pada Gambar 8.




Gambar 8. Standard Error Hasil Pengujian K-Means *Clustering*

Berdasarkan pengujian, standard error maksimum dari 100 data adalah sebesar 0.017842 dan minimum sebesar 0.003124.Sedangkan, rata-rata standard error dari masing-masing data adalah sebesar 0.009182.

Hasil standard error dari 100 kali pengelompokkan data adalah 0.009182 atau 0.9182%. Hal ini menunjukkan bahwa pengelompokkan yang dihasilkan oleh proses K-Means *clustering* tidak banyak mengalami perubahan. Hal ini sejalan dengan kebutuhan yang menginginkan hasil pengelompokkan yang konsisten.

Berdasarkan hasil dan pembahasan dari penelitian yang telah dilakukan, dapat disimpulkan bahwa:

1. Metode K-Means *Clustering* dapat digunakan untuk melakukan pengelompokkan performa akademik mahasiswa.

2. Standard Error rata rata dari data yang diuji adalah sebesar 0.92% yang menunjukkan bahwa metode K–Means *Clustering* dapat melakukan pengelompokkan performa akademik mahasiswa dengan baik.

3. *Cluster* akhir dari suatu data berubah – ubah sesuai dengan jarak terdekat data terhadap *centroid*.

Sumber :

**PENGELOMPOKKAN PERFORMA AKADEMIK MAHASISWA ...**

www.seminar.ilkom.unsri.ac.id

# TUGAS 05

**Cara menghitung perfomance sebuah cluster dengan Study Kasus**

## Clustering Penilaian Kinerja Dosen Menggunakan Algoritma K-Means (Studi Kasus: Universitas Dehasen Bengkulu)

*https://prosiding.seminar-id.com/index.php/sainteks/article/download/218/213*

*Clustering* atau pengklasteran adalah suatu teknik data mining yang digunakan untuk menganalisis data untuk memecahkan permasalahan dalam pengelompokkan data atau lebih tepatnya mempartisi dari dataset ke dalam subset. Pada teknik clustering targetnya adalah untuk kasus pendistribusian (objek, orang, peristiwa dan lainnya) ke dalam suatu kelompok, hingga derajat tingkat keterhubungan antar anggota cluster yang sama adalah kuat dan lemah antara anggota cluster yang berbeda

Analisa Data:

Data dosen akan diolah menggunakan algoritma *K-Mean* untuk mengelompokkan data dosen tersebut kedalam kelompok "Dosen Sangat Baik", "Dosen Baik", "Dosen Cukup Baik" atau "Dosen Kurang Baik" berdasarkan variabel kuisoner sebagai input.

**Tabel 1.** Data Rekapitulasi Kegiatan Dosen Mengajar

| NO. | NAMA DOSEN | EVALUASI DOSEN (BERDASARKAN KUISONER) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 1 | AJI SUDARSONO, | **4,00** | **1,11** | **2,89** | **1,95** | **1,32** | **1,68** | **2,84** | **2,74** | **2,84** | **1,11** | **3,79** | **3,42** | **2,42** | **3,05** |
| 2 | Dr. MESTERJON | 4,00 | 3,00 | 2,00 | 1,00 | 2,00 | 3,00 | 1,08 | 3,88 | 2,08 | 1,04 | 4,00 | 2,00 | 2,00 | 3,00 |
| 3 | DEWI SURANTI | 4,00 | 3,00 | 2,70 | 1,80 | 2,05 | 2,60 | 2,40 | 2,70 | 2,10 | 2,55 | 3,50 | 2,40 | 2,40 | 1,45 |
| 4 | DIAN MARDIATI SARI | 4,00 | 3,70 | 4,00 | 2,70 | 2,00 | 2,50 | 2,40 | 2,60 | 2,50 | 2,35 | 3,00 | 2,70 | 2,90 | 3,00 |
| 5 | DIMAS AULIA | 4,00 | 3,91 | 4,00 | 4,00 | 4,00 | 4,00 | 4,00 | 4,00 | 4,00 | 4,00 | 4,00 | 4,00 | 3,82 | 3,55 |
| 6 | MARYANINGSIH | 4,00 | 2,84 | 2,53 | 2,42 | 2,58 | 2,63 | 2,68 | 2,63 | 2,63 | 2,32 | 2,58 | 2,74 | 2,79 | 2,74 |
| 7 | EKO PRASETYO | 4,00 | 2,13 | 2,30 | 2,04 | 1,87 | 2,39 | 2,43 | 2,13 | 2,35 | 2,39 | 3,65 | 3,52 | 2,74 | 3,26 |

| NO. | NAMA DOSEN | EVALUASI DOSEN (BERDASARKAN KUISONER) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 8 | PRAHASTI | 4,00 | 3,00 | 4,00 | 3,96 | 4,00 | 3,96 | 4,00 | 4,00 | 3,96 | 3,08 | 4,00 | 3,92 | 3,92 | 3,08 |
| 9 | HARI ASPRIYONO | 4,00 | 3,10 | 3,90 | 2,60 | 2,80 | 3,70 | 2,90 | 3,60 | 3,40 | 3,30 | 3,60 | 2,40 | 3,40 | 3,90 |
| 10 | HERLINA LATIPA SARI, | 4,00 | 2,62 | 2,57 | 2,57 | 2,43 | 2,76 | 2,14 | 2,29 | 2,19 | 2,62 | 2,76 | 2,95 | 2,86 | 2,90 |
| 11 | JUJU JUMADI | 4,00 | 2,68 | 2,86 | 2,91 | 2,91 | 2,95 | 2,86 | 2,55 | 2,68 | 2,68 | 2,95 | 2,68 | 2,73 | 3,05 |
| 12 | Ir. H. JUSUF WAHYUDI | 4,00 | 1,70 | 2,96 | 2,00 | 2,17 | 2,39 | 2,13 | 2,52 | 2,35 | 2,17 | 3,83 | 3,39 | 2,57 | 3,48 |
| 13 | LENI NATALIA | 4,00 | 3,00 | 3,14 | 3,66 | 3,48 | 3,86 | 3,72 | 3,86 | 3,90 | 3,38 | 3,86 | 2,93 | 3,90 | 3,97 |
| 14 | LIZA YULIANTI | 4,00 | 3,05 | 3,00 | 3,05 | 2,80 | 3,00 | 3,05 | 2,55 | 2,80 | 2,85 | 3,05 | 2,95 | 2,95 | 2,70 |
| 15 | YUPIANTI | 4,00 | 1,32 | 2,89 | 1,95 | 1,47 | 1,79 | 2,84 | 2,74 | 2,84 | 1,26 | 3,79 | 3,42 | 2,42 | 3,05 |
| 16 | SAPRI | 4,00 | 3,93 | 3,93 | 2,67 | 3,41 | 3,52 | 4,00 | 3,96 | 2,85 | 3,59 | 2,30 | 3,00 | 2,19 | 3,00 |
| 17 | YODE ARLIANDO | 4,00 | 3,00 | 3,80 | 2,40 | 2,90 | 3,20 | 3,80 | 4,00 | 2,30 | 3,00 | 4,00 | 2,00 | 2,00 | 2,00 |

Data pada tabel 2. ini yang akan digunakan sebagai data sampel dalam mengelompokkan dosen ke dalam dosen "Sangat Baik ($C_1$)", "Baik ($C_2$)", "Cukup Baik ($C_3$)", dan "Kurang Baik ($C_4$)". Pengelompokan tersebut berdasarkan atribut point pernyataan kuisoner seperti pada tabel 1. di atas. Untuk dapat melakukan pengelompokkan data-data tersebut menjadi beberapa *cluster* perlu dilakukan beberapa langkah langkah yaitu:

*1.* Titik pusat awal dari setiap *cluster*

**Tabel 2.** Titik pusat awal dari setiap *cluster*

| NAMA DOSEN | EVALUASI DOSEN (BERDASARKAN KUISONER) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| MARYANINGSIH | 4,00 | 2,84 | 2,53 | 2,42 | 2,58 | 2,63 | 2,68 | 2,63 | 2,63 | 2,32 | 2,58 | 2,74 | 2,79 | 2,74 |
| HARI ASPRIYONO | 4,00 | 3,10 | 3,90 | 2,60 | 2,80 | 3,70 | 2,90 | 3,60 | 3,40 | 3,30 | 3,60 | 2,40 | 3,40 | 3,90 |
| JUJU JUMADI | 4,00 | 2,68 | 2,86 | 2,91 | 2,91 | 2,95 | 2,86 | 2,55 | 2,68 | 2,68 | 2,95 | 2,68 | 2,73 | 3,05 |
| YODE ARLIANDO | 4,00 | 3,00 | 3,80 | 2,40 | 2,90 | 3,20 | 3,80 | 4,00 | 2,30 | 3,00 | 4,00 | 2,00 | 2,00 | 2,00 |

*2.* Perhitungan Jarak Pusat *Cluster*

Untuk menghitung jarak antara data dengan pusat *cluster* awal menggunakan persamaan *Eucliden Distance* sebagai berikut:

$$D_{i,j} = \sqrt{(x_{1i} - x_{ij})^2 + (x_{2i} - x_{2j})^2 + \ldots + (x_{ki} - x_{kj})^2}$$

Maka di dapatkan nilai matrik sebagai berikut :

$$D_{11} = \sqrt{\begin{matrix}(4{,}00-4{,}00)^2+(1{,}11-1{,}11)^2+(2{,}89-2{,}89)^2+(1{,}95-1{,}95)^2\\+(1{,}32-1{,}32)^2+(1{,}68-1{,}68)^2+(2{,}84-2{,}84)^2+(2{,}74-2{,}74)^2\\+(2{,}84-2{,}84)^2+(1{,}11-1{,}11)^2+(3{,}79-3{,}79)^2+(3{,}42-3{,}42)^2\\+(2{,}42-2{,}42)^2+(3{,}05-3{,}05)^2\end{matrix}} = 0{,}000$$

$$D_{12} = \sqrt{\begin{matrix}(4{,}00-4{,}00)^2+(3{,}00-1{,}11)^2+(2{,}00-2{,}89)^2+(1{,}00-1{,}95)^2\\+(2{,}00-1{,}32)^2+(3{,}00-1{,}68)^2+(1{,}08-2{,}84)^2+(3{,}88-2{,}74)^2\\+(2{,}08-2{,}84)^2+(1{,}04-1{,}11)^2+(4{,}00-3{,}79)^2+(2{,}00-3{,}42)^2\\+(2{,}00-2{,}42)^2+(3{,}00-3{,}05)^2\end{matrix}} = 3{,}838$$

$$D_{13} = \sqrt{\begin{matrix}(4{,}00-4{,}00)^2+(3{,}00-1{,}11)^2+(2{,}70-2{,}89)^2+(1{,}80-1{,}95)^2\\+(2{,}05-1{,}32)^2+(2{,}60-1{,}68)^2+(2{,}40-2{,}84)^2+(2{,}70-2{,}74)^2\\+(2{,}10-2{,}84)^2+(2{,}55-1{,}11)^2+(3{,}50-3{,}79)^2+(2{,}40-3{,}42)^2\\+(2{,}40-2{,}42)^2+(1{,}45-3{,}05)^2\end{matrix}} = 3{,}400$$

$$D_{14} = \sqrt{\begin{matrix}(4{,}00-4{,}00)^2+(3{,}70-1{,}11)^2+(4{,}00-2{,}89)^2+(2{,}70-1{,}95)^2\\+(2{,}00-1{,}32)^2+(2{,}50-1{,}68)^2+(2{,}40-2{,}84)^2+(2{,}60-2{,}74)^2\\+(2{,}50-2{,}84)^2+(2{,}35-1{,}11)^2+(3{,}00-3{,}79)^2+(2{,}70-3{,}42)^2\\+(2{,}90-2{,}42)^2+(3{,}00-3{,}05)^2\end{matrix}} = 3{,}593$$

Dan seterusnya dilanjutkan menghitung untuk data ke-2 sampai ke -17 terhadap pusat awal *cluster.* Hasil perhitungan selengkapnya untuk 17 data dosen dapat dilihat pada tabel berikut:

**Tabel 4.** Hasil Perhitungan Jarak Iterasi 1.

| Data | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
|---|---|---|---|---|
| 1 | 0,000 | 3,384 | 4,507 | 4,378 |
| 2 | 3,838 | 3,928 | 4,543 | 4,284 |
| 3 | 3,400 | 2,418 | 3,767 | 2,724 |
| 4 | 3,593 | 1,949 | 2,633 | 3,099 |
| 5 | 6,411 | 4,174 | 3,065 | 4,347 |
| 6 | 3,096 | 1,005 | 2,933 | 3,082 |
| 7 | 2,196 | 2,134 | 3,576 | 3,891 |
| 8 | 5,644 | 3,731 | 2,915 | 3,946 |
| 9 | 4,507 | 2,367 | 0,000 | 2,912 |
| 10 | 3,175 | 1,204 | 3,088 | 3,454 |
| 11 | 3,384 | 0,000 | 2,367 | 2,778 |
| 12 | 1,933 | 2,247 | 3,371 | 3,726 |
| 13 | 5,252 | 3,069 | 1,949 | 3,722 |
| 14 | 3,639 | 0,718 | 2,332 | 2,693 |
| 15 | 0,324 | 3,102 | 4,242 | 4,130 |
| 16 | 5,366 | 2,879 | 2,674 | 2,634 |
| 17 | 4,378 | 2,778 | 2,912 | 0,000 |

Selanjutnya melakukan perbandingan dan memilih jarak yang paling dekat dengan data dengan pusat *cluster*, jarak ini menunjukan bahwa data memiliki jarak terdekat berada dalam satu kelompok dengan pusat *cluster* terdekat .

**Tabel 5.** Hasil Pengelompokan Iterasi 1

| Kelompok Cluster | Anggota Kelompok | Jumlah |
|---|---|---|
| $C_1$ | 1,2,12,15 | 4 |
| $C_2$ | 3,4,6,7,10,11,14 | 7 |
| $C_3$ | 5,8,9,13 | 4 |
| $C_4$ | 16,17 | 2 |

3. Penentuan Pusat *Cluster* Baru
   Setelah didapatkan anggota dari setiap *cluster* kemudian pusat *cluster* baru dihitung berdasarkan tiap-tiap anggota *cluster* yang sudah didapatkan dengan menggunakan rumus sebagai berikut:

$$C_1 = \frac{(4{,}00 + 4{,}00 + 4{,}00 + 4{,}0)}{4} = 4{,}00$$
$$= \frac{(1{,}11 + 3{,}00 + 1{,}70 + 1{,}32)}{4} = 1{,}779$$
$$= \frac{(2{,}89 + 2{,}00 + 2{,}95 + 2{,}89)}{4} = 2{,}686$$
$$= \frac{(1{,}95 + 1{,}00 + 2{,}00 + 1{,}95)}{4} = 1{,}724$$
$$C_2 = \frac{(4{,}00 + 4{,}00 + 4{,}00 + 4{,}00 + 4{,}00 + 4{,}00 + 4{,}00)}{7} = 4{,}00$$
$$= \frac{(3{,}00 + 3{,}00 + 3{,}00 + 3{,}70 + 2{,}84 + 2{,}13 + 3{,}05)}{7} = 2{,}860$$
$$= \frac{(2{,}70 + 4{,}00 + 2{,}54 + 2{,}30 + 2{,}57 + 2{,}86 + 3{,}00)}{7} = 2{,}852$$
$$= \frac{(1{,}80 + 2{,}70 + 2{,}42 + 2{,}04 + 2{,}57 + 2{,}91 + 3{,}05)}{7} = 2{,}499$$
$$C_3 = \frac{(4{,}00 + 4{,}00 + 4{,}00 + 4{,}00)}{4} = 4{,}00$$
$$= \frac{(3{,}90 + 3{,}00 + 3{,}10 + 3{,}00)}{4} = 3{,}252$$
$$= \frac{(4{,}00 + 4{,}00 + 3{,}00 + 3{,}14)}{4} = 3{,}759$$
$$= \frac{(4{,}00 + 3{,}96 + 2{,}60 + 3{,}66)}{4} = 3{,}553$$
$$C_4 = \frac{(4{,}00 + 4{,}00)}{2} = 4{,}00$$
$$= \frac{(3{,}93 + 3{,}00)}{2} = 3{,}465$$
$$= \frac{(3{,}93 + 3{,}80)}{2} = 3{,}865$$
$$= \frac{(2{,}67 + 2{,}40)}{2} = 2{,}535$$

Setelah semua nilai *centroid* baru diperoleh seperti yang terlihat pada tabel berikut ini:

**Tabel 5.** Pusat Cluster Baru

| CENTROID BARU | DATA 1 | DATA 2 | DATA 3 | DATA 4 | DATA 5 | DATA 6 | DATA 7 | DATA 8 | DATA 9 | DATA 10 | DATA 11 | DATA 12 | DATA 13 | DATA 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **C1** | 4,000 | 1,779 | 2,686 | 1,724 | 1,741 | 2,216 | 2,224 | 2,969 | 2,528 | 1,396 | 3,851 | 3,058 | 2,352 | 3,146 |
| **C2** | 4,000 | 2,860 | 2,852 | 2,499 | 2,377 | 2,691 | 2,568 | 2,492 | 2,465 | 2,537 | 3,071 | 2,849 | 2,766 | 2,728 |
| **C3** | 4,000 | 3,252 | 3,759 | 3,553 | 3,571 | 3,880 | 3,656 | 3,866 | 3,814 | 3,441 | 3,866 | 3,312 | 3,758 | 3,624 |
| **C4** | 4,000 | 3,465 | 3,865 | 2,535 | 3,155 | 3,360 | 3,900 | 3,980 | 2,575 | 3,295 | 3,150 | 2,500 | 2,095 | 2,500 |

Langkah selanjutnya hitung nilai *Euclidean Distance* dari semua data ke titik pusat *cluster* baru ($C_1$,$C_2$,$C_3$,$C_4$) seperti yang dilakukan seperti pada langkah ke-2.

**Tabel 6.** Hasil Pengelompokan *Cluster* 2

| Kelompok Cluster | Anggota Kelompok | Jumlah |
|---|---|---|
| $C_1$ | 1,2,7,12,15 | 5 |
| $C_2$ | 3,4,6,10,11,14 | 6 |
| $C_3$ | 5,8,9,13 | 4 |
| $C_4$ | 16,17 | 2 |

**Tabel 7.** Hasil Pengelompokan *Cluster* 3

| Kelompok Cluster | Anggota Kelompok | Jumlah |
|---|---|---|
| $C_1$ | 1,2,7,12,15 | 5 |
| $C_2$ | 3,4,6,10,11,14 | 6 |
| $C_3$ | 5,8,9,13 | 4 |
| $C_4$ | 16,17 | 2 |

Pada pengujian iterasi ke-2 dan iterasi ke-3 tidak mengalami perubahan (sama dengan *centroid* sebelumnya) maka proses iterasi selesai dan diperoleh 4 *cluster* dengan 3 iterasi. Berikut tabel hasil akhir analisa perhitungan kinerja dosen dengan menggunakan algoritma *K-Means*

**Tabel 8.** Hasil Analisa *K-Means Clustering*

| *Cluster* | Hasil Akhir | | |
|---|---|---|---|
| | **Centroid Akhir** | **Jumlah Anggota** | **Nama Dosen** |
| 1 | 1,2,7,12,15 | 5 | Aji Sudarsono, M.Kom<br>Mesterjon, M.Kom<br>Eko Prasetyo.R, M.Kom<br>Ir. H. Jusuf Wahyudi, M.Kom<br>Yupianti, M.Kom |
| 2 | 3,4,6,10,11,14 | 6 | Dewi Suranti, M.Kom<br>Dian Mardiati Sari, M.Si<br>Dra. Hj. Maryaningsih, M.Kom<br>Herlina Latipa Sari, M.Kom<br>Juju Jumadi, M.Kom<br>Liza Yulianti, M.Kom |
| 3 | 5,8,9,13 | 4 | Dimas Aulia.T, M.Kom<br>Prahasti, M. Kom<br>Hari Aspriyono, M.Kom<br>Leni Natalia Z. M. Kom |

| | | | |
|---|---|---|---|
| 4 | 16,17 | 2 | Sapri, M.Kom<br>Yode Arlindo, M.Kom |

Nama : Andry Meylani
NIM : 202420009
TUGAS 05

Pelajari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format ms word dan dilengkapi dengan gambar dan keterangan yang jelas.

Jawaban :
Pada tugas ini saya mengambil contoh Penerapan Metode K-Means Clustering untuk Menentukan Persediaan Stok Barang Pada Toko.

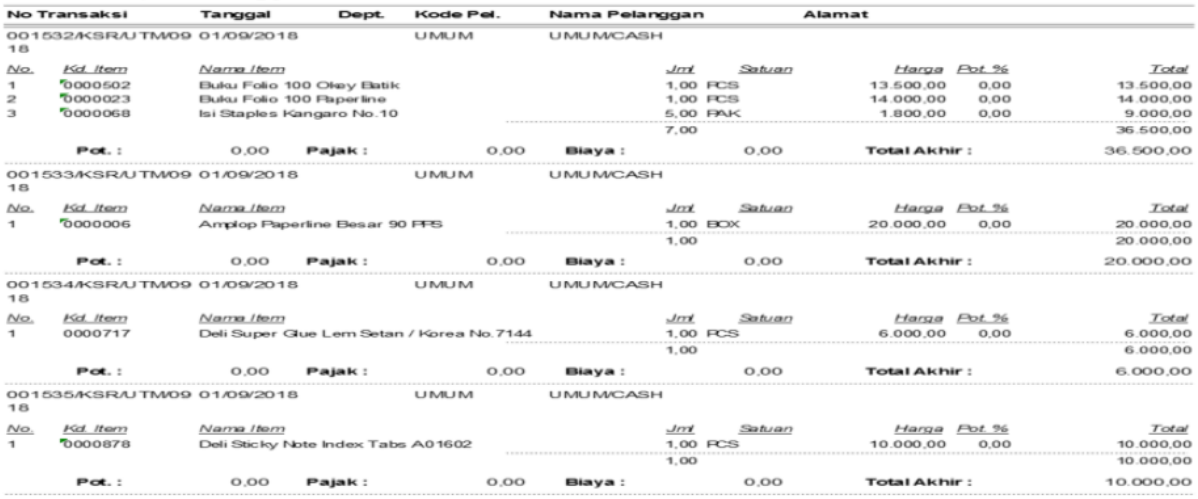
1. Data Penjualan
   Dibawah ini adalah tampilan data penjualan pada Toko;

| Kode Item | Nama Item | Jenis | Merek | Jumlah Terjual | Satuan |
|---|---|---|---|---|---|
| 0000006 | Amplop Paperline Besar 90 PPS | Amplop | Paperline | 52,00 | PAK |
| 0000007 | Amplop Paperline Kecil 104 PPS | Amplop | Paperline | 82,00 | PAK |
| 0000016 | Binder Clip Kenko No.107 | Misc ATK | Nomerk | 10,00 | BOX |
| 0000017 | Binder Clip Joyko No.111 | Misc ATK | Joyko | 8,00 | BOX |
| 0000021 | Buku Tulis Big Boss 42 | Buku | Bigboss | 70,00 | PCS |
| 0000023 | Buku Folio 100 Paperline | Buku | Paperline | 9,00 | PCS |
| ..... | | | | | |
| 0000822 | Map Dokumen Clear Holder 40 Pocket | Map / File | Nomerk | 5,00 | PCS |
| 0000825 | Buku Tulis Sidu Kotak 38 Lembar | Buku | Sidu | 50,00 | PCS |
| 0000826 | Buku Tulis Sidu 100 Lembar | Buku | Sidu | 15,00 | PCS |

2. Data Persediaan
   Dibawah ini tampilan data persediaan barang pada Toko;

| Kode Item | Nama Item | Jenis | Merek | Jumlah Stok | Satuan |
|---|---|---|---|---|---|
| 0000006 | Amplop Paperline Besar 90 PPS | Amplop | Paperline | 72,00 | PAK |
| 0000007 | Amplop Paperline Kecil 104 PPS | Amplop | Paperline | 120,00 | PAK |
| 0000013 | Bantal Stempel Besar Hero E1460 Violet | Misc ATK | Hero | 12,00 | PCS |
| 0000016 | Binder Clip Kenko No.107 | Misc ATK | Nomerk | 24,00 | BOX |
| 0000017 | Binder Clip Joyko No.111 | Misc ATK | Joyko | 12,00 | BOX |
| 0000029 | Buku Kwitansi Sedang Sidu / Paperline 40 lembar | Buku Kwintansi | Sidu | 24,00 | PCS |
| 0000037 | Buku Nota Besar 3 Ply Paperline 25 set | Buku Nota | Paperline | 60,00 | PCS |
| 0000038 | Buku Nota Kecil 1 Ply Paperline 50 lbr | Buku Nota | Paperline | 84,00 | PCS |
| ..... | | | | | |
| 0000760 | Deli Pena Ballpoint No. 6546 | Pena Pulpen | Deli | 24,00 | PCS |
| 0000764 | Bantex Ordner 1465V Blue | File | Bantex | 12,00 | PCS |
| 0000772 | Bantex Jumbo Box File Blue 4011-62 | File | Bantex | 12,00 | PCS |
| 0000773 | Pembolong Kertas Cadwell CD-30S | Misc ATK | Cadwell | 12,00 | PCS |
| 0000775 | Jangka Set Nikiko TS-688 | Alat Gambar | Nomerk | 12,00 | PCS |
| 0000776 | Pensil Warna JOYKO Kaleng CP-12RT | Alat Gambar | Nomerk | 12,00 | PCS |
| 0000777 | Lakban Bening 1" Daimaru | Lakban | Daimaru | 48,00 | PCS |

3. Data Transaksi
   Di bawah ini tampilan data transaksi pada Toko;



| No Transaksi | Tanggal | Dept. | Kode Pel. | Nama Pelanggan | Alamat |
|---|---|---|---|---|---|
| 001532/KSR/UTM/09 18 | 01/09/2018 | | UMUM | UMUM/CASH | |

| No. | Kd. Item | Nama Item | Jml | Satuan | Harga | Pot. % | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0000502 | Buku Folio 100 Okey Batik | 1,00 | PCS | 13.500,00 | 0,00 | 13.500,00 |
| 2 | 0000023 | Buku Folio 100 Paperline | 1,00 | PCS | 14.000,00 | 0,00 | 14.000,00 |
| 3 | 0000068 | Isi Staples Kangaro No.10 | 5,00 | PAK | 1.800,00 | 0,00 | 9.000,00 |
| | | | 7,00 | | | | 36.500,00 |

Pot. : 0,00 Pajak : 0,00 Biaya : 0,00 Total Akhir : 36.500,00

| 001533/KSR/UTM/09 18 | 01/09/2018 | | UMUM | UMUM/CASH | |
|---|---|---|---|---|---|

| No. | Kd. Item | Nama Item | Jml | Satuan | Harga | Pot. % | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0000006 | Amplop Paperline Besar 90 PPS | 1,00 | BOX | 20.000,00 | 0,00 | 20.000,00 |
| | | | 1,00 | | | | 20.000,00 |

Pot. : 0,00 Pajak : 0,00 Biaya : 0,00 Total Akhir : 20.000,00

| 001534/KSR/UTM/09 18 | 01/09/2018 | | UMUM | UMUM/CASH | |
|---|---|---|---|---|---|

| No. | Kd. Item | Nama Item | Jml | Satuan | Harga | Pot. % | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0000717 | Deli Super Glue Lem Setan / Korea No.7144 | 1,00 | PCS | 6.000,00 | 0,00 | 6.000,00 |
| | | | 1,00 | | | | 6.000,00 |

Pot. : 0,00 Pajak : 0,00 Biaya : 0,00 Total Akhir : 6.000,00

| 001535/KSR/UTM/09 18 | 01/09/2018 | | UMUM | UMUM/CASH | |
|---|---|---|---|---|---|

| No. | Kd. Item | Nama Item | Jml | Satuan | Harga | Pot. % | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0000878 | Deli Sticky Note Index Tabs A01602 | 1,00 | PCS | 10.000,00 | 0,00 | 10.000,00 |
| | | | 1,00 | | | | 10.000,00 |

Pot. : 0,00 Pajak : 0,00 Biaya : 0,00 Total Akhir : 10.000,00

4. Data Cleaning dan Penggabungan Data

Hasil gabungan dari ketiga data tersebut dapat dilihat pada tampilan data gabungan;

| Kode Item | Nama Item | Jenis | Merek | Jumlah Terjua | Jumlah Stok | Volume Penjual | Satuan |
|---|---|---|---|---|---|---|---|
| 0000006 | Amplop Paperline Besar 90 PPS | Amplop | Paperline | 52,00 | 72,00 | 6,00 | PAK |
| 0000007 | Amplop Paperline Kecil 104 PPS | Amplop | Paperline | 82,00 | 120,00 | 18,00 | PAK |
| 0000023 | Buku Folio 100 Paperline | Buku | Paperline | 9,00 | 24,00 | 7,00 | PCS |
| 0000027 | Buku Kwitansi Besar Sidu 50 lembar | Buku Kwintansi | Sidu | 7,00 | 24,00 | 5,00 | PCS |
| 0000028 | Buku Kwitansi Kecil Sidu/PPL 40 lembar | Buku Kwintansi | Sidu | 13,00 | 36,00 | 8,00 | PCS |
| 0000029 | Buku Kwitansi Sedang Sidu / Paperline 40 lembar | Buku Kwintansi | Sidu | 4,00 | 24,00 | 3,00 | PCS |
| ..... | | | | | | | |
| 0000440 | Buku Milimeter Folio Cox | Buku | Cox | 3,00 | 12,00 | 1,00 | PCS |
| 0000441 | Peruncing Daun Fancy 033 041 | Alat Tulis | Nomerk | 5,00 | 24,00 | 5,00 | PCS |
| 0000442 | Pena Joyko Lino BP-249 | Pena Pulpen | Joyko | 47,00 | 60,00 | 13,00 | PCS |
| 0000443 | Tas Ransel Smiley Emoticon AD-200 | Tas | Nomerk | 5,00 | 6,00 | 5,00 | PCS |
| 0000444 | Paper Spear Dingli DL 85201 | Misc ATK | Nomerk | 4,00 | 12,00 | 3,00 | PCS |
| 0000446 | Jangka Besi Space Ball SP4001 | Alat Gambar | Nomerk | 4,00 | 6,00 | 4,00 | PCS |
| 0000447 | Stapler HD-45 V-tec | Stapler / Staples | V-tec | 2,00 | 12,00 | 2,00 | PCS |
| 0000448 | Paper Fastener Eco-nomical | Misc ATK | Nomerk | 5,00 | 12,00 | 2,00 | BOX |
| 0000450 | Pensil Warna 12W Panjang Disney | Alat Gambar | Nomerk | 1,00 | 12,00 | 1,00 | SET |

Pada penelitian ini data yang diambil sebagai bahan penilitian yaitu ± 1000 data item, kemudian dari data tersebut penulis kembali melakukan proses data cleaning berikutnya dengan memilih 62 data item saja secara random untuk dijadikan data sampel yang akan diolah selanjutnya, dan beberapa atribut yang berpengaruh, atribut tersebut meliputi kode item, jenis item, jumlah terjual, jumlah stok, volume penjualan, rata-rata.

Berikut hasil Data Sample;

| No | Kode Item | Jenis | Jumlah Terjual | Jumlah Stok | Volume Penjualan | Rata-Rata |
|---|---|---|---|---|---|---|
| 1 | 0000369 | Alat Gambar | 17,00 | 60,00 | 10,00 | 14,50 |
| 2 | 0001174 | Alat Gambar | 12,00 | 24,00 | 12,00 | 8,00 |
| 3 | 0000466 | Alat Tulis | 138,00 | 240,00 | 62,00 | 73,33 |
| 4 | 0000882 | Alat Tulis | 33,00 | 60,00 | 19,00 | 18,67 |
| 5 | 0000007 | Amplop | 82,00 | 120,00 | 18,00 | 36,67 |
| 6 | 0000006 | Amplop | 52,00 | 72,00 | 6,00 | 21,67 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 7 | 0000635 | Botol Minum | 6,00 | 10,00 | 4,00 | 3,33 |
| 8 | 0001404 | Botol Minum | 3,00 | 5,00 | 2,00 | 1,67 |
| 9 | 0000263 | Buku | 29,00 | 48,00 | 9,00 | 14,33 |
| 10 | 0000021 | Buku | 70,00 | 96,00 | 12,00 | 29,67 |
| 11 | 0000613 | Buku Binder | 4,00 | 5,00 | 4,00 | 2,17 |
| 12 | 0000606 | Buku Binder | 1,00 | 5,00 | 1,00 | 1,17 |
| 13 | 0001393 | Buku Diary | 6,00 | 12,00 | 6,00 | 4,00 |
| 14 | 0001395 | Buku Diary | 10,00 | 12,00 | 8,00 | 5,00 |
| 15 | 0000028 | Buku Kwintansi | 13,00 | 36,00 | 8,00 | 9,50 |
| 16 | 0000027 | Buku Kwintansi | 7,00 | 24,00 | 5,00 | 6,00 |
| 17 | 0000039 | Buku Nota | 23,00 | 60,00 | 3,00 | 14,33 |
| 18 | 0000794 | Buku Nota | 20,00 | 60,00 | 5,00 | 14,17 |
| 19 | 0001350 | Buku Sampul | 69,00 | 100,00 | 7,00 | 29,33 |
| 20 | 0000783 | Buku Sampul | 11,00 | 20,00 | 6,00 | 6,17 |
| 21 | 0000503 | File | 34,00 | 48,00 | 6,00 | 14,67 |
| 22 | 0001011 | File | 2,00 | 12,00 | 2,00 | 2,67 |
| 23 | 0000347 | Gunting | 12,00 | 24,00 | 8,00 | 7,33 |
| 24 | 0000832 | Gunting | 12,00 | 24,00 | 12,00 | 8,00 |
| 25 | 0000674 | Kalkulator | 1,00 | 4,00 | 1,00 | 1,00 |
| 26 | 0000972 | Kalkulator | 3,00 | 4,00 | 3,00 | 1,67 |
| 27 | 0000095 | Kertas | 105,00 | 200,00 | 14,00 | 53,17 |
| 28 | 0001169 | Kertas | 52,00 | 60,00 | 45,00 | 26,17 |
| 29 | 0000360 | Kertas HVS | 1,00 | 5,00 | 1,00 | 1,17 |
| 30 | 0000358 | Kertas HVS | 32,00 | 40,00 | 15,00 | 14,50 |
| 31 | 0001566 | Kotak Makan | 2,00 | 5,00 | 2,00 | 1,50 |
| 32 | 0000651 | Kotak Makan | 2,00 | 6,00 | 2,00 | 1,67 |
| 33 | 0001511 | Kotak Pensil | 5,00 | 12,00 | 5,00 | 3,67 |
| 34 | 0001512 | Kotak Pensil | 3,00 | 12,00 | 2,00 | 2,83 |
| 35 | 0000296 | Lakban | 36,00 | 60,00 | 14,00 | 18,33 |
| 36 | 0001311 | Lakban | 46,00 | 60,00 | 14,00 | 20,00 |
| 37 | 0000352 | Lem | 24,00 | 36,00 | 14,00 | 12,33 |
| 38 | 0000717 | Lem | 19,00 | 24,00 | 16,00 | 9,83 |
| 39 | 0000165 | Map / File | 52,00 | 200,00 | 3,00 | 42,50 |
| 40 | 0000785 | Map / File | 7,00 | 100,00 | 2,00 | 18,17 |
| 41 | 0000697 | Misc | 1,00 | 6,00 | 1,00 | 1,33 |
| 42 | 0000892 | Misc | 2,00 | 24,00 | 1,00 | 4,50 |
| 43 | 0001578 | Misc Ultah | 8,00 | 36,00 | 4,00 | 8,00 |
| 44 | 0000711 | Misc Ultah | 8,00 | 24,00 | 5,00 | 6,17 |
| 45 | 0000083 | Misc ATK | 9,00 | 24,00 | 9,00 | 7,00 |
| 46 | 0000364 | Misc ATK | 9,00 | 12,00 | 8,00 | 4,83 |
| 47 | 0001419 | Pena Fancy | 60,00 | 60,00 | 46,00 | 27,67 |
| 48 | 0001529 | Pena Fancy | 31,00 | 36,00 | 26,00 | 15,50 |
| 49 | 0000242 | Pena Pulpen | 203,00 | 240,00 | 27,00 | 78,33 |
| 50 | 0000505 | Pena Pulpen | 95,00 | 120,00 | 50,00 | 44,17 |
| 51 | 0000782 | Pensil | 44,00 | 60,00 | 28,00 | 22,00 |
| 52 | 0000619 | Pensil | 43,00 | 60,00 | 21,00 | 20,67 |
| 53 | 0001486 | Post It / Stick Note / Memo | 4,00 | 12,00 | 4,00 | 3,33 |
| 54 | 0001489 | Post It / Stick Note / Memo | 3,00 | 12,00 | 3,00 | 3,00 |
| 55 | 0000715 | Stapler / Staples | 37,00 | 48,00 | 8,00 | 15,50 |
| 56 | 0000068 | Stapler / Staples | 91,00 | 120,00 | 27,00 | 39,67 |
| 57 | 0001142 | Tas | 13,00 | 16,00 | 12,00 | 6,83 |
| 58 | 0001134 | Tas | 13,00 | 16,00 | 13,00 | 7,00 |
| 59 | 0001227 | Tip Ex | 28,00 | 36,00 | 23,00 | 14,50 |
| 60 | 0000330 | Tip Ex | 26,00 | 36,00 | 21,00 | 13,83 |
| 61 | 0001129 | Tissue | 19,00 | 24,00 | 17,00 | 10,00 |
| 62 | 0001131 | Tissue | 4,00 | 12,00 | 4,00 | 3,33 |

Dari data yang telah ada pada tabel sampel diatas, selanjutnya akan dilakukan perhitungan manual k-means clustering terlebih dahulu. Adapun beberapa langkah yang dapat dilakukan dalam perhitungan k-means clustering ini yaitu antara lain :

1. Tahap Pertama

Menentukan jumlah cluster (K) terhadap data sampel yang sudah ada, maka untuk penentuan jumlah cluster (K), penulis menentukan dengan 5 cluster (K) seperti yang sudah dijelaskan sebelumnya, untuk proses perhitungan manual algoritma k-means yang akan dilakukan.

2. Tahap 2

Inisialisasi nilai centroid yaitu dengan menentukan nilai centroid awal pada data sampel yang sudah ada secara acak yaitu dipilih data dengan nomor urut ke-10 dengan kode item 0000021 dan jenis item buku, nomor urut ke-42 dengan 0000892 dan jenis item misc, nomor urut ke-34 dengan kode item 0001512 dan jenis item kotak pensil, nomor urut ke-19 dengan kode item 0001350 dan jenis item buku sampul, dan nomor urut ke-1 dengan kode item 0000369 dan jenis item alat gambar. Berikut dapat dilihat pada tabel dibawah ini nilai pada centroid awal yang ditentukan

| Cluster | Centroid Awal | | | |
|---|---|---|---|---|
| 1 | 70,00 | 96,00 | 12,00 | 29,67 |
| 2 | 2,00 | 24,00 | 1,00 | 4,50 |
| 3 | 3,00 | 12,00 | 2,00 | 2,83 |
| 4 | 69,00 | 100,00 | 7,00 | 29,33 |
| 5 | 17,00 | 60,00 | 10,00 | 14,50 |

3. Tahap Ketiga

Menghitung jarak antara centroid dengan data, yaitu menghitung jarak dari setiap data ke centroid terdekat. Centroid terdekat akan menjadi cluster yang diikuti oleh data tersebut. Untuk perhitungan jarak ke setiap centroid digunakan rumus Euclidean Distance seperti dibawah ini :

$d(xi, \mu j) = \sqrt{\sum}(xi - \mu j)2$ (4)

Dimana : d = jarak

$xi$ = data kriteria

$\mu j$ = centroid pada cluster ke-j

Jarak centroid data ke-1 pada cluster 1 adalah :

$d(xi, \mu j) = \sqrt{\sum}(xi - \mu j)2$ = √((17,00-70,00)^2+(60,00-96,00)^2+(10,00-12,00)^2+(14,50-29,67)^2) = 65,87

Jarak centroid data ke-1 pada cluster 2 adalah :

$d(xi, \mu j) = \sqrt{\sum}(xi - \mu j)2$ = √((17,00-2,00)^2+(60,00-24,00)^2+(10,00-1,00)^2+(14,50-4,50)^2 ) = 41,26

Jarak centroid data ke-1 pada cluster 3 adalah :

$d(xi, \mu j) = \sqrt{\sum}(xi - \mu j)2$ = √((17,00-3,00)^2+(60,00-12,00)^2+(10,00-2,00)^2+(14,50-2,83)^2 ) = 51,96

Jarak centroid data ke-1 pada cluster 4 adalah :

$d(xi, \mu j) = \sqrt{\sum}(xi - \mu j)2$ = √((17,00-69,00)^2+(60,00-100,00)^2+(10,00-7,00)^2+(14,50-29,33)^2 ) = 67,33

Jarak centroid data ke-1 pada cluster 5 adalah :

$d(xi, \mu j) = \sqrt{\sum}(xi - \mu j)2$ = √((17,00-17,00)^2+(60,00-60,00)^2+(10,00-10,00)^2+(14,50-14,50)^2 ) = 0

Untuk jarak centroid selanjutnya dapat dilihat pada tabel dibawah ini, yang mana penulis memasukkan tabel dari hasil proses perhitungan keseluruhan data sampel untuk jarak centroid iterasi yang pertama.

| No | Kode Item | Jenis | Jarak Centroid | | | | | Jarak Terdekat | Iterasi yang diikuti |
|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | | |
| 1 | 0000369 | Alat Gambar | 65,87 | 41,26 | 51,96 | 67,33 | 0,00 | 0,00 | 5 |
| 2 | 0001174 | Alat Gambar | 94,96 | 15,27 | 18,75 | 97,49 | 36,98 | 15,27 | 2 |
| 3 | 0000466 | Alat Tulis | 172,53 | 271,31 | 280,68 | 171,24 | 230,67 | 171,24 | 4 |
| 4 | 0000882 | Alat Tulis | 53,24 | 52,74 | 61,19 | 56,16 | 18,82 | 18,82 | 5 |
| 5 | 0000007 | Amplop | 28,37 | 130,15 | 138,94 | 27,27 | 91,54 | 27,27 | 4 |
| 6 | 0000006 | Amplop | 31,62 | 71,58 | 79,82 | 33,66 | 37,90 | 31,62 | 1 |
| 7 | 0000635 | Botol Minum | 110,68 | 14,91 | 4,15 | 112,93 | 52,74 | 4,15 | 3 |
| 8 | 0001404 | Botol Minum | 116,85 | 19,26 | 7,10 | 119,04 | 58,73 | 7,10 | 3 |
| 9 | 0000263 | Buku | 65,03 | 38,28 | 46,40 | 67,33 | 17,00 | 17,00 | 5 |
| 10 | 0000021 | Buku | 0,00 | 102,77 | 111,20 | 6,49 | 65,87 | 0,00 | 1 |
| 11 | 0000613 | Buku Binder | 116,01 | 19,48 | 7,38 | 118,31 | 58,16 | 7,38 | 3 |
| 12 | 0000606 | Buku Binder | 118,22 | 19,32 | 7,54 | 120,33 | 59,50 | 7,54 | 3 |
| 13 | 0001393 | Buku Diary | 108,84 | 13,61 | 5,13 | 111,16 | 50,51 | 5,13 | 3 |
| 14 | 0001395 | Buku Diary | 106,21 | 16,04 | 9,47 | 108,71 | 49,47 | 9,47 | 3 |
| 15 | 0000028 | Buku Kwintansi | 85,27 | 18,41 | 27,50 | 87,33 | 24,92 | 18,41 | 2 |
| 16 | 0000027 | Buku Kwintansi | 98,80 | 6,58 | 13,38 | 100,84 | 38,64 | 6,58 | 2 |
| 17 | 0000039 | Buku Nota | 61,82 | 42,87 | 53,27 | 62,90 | 9,22 | 9,22 | 5 |
| 18 | 0000794 | Buku Nota | 63,92 | 41,59 | 52,25 | 65,08 | 5,84 | 5,84 | 5 |
| 19 | 0001350 | Buku Sampul | 6,49 | 104,49 | 113,26 | 0,00 | 67,33 | 0,00 | 4 |
| 20 | 0000783 | Buku Sampul | 99,22 | 11,17 | 12,45 | 101,50 | 41,49 | 11,17 | 2 |
| 21 | 0000503 | File | 62,14 | 41,57 | 49,12 | 64,38 | 21,19 | 21,19 | 5 |
| 22 | 0001011 | File | 111,84 | 12,18 | 1,01 | 113,88 | 52,28 | 1,01 | 3 |
| 23 | 0000347 | Gunting | 95,20 | 12,53 | 16,77 | 97,52 | 37,10 | 12,53 | 2 |
| 24 | 0000832 | Gunting | 94,96 | 15,27 | 18,75 | 97,49 | 36,98 | 15,27 | 2 |
| 25 | 0000674 | Kalkulator | 119,03 | 20,33 | 8,51 | 121,16 | 60,46 | 8,51 | 3 |
| 26 | 0000972 | Kalkulator | 117,55 | 20,32 | 8,15 | 119,81 | 59,55 | 8,15 | 3 |
| 27 | 0000095 | Kertas | 112,24 | 210,05 | 220,06 | 109,15 | 169,87 | 109,15 | 4 |
| 28 | 0001169 | Kertas | 52,17 | 78,75 | 84,25 | 57,82 | 50,85 | 50,85 | 5 |
| 29 | 0000360 | Kertas HVS | 118,22 | 19,32 | 7,54 | 120,33 | 59,50 | 7,54 | 3 |
| 30 | 0000358 | Kertas HVS | 69,42 | 38,11 | 43,93 | 72,48 | 25,50 | 25,50 | 5 |
| 31 | 0001566 | Kotak Makan | 117,47 | 19,26 | 7,20 | 119,64 | 59,02 | 7,20 | 3 |
| 32 | 0000651 | Kotak Makan | 116,65 | 18,25 | 6,19 | 118,81 | 58,05 | 6,19 | 3 |
| 33 | 0001511 | Kotak Pensil | 109,57 | 13,03 | 3,70 | 111,82 | 50,90 | 3,70 | 3 |
| 34 | 0001512 | Kotak Pensil | 111,20 | 12,20 | 0,00 | 113,26 | 51,96 | 0,00 | 3 |
| 35 | 0000296 | Lakban | 50,84 | 53,03 | 61,46 | 53,47 | 19,79 | 19,79 | 5 |
| 36 | 0001311 | Lakban | 44,38 | 60,34 | 67,76 | 47,59 | 29,79 | 29,79 | 5 |
| 37 | 0000352 | Lem | 77,59 | 29,30 | 35,37 | 80,37 | 25,41 | 25,41 | 5 |
| 38 | 0000717 | Lem | 90,52 | 23,29 | 25,40 | 93,47 | 36,85 | 23,29 | 2 |
| 39 | 0000165 | Map / File | 106,70 | 186,88 | 198,29 | 102,36 | 147,17 | 102,36 | 4 |
| 40 | 0000785 | Map / File | 64,94 | 77,39 | 89,42 | 63,20 | 42,16 | 42,16 | 5 |
| 41 | 0000697 | Misc | 117,41 | 18,30 | 6,58 | 119,50 | 58,54 | 6,58 | 3 |
| 42 | 0000892 | Misc | 102,77 | 0,00 | 12,20 | 104,49 | 41,26 | 0,00 | 2 |
| 43 | 0001578 | Misc Ultah | 89,32 | 14,19 | 25,13 | 91,00 | 27,12 | 14,19 | 2 |
| 44 | 0000711 | Misc Ultah | 98,13 | 7,40 | 13,75 | 100,19 | 38,36 | 7,40 | 2 |
| 45 | 0000083 | Misc ATK | 97,10 | 10,92 | 15,70 | 99,39 | 37,65 | 10,92 | 2 |
| 46 | 0000364 | Misc ATK | 106,82 | 15,56 | 8,72 | 109,29 | 49,65 | 8,72 | 3 |
| 47 | 0001419 | Pena Fancy | 50,56 | 84,98 | 90,03 | 56,61 | 57,61 | 50,56 | 1 |
| 48 | 0001529 | Pena Fancy | 74,28 | 41,61 | 45,79 | 78,05 | 32,08 | 32,08 | 5 |
| 49 | 0000242 | Pena Pulpen | 202,53 | 305,26 | 313,54 | 200,89 | 267,13 | 200,89 | 4 |
| 50 | 0000505 | Pena Pulpen | 53,43 | 147,78 | 155,37 | 56,08 | 110,29 | 53,43 | 1 |
| 51 | 0000782 | Pensil | 47,82 | 63,99 | 70,91 | 52,15 | 33,31 | 33,31 | 5 |
| 52 | 0000619 | Pensil | 46,77 | 60,32 | 67,70 | 50,47 | 28,90 | 28,90 | 5 |
| 53 | 0001486 | Post It / Stick Note / Memo | 110,32 | 12,58 | 2,29 | 112,49 | 51,32 | 2,29 | 3 |
| 54 | 0001489 | Post It / Stick Note / Memo | 111,07 | 12,30 | 1,01 | 113,18 | 51,78 | 1,01 | 3 |
| 55 | 0000715 | Stapler / Staples | 60,08 | 44,40 | 51,46 | 62,61 | 23,43 | 23,43 | 5 |
| 56 | 0000068 | Stapler / Staples | 36,63 | 138,02 | 146,25 | 37,29 | 99,99 | 36,63 | 1 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 57 | 0001142 | Tas | 100,85 | 17,65 | 15,23 | 103,55 | 44,89 | 15,23 | 3 |
| 58 | 0001134 | Tas | 100,82 | 18,31 | 15,95 | 103,57 | 44,91 | 15,95 | 3 |
| 59 | 0001227 | Tip Ex | 75,60 | 37,47 | 42,17 | 79,08 | 29,43 | 29,43 | 5 |
| 60 | 0000330 | Tip Ex | 76,60 | 34,74 | 39,84 | 79,88 | 27,90 | 27,90 | 5 |
| 61 | 0001129 | Tissue | 90,54 | 23,98 | 26,01 | 93,54 | 37,00 | 23,98 | 2 |
| 62 | 0001131 | Tissue | 110,32 | 12,58 | 2,29 | 112,49 | 51,32 | 2,29 | 3 |

4. Tahap Keempat

Untuk setiap cluster membuat pilihan centroid baru

5. Tahap Kelima

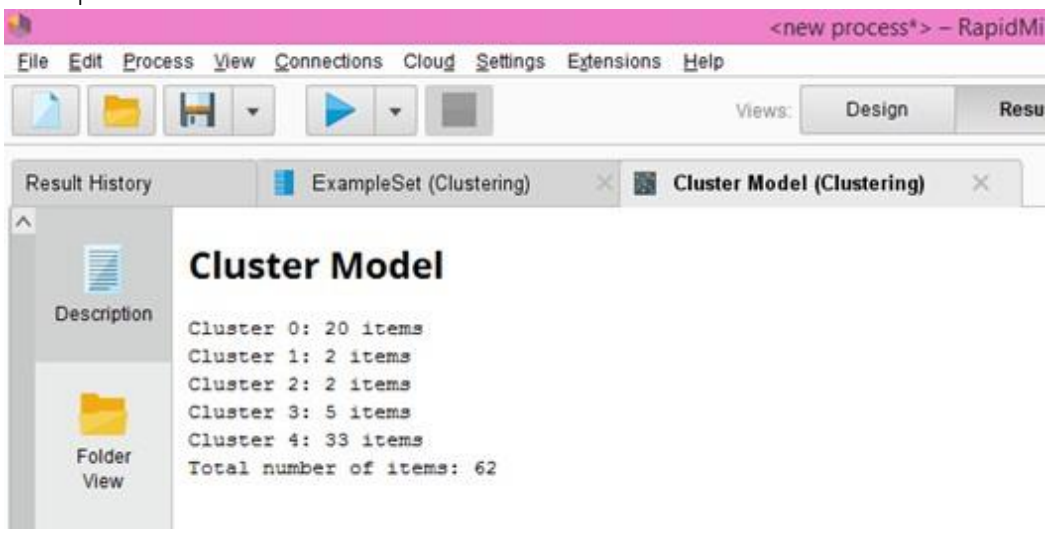
Ulangi langkah 2 dan 3 sampai nilai dari titik centroid tidak berubah lagi, maksudnya dengan melakukan perhitungan kembali (perulangan) seperti langkah ke-2 dan ke-3 yang sudah dilakukan diatas sampai benar-benar menemukan persamaan nilai dari titik centroid yang sudah tidak berubah atau bergeser lagi.

Setelah kelima tahapan selesai dilakukan, maka diperoleh hasil akhirnya yaitu untuk proses perhitungan k-means clustering secara manual yang telah selesai dilakukan pada data sampel yang ada, dan setelah perhitungan centroid yang baru berhenti pada iterasi ke-7, maka dapat diketahui untuk hasil dari 5 cluster yang telah ditentukan sebelumnya, yaitu sebagai berikut :

| No | Kode Item | Jenis | Jumlah Terjual | Jumlah Stok | Volume Penjualan | Rata-rata |
|---|---|---|---|---|---|---|
| 1 | 0000007 | Amplop | 82,00 | 120,00 | 18,00 | 36,67 |
| 2 | 0000021 | Buku | 70,00 | 96,00 | 12,00 | 29,67 |
| 3 | 0001350 | Buku Sampul | 69,00 | 100,00 | 7,00 | 29,33 |
| 4 | 0000505 | Pena Pulpen | 95,00 | 120,00 | 50,00 | 44,17 |
| 5 | 0000068 | Stapler / Staples | 91,00 | 120,00 | 27,00 | 39,67 |
| 6 | Jumlah | 5 | 407,00 | 556,00 | 114,00 | 179,50 |
| 7 | Rata-rata | - | 81,40 | 111,20 | 22,80 | 35,90 |

Berikut Hasil Pengukuran Performance terhadap clustering di atas menggunakan tools RapidMiner :

Description Cluster Model



Data Cluster 0

eSet (62 examples, 4 special attributes, 4 regular attributes) Filter (62 / 62 examples): all

| cluster | Kode Item | Jenis | Jumlah Terj... | Jumlah Stok | Volume Penj... | Rata-Rata |
|---|---|---|---|---|---|---|
| cluster_0 | 0000369 | Alat Gambar | 17 | 60 | 10 | 14.500 |
| cluster_0 | 0000882 | Alat Tulis | 33 | 60 | 19 | 18.667 |
| cluster_0 | 0000006 | Amplop | 52 | 72 | 6 | 21.667 |
| cluster_0 | 0000263 | Buku | 29 | 48 | 9 | 14.333 |
| cluster_0 | 0000039 | Buku Nota | 23 | 60 | 3 | 14.333 |
| cluster_0 | 0000794 | Buku Nota | 20 | 60 | 5 | 14.167 |
| cluster_0 | 0000503 | File | 34 | 48 | 6 | 14.667 |
| cluster_0 | 0001169 | Kertas | 52 | 60 | 45 | 26.167 |
| cluster_0 | 0000358 | Kertas HVS | 32 | 40 | 15 | 14.500 |
| cluster_0 | 0000296 | Lakban | 36 | 60 | 14 | 18.333 |
| cluster_0 | 0001311 | Lakban | 46 | 60 | 14 | 20 |
| cluster_0 | 0000352 | Lem | 24 | 36 | 14 | 12.333 |
| cluster_0 | 0000785 | Map / File | 7 | 100 | 2 | 18.167 |
| cluster_0 | 0001419 | Pena Fancy | 60 | 60 | 46 | 27.667 |
| cluster_0 | 0001529 | Pena Fancy | 31 | 36 | 26 | 15.500 |

Data *Cluster* 1,2,3, dan 4

eSet (62 examples, 4 special attributes, 4 regular attributes) Filter (62 / 62 examples): all

| cluster | Kode Item | Jenis | Jumlah Terj... | Jumlah Stok | Volume Penj... | Rata-Rata |
|---|---|---|---|---|---|---|
| cluster_1 | 0000466 | Alat Tulis | 138 | 240 | 62 | 73.333 |
| cluster_1 | 0000242 | Pena Pulpen | 203 | 240 | 27 | 78.333 |
| cluster_2 | 0000095 | Kertas | 105 | 200 | 14 | 53.167 |
| cluster_2 | 0000165 | Map / File | 52 | 200 | 3 | 42.500 |
| cluster_3 | 0000007 | Amplop | 82 | 120 | 18 | 36.667 |
| cluster_3 | 0000021 | Buku | 70 | 96 | 12 | 29.667 |
| cluster_3 | 0001350 | Buku Sampul | 69 | 100 | 7 | 29.333 |
| cluster_3 | 0000505 | Pena Pulpen | 95 | 120 | 50 | 44.167 |
| cluster_3 | 0000068 | Stapler / Stap... | 91 | 120 | 27 | 39.667 |
| cluster_4 | 0001174 | Alat Gambar | 12 | 24 | 12 | 8 |
| cluster_4 | 0000635 | Botol Minum | 6 | 10 | 4 | 3.333 |
| cluster_4 | 0001404 | Botol Minum | 3 | 5 | 2 | 1.667 |
| cluster_4 | 0000613 | Buku Binder | 4 | 5 | 4 | 2.167 |
| cluster_4 | 0000606 | Buku Binder | 1 | 5 | 1 | 1.167 |
| cluster_4 | 0001393 | Buku Diary | 6 | 12 | 6 | 4 |

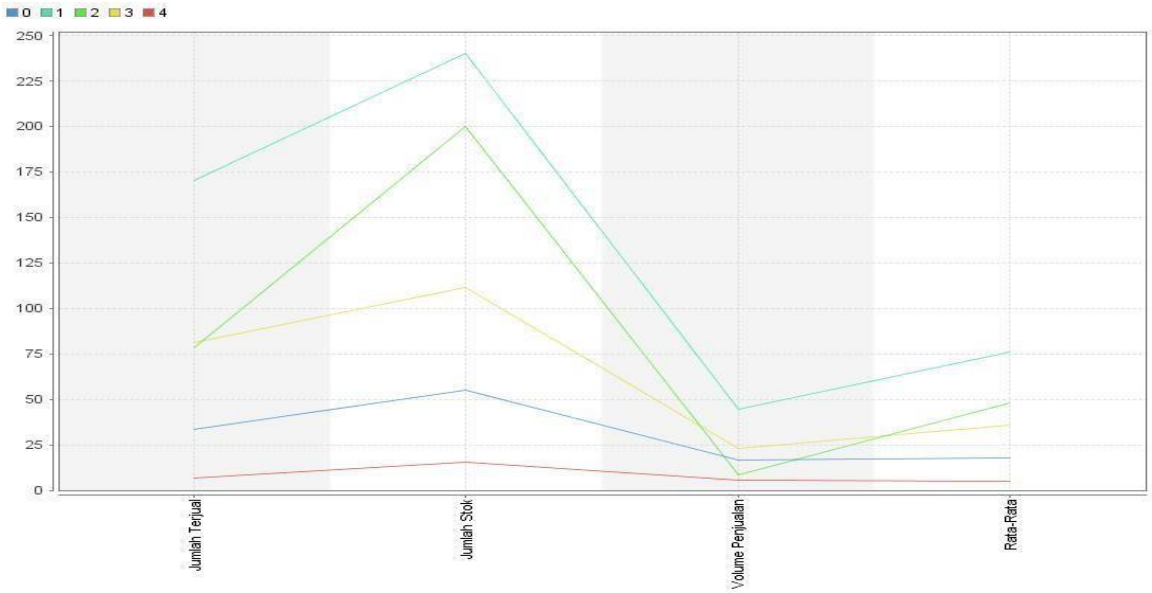
Hasil Centroid Akhir

Description

Folder View

| Attribute | cluster_0 | cluster_1 | cluster_2 | cluster_3 | cluster_4 |
|---|---|---|---|---|---|
| Jumlah Terjual | 33.700 | 170.500 | 78.500 | 81.400 | 6.909 |
| Jumlah Stok | 55 | 240 | 200 | 111.200 | 15.242 |
| Volume Penjualan | 16.750 | 44.500 | 8.500 | 22.800 | 5.788 |
| Rata-Rata | 17.575 | 75.833 | 47.833 | 35.900 | 4.657 |

Berikut hasil Grafik menggunakan Rapid Miner :



# TUGAS 05

# ADVANCDED DATABASE

# KELAS MTI 23 A


Di Susun Oleh :

Ari Hardiyantoro Susanto

NIM : 202420015

Dosen Pengasuh :

Tri Basuki Kurniawan , S.Kom., M.Eng. Ph.D

**Program Pasca Sarjana**

**Universitas Bina Darma Palembang**

2020/2021

Soal :

Pelajari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format ms word dan dilengkapi dengan gambar dan keterangan yang jelas untuk tiap langkah-langkahnya. Sertkan sumber sumber rujukan yang anda gunakan. Kumpulkan sebelum waktu batas akhir pengumpulan tugas.

Jawaban :

Menerapkan K-Means Clustering Pada Data Pengemudi Go-Track

Sebagai contoh, saya akan menunjukkan cara kerja algoritma K-means dengan dataset sampel dari data pengemudi (driver) pada aplikasi Go-Track. Disini saya hanya akan menggunakan dua fitur driver pada aplikasi Go-Tracks untuk dijadikan variabel : jarak rata-rata per hari dan kecepatan dalam mengemudi per-hari . Data ini bersumber dari http://archive.ics.uci.edu/ml/datasets/GPS+Trajectories

Secara umum, algoritma ini dapat digunakan untuk sejumlah variabel , asalkan jumlah sampel data jauh lebih besar dari jumlah variabel. Mari kita lihat langkah-langkah tentang bagaimana algoritma K-means Clustering bekerja menggunakan bahasa pemrograman Python.

## **L**angkah 1 : Import Libraries

Pada kasus ini , Kides akan menggunakan library Scikit-learn dan beberapa data acak untuk mengilustrasikan penjelasan sederhana pengelompokan K-means.

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as pltfrom sklearn.cluster

import KMeans
from sklearn.preprocessing import MinMaxScaler
```

Keterangan

>> Panda untuk membaca dan menulis spreadsheet

>> Numpy untuk melakukan perhitungan yang efisien

>>Matplotlib untuk visualisasi data

## Langkah 2 : Menginput Data

Data yang digunakan yakni dataset sampel data pengemudi (driver) pada aplikasi Go-Track. Dibawah ini adalah script untuk :

>>menginput data pada (baris 1)

>>membaca data pada (baris 2)

```
--- Membaca Data ---

driver=pd.read_csv("go_track_tracks.csv")
driver.head()
```

| | id | id_android | speed | time | distance | rating | rating_bus | rating_weather | car_or_bus | linha |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 19.210586 | 0.138049 | 2.652 | 3 | 0 | 0 | 1 | NaN |
| **1** | 2 | 0 | 30.848229 | 0.171485 | 5.290 | 3 | 0 | 0 | 1 | NaN |
| **2** | 3 | 1 | 13.560101 | 0.067699 | 0.918 | 3 | 0 | 0 | 2 | NaN |
| **3** | 4 | 1 | 19.766679 | 0.389544 | 7.700 | 3 | 0 | 0 | 2 | NaN |
| **4** | 8 | 0 | 25.807401 | 0.154801 | 3.995 | 2 | 0 | 0 | 1 | NaN |

Dari data yang di miliki terdapat 10 variabel pada data set yang ada . Ada beberapa variabel yang tidak dibutuhkan sehingga harus dihapuskan

## Langkah 3 : Menghilangkan kolom yang tidak diperlukan

Untuk contoh ini, saya telah menghilangkan beberapa kolom yang tidak diperlukan . Sehingga hanya menyisahkan 4 kolom yakni id , **id_android,speed dan distance** seperti ditunjukkan pada gambar dibawah :

```
--- Menghilangkan Kolom Yang Tidak Perlu ---

driver=driver.drop(["linha","car_or_bus","rating_weather",
"rating_bus","rating","time"],axis=1)
driver.head()
```

| | id | id_android | speed | distance |
|---|---|---|---|---|
| 0 | 1 | 0 | 19.210586 | 2.652 |
| 1 | 2 | 0 | 30.848229 | 5.290 |
| 2 | 3 | 1 | 13.560101 | 0.918 |
| 3 | 4 | 1 | 19.766679 | 7.700 |
| 4 | 8 | 0 | 25.807401 | 3.995 |

Pada gambar diatas menunjukkan dataset untuk 38092 driver . Selanjutnya menentukan variabel yang diklusterkan , disini menggunakan variabel jarak pada sumbu X dan variabel kecepatan pada sumbu Y .

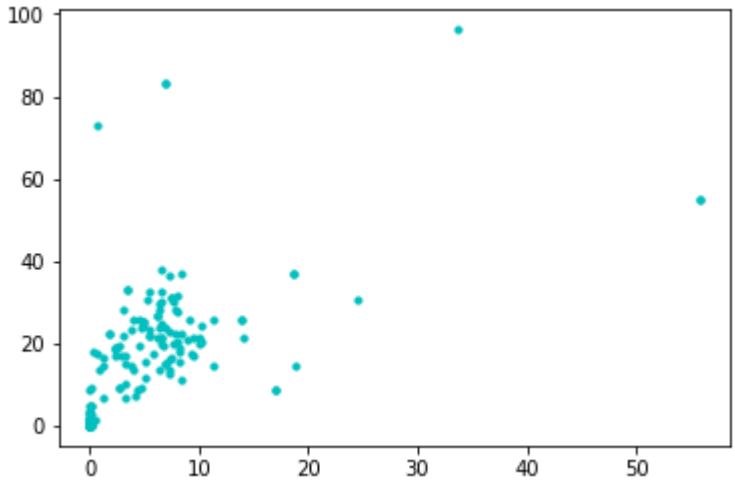
```
-- Menentukan variabel yang akan di klusterkan ---

driver_x=driver.iloc[:,1:3]
driver_x.head()
```

| | speed | distance |
|---|---|---|
| 0 | 19.210586 | 2.652 |
| 1 | 30.848229 | 5.290 |
| 2 | 13.560101 | 0.918 |
| 3 | 19.766679 | 7.700 |
| 4 | 25.807401 | 3.995 |

Dataset diatas divisualisasikan persebaran datanya sebagai berikut :

```
--- Memvisualkan persebaran data ---
plt.scatter(driver.distance, driver.speed, s =10, c = "c", marker =
"o", alpha = 1)
plt.show()
```




## **L**angkah 4 : Menentukan nilai K

Sebelum menetukan nilai K , saya harus mengubah variabel data frame menjadi array terlebih dahulu :

```
--- Mengubah Variabel Data Frame Menjadi Array ---
x_array=np.array(driver_x)
print(x_array)
```

```
[[1.92105856e+01 2.65200000e+00]
 [3.08482291e+01 5.29000000e+00]
 [1.35601009e+01 9.18000000e-01]
 [1.97666790e+01 7.70000000e+00]
 [2.58074009e+01 3.99500000e+00]
 [1.34691332e+00 9.00000000e-03]
 [3.68507874e+01 8.40200000e+00]
 [1.74051313e+01 6.75000000e-01]
 [1.53954361e+01 8.11100000e+00]
 [8.90272944e+00 2.70000000e-02]
 [1.50413480e+01 3.27700000e+00]
 [1.44400981e+01 3.87200000e+00]
 [1.63567325e+01 1.26000000e+00]
 [1.75427999e+01 5.85700000e+00]
 [9.45181557e+00 2.61600000e+00]
 [9.45181557e+00 2.61600000e+00]
 [1.62635039e+01 7.33400000e+00]
 [2.12235944e+01 6.14900000e+00]
 [1.94236545e+01 4.59500000e+00]
```

Kemudian saya harus menstandarkan kembali ukuran variabel . Agar data dapat kembali seperti jenis data diawal sebelum di array kan

```
--- Menstandarkan Ukuran Variabel ---
scaler = MinMaxScaler()
x_scaled = scaler.fit_transform(x_array)
x_scaled
```

```
array([[1.99600362e-01, 4.75353691e-02],
       [3.20578504e-01, 9.48376338e-02],
       [1.40861225e-01, 1.64428267e-02],
       [2.05381184e-01, 1.38051606e-01],
       [2.68176999e-01, 7.16168481e-02],
       [1.39000630e-02, 1.43448869e-04],
       [3.82977592e-01, 1.50639244e-01],
       [1.80831914e-01, 1.20855673e-02],
       [1.59940297e-01, 1.45421291e-01],
       [9.24459124e-02, 4.66208826e-04],
       [1.56259404e-01, 5.87423120e-02],
       [1.50009162e-01, 6.94113217e-02],
       [1.69933373e-01, 2.25752658e-02],
       [1.82263037e-01, 1.05004572e-01],
       [9.81538910e-02, 4.68898492e-02],
       [9.81538910e-02, 4.68898492e-02],
       [1.68964223e-01, 1.31488820e-01],
       [2.20526426e-01, 1.10240456e-01],
       [2.01815302e-01, 8.23755133e-02],
```

Dalam hal tersebut, nilai K (n_clusters) atau nilai arbitrer bebas ditentukan. Untuk ini , saya akan mengkonfigurasi dan menentukan nilai K sebesar 3 cluster. Selain itu saya juga menetukan kluster dari data yang telah di standarkan .

```
--- Menentukan dan mengkonfigurasi fungsi kmeans ---
kmeans = KMeans(n_clusters = 3, random_state=123)---

Menentukan kluster dari data ---
kmeans.fit(x_scaled)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=2, n_init=10, n_jobs=1, precompute_distances='auto',
    random_state=123, tol=0.0001, verbose=0)
```

## **L**angkah 5 : Memvisualisasikan Kluster

Untuk memvisualisasikan kluster, harus menampilkan centroid terlebih dahulu seperti gambar dibawah ini :

```
--- Menampilkan pusat cluster ---
print(kmeans.cluster_centers_)
```

```
[[0.23246003 0.12863014]
 [0.02863662 0.00851813]
 [0.77224472 0.47782221]]
```

Selanjutnya menampilkan hasil cluster dan Menambahkan kolom data frame driver .

```
--- Menampilkan Hasil Kluster ---
print(kmeans.labels_)
-- Menambahkan Kolom "kluster" Dalam Data Frame Driver --
driver["kluster"] = kmeans.labels_
```

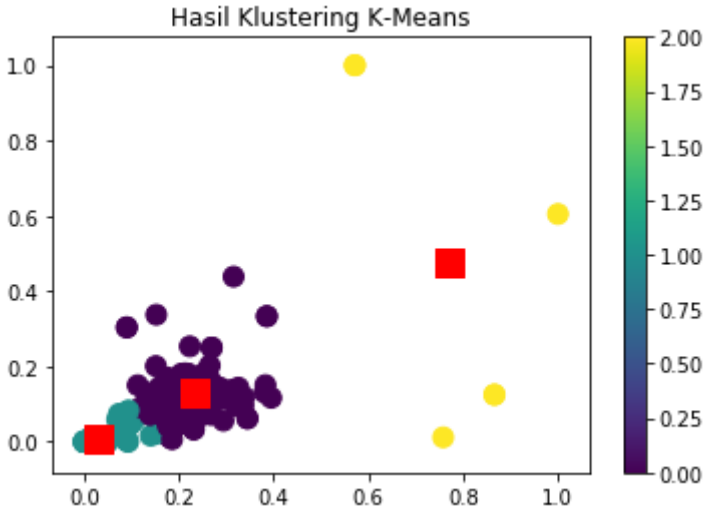
```
[0 0 1 0 0 1 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 1 2 2 0 1 1 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 2 1 2 2 2 1 1 1 1 1 1 1 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 1 1
 0 0 0 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0
 0 1 0 1 1 1 1 1 1 0 1 0 1 1 0 0 1 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1 1 1 1 0 0
 1 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1]
```

Kemudian yang terakhir memvisualisasikan hasil cluster seperti gambar dibawah ini :

```
--- Memvisualkan hasil kluster ---
output = plt.scatter(x_scaled[:,0], x_scaled[:,1], s =
100, c = driver.kluster, marker = "o", alpha = 1, )

centers = kmeans.cluster_centers_
plt.scatter(centers[:,0], centers[:,1], c='red', s=200,
alpha=1 , marker="s");

plt.title("Hasil Klustering K-Means")
plt.colorbar (output)
plt.show()
```




Dari gambar diatas dapat kita lihat bahwa hasil dari data pengemudi Go-Track telah ter cluster menjadi 3.

Dalam mengukur *performance* atau kualitas dari cluster yang dihasilkan oleh sebuah algoritma *Clustering* dapat menggunakan beberapa metode di antaranya, metode *entropy*, *purity*, *recall* dan *F-Score.*

Pada contoh berikut ini akan digunakan 2 (dua) metode dalam menguji *performance Clustering ,* yaitu metode *entropy* dan *purity* untuk menguji kualitas Pengelompokan dokumen (*Document Clustering*) dengan menggunakan metode **K-Means** dan **K-Medians**

Untuk setiap cluster, entropy dapat di ukur sebagai berikut :

$$Entropy\ (D_i) = -\sum_{j=1}^{k} \Pr_i(c_j) \log_2 \Pr_i(c_j),$$

Dimana, $\Pr_i(c_j)$ adalah rasio data dengan kategori $c_j$ dalam cluster $D_i$. Total *entropy* dari seluruh cluster dihitung dengan :

$$Entropy_{total}(D_i) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} \text{ x } entropy(D_i),$$

dengan $D_i$ adalah jumlah data *cluster i* dan $D$ adalah total jumlah data.

Nilai *purity* dan total *purity* setiap *cluster* dapat diukur sebagai berikut.

$$Purity\ (D_i) = \max(\Pr_i(c_j)) \text{ dan}$$

$$Purity_{total}(D_i) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} \text{ x } purity(D_i),$$

dimana, $max(Pr_i(c_j))$ adalah rasio kategori tertinggi dalam *cluster* $D_i$ [8].

a. **Pengujian dengan data IRIS**

   Data IRIS merupakan data multivariat yang diperkenalkan oleh Sir Ronald Aylmer Fisher (1936) sebagai contoh analisis diskriminan. Data ini juga disebut Anderson IRIS dataset karena Edgar Anderson mengumpulkan data ini untuk mengukur variasi geografis bunga IRIS di Semenanjung Gaspé. Data IRIS berasal dari 3 (tiga) jenis bunga IRIS yaitu setosa, versicolor, dan virginica. Masing-masing jenis bunga terdiri dari 50 sampel. Ada empat fiturn yang diukur dari masing-masing

   Data yang pertama diuji dalam penelitian ini adalah data IRIS. Pengujian tehadap data ini dilakukan sebanyak 6 (enam) kali percobaan. Seperti yang telah dijelaskan pada bab sebelumnya, ada 3 (tiga) jenis data IRIS yang digunakan yaitu data IRIS All (data IRIS asli) dan 2 jenis data IRIS yang diperkecil dimensinya dengan metode *SVD* (IRIS 3D dan IRIS 2D).

Tabel 1 menunjukkan perbandingan kualitas *cluster* yang didapatkan dengan data IRIS untuk kedua metode *clustering* yang digunakan. Pada baris pertama memperlihatkan kualitas *cluster* untuk data IRIS All, sedangkan baris kedua dan ketiga adalah kualitas *cluster* untuk data IRIS 3D dan 2D.

Tabel 1 Hasil rata-rata perbandingan kualitas *cluster* dengan data IRIS untuk K=3

| No | DATA | Metode K-means | | | | Metode K-medians | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Entropy Total** | **Purity Total** | **Jumlah Iterasi** | **Waktu (dtk)** | **Entropy Total** | **Purity Total** | **Jumlah Iterasi** | **Waktu (dtk)** |
| 1 | IRIS All (Rank=4) | 0.453 | 0.858 | 6.143 | 0.054 | 0.429 | 0.866 | 8.714 | 0.079 |
| 2 | IRIS 3D (Rank=3) | 0.457 | 0.857 | 6.857 | 0.048 | 0.424 | 0.861 | 6.429 | 0.046 |
| 3 | IRIS 2D (Rank=2) | 0.464 | 0.855 | 6.571 | 0.032 | 0.436 | 0.855 | 5.714 | 0.030 |
| Rata-rata | | 0.458 | 0.857 | 6.524 | 0.045 | 0.414 | 0.872 | 6.952 | 0.052 |

**b. Pengujian dengan data Reuters-21578**

Reuters-21578 (V 1.3) dataset adalah data dokumen teks yang sudah dikategorikan dan dipublikasikan oleh David D. Lewis pada tanggal 14 Mei 2004 (http://www.research.att.com/~lewis). Reuters-21578 dataset memiliki enam kategori asli yaitu exchanges, orgs, people, places, topics, dan companies, tetapi karena jumlah artikel untuk kategori companies hanya berjumlah 6 dan tidak memiliki isi maka kategori ini tidak disertakan dalam penelitian ini. Dokumendokumen dalam data Reuters disusun dalam format seperti yang terlihat pada Gambar. 2.

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET"
OLDID="5544" NEWID="1">
<DATE>26-FEB-1987 15:01:01.79</DATE>
<TOPICS><D>cocoa</D></TOPICS>
<PLACES><D>el-salvador</D><D>usa</D><D>uruguay</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;C T
&#22;&#22;&#1;f0704&#31;reute
u f BC-BAHIA-COCOA-REVIEW   02-26 0105</UNKNOWN>
<TEXT>&#2;
<TITLE>BAHIA COCOA REVIEW</TITLE>
<DATELINE>    SALVADOR, Feb 26 - </DATELINE><BODY>Showers continued
throughout the week in
the Bahia cocoa zone, alleviating the drought since early
January and improving prospects for the coming temporao,
although normal humidity levels have not been restored,
Comissaria Smith said in its weekly review…
    Reuter
&#3;</BODY></TEXT>
</REUTERS>
```

Pengujian yang kedua dilakukan dengan data Reuters-21578. Sama seperti data IRIS, pengujian dengan data ini juga dilakukan sebanyak 6 kali percobaan untuk tiap-tiap jenis data, kecuali pada data Reuters All, hanya dilakukan sebannyak 2 kali percobaan. Tabel 2 menunjukkan kualitas *cluster* yang diperoleh dari setiap metode *clustering* yang digunakan. Baris pertama memperlihatkan kualitas *cluster* ketika data Reuters All (32675 kata dari 8738 dokumen), sedangkan baris kedua, ketiga, dan keempat menunjukkan hasil ketika data diperkecil dimensinya (D600, D400, dan D200).

Kualitas *cluster* yang diperoleh dari pengujian dengan data Reuters-21578 untuk kedua jenis metode *clustering* tidak begitu bagus. Hal ini ditunjukkan dengan tingginya nilai *entropy* dan rendahnya nilai *purity* yang diperoleh dari kedua jenis metode *clustering* untuk tiap-tiap data *Reuters* yang digunakan. Salah satu penyebab tidak bagusnya nilai *entropy* dan *purity* yang diperoleh dari pengujian dengan data *Reuters* diduga karena banyaknya dokumen yang terdapat dalam data tersebut memiliki lebih dari satu jenis kategori.

Tabel 2 Hasil rata-rata perbandingan kualitas *cluster* dengan data Reuters-21578 untuk K=5

| No | Data | Metode K-means | | | | Metode K-medians | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Entropy Total | Purity Total | Jumlah Iterasi | Waktu (dtk) | Entropy Total | Purity Total | Jumlah Iterasi | Waktu (dtk) |
| 1 | Data Reuters All (Rank=32675) | 1.078 | 0.527 | 28.00 | 2100 | 1.073 | 0.584 | 62.50 | 6144.7 |
| 2 | Data Reuters D600 (Rank=600) | 1.107 | 0.511 | 54.14 | 70 | 1.108 | 0.512 | 101.00 | 197.0 |
| 3 | Data Reuters D400 (Rank=400) | 1.107 | 0.513 | 56.28 | 48 | 1.108 | 0.512 | 101.00 | 108 |
| 4 | Data Reuters D200 (Rank=200) | 1.107 | 0.511 | 49.71 | 21 | 1.105 | 0.511 | 101.00 | 61 |

**Kesimpulan :**

Berdasarkan hasil penelitian yang dilakukan beserta pembahasan pada bab sebelumnya maka dapat disimpulkan bahwa:

1. Pengujian dengan data IRIS jelas memperlihatkan bahwa kualitas *cluster* yang diperoleh dari metode K-medians lebih baik dibandingkan dengan metode Kmeans.
2. Berdasarkan pengujian pada data Reuters All, waktu yang dibutuhkan metode K-means jauh labih cepat dibandingkan dengan metode K-medians. K-means hanya membutuhkan waktu sekitar 35 jam sedangkan metode Kmedians membutuhkan waktu sebanyak 102,411 jam (4 hari 6 jam).
3. Penggunaan metode *SVD* dalam kasus pengelompokan dokumen adalah solusi yang baik untuk memperkecil waktu dalam melakukan *clustering.*
4. Data Reuters-21578 V.1 kurang cocok digunakan untuk kasus *clustering* dokumen karena terdapat dokumen-dokumen yang tidak memiliki kategori tunggal.
5. Kualitas *cluster* yang diperoleh dari metode clustering juga dipengaruhi oleh kualitas dokumen.

**Referensi :**

***https://jurnal.ar-raniry.ac.id/index.php/elkawnie/article/download/536/2699***

Nama : Bhijanta Wyasa WM
NIM : 202420019
Kelas : MTI23

# Tugas 05

Pelari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format ms word dan dilengkapi dengan gambar dan keterangan yang jelas untuk tiap langkah-langkahnya. Sertkan sumber sumber rujukan yang anda gunakan. Kumpulkan sebelum waktu batas akhir pengumpulan tugas.

**Jawaban :**

Pada tugas ini saya coba mengambil sample untuk mengukur performance Clustering dengan menggunakan K-Means, pengukuran performance clustering saat ini menggunakan clusterisasi sampai didapatkan nilai stabil dan dengan rapidminer saya coba untuk bandingkan hasilnya dengan sumber data yang sama, adapun datanya sebagai berikut :

**Data Komsumsi Listrik bulan Mei 2020 di Prov Lampung :**

| | Kabupaten/Kota | Latitude | Longitude | Mei 2020 | | |
|---|---|---|---|---|---|---|
| | | | | Rumah Tangga | Industri | Bisnis |
| | Kabupaten Tulang Bawang Barat | -4.52569 | 105.07912 | 8647851 | 1424419 | 771600 |
| | Kabupaten Way Kanan | -4.49636 | 104.56552 | 8922355 | 904997 | 580164 |
| | Kabupaten Lampung Barat | -5.10952 | 104.1466 | 7870512 | 908299 | 1508881 |
| | Kabupaten Lampung Selatan | -5.56226 | 105.54743 | 39811088 | 28596132 | 6923204 |
| | Kabupaten Lampung Tengah | -4.8008 | 105.31311 | 43819603 | 17071293 | 5099183 |
| | Kabupaten Lampung Timur | -5.11349 | 105.68817 | 17150724 | 4651063 | 3006046 |
| | Kabupaten Lampung Utara | -4.81339 | 104.75209 | 21682207 | 3629202 | 2188612 |
| | Kabupaten Pringsewu | -5.33311 | 104.98561 | 17504659 | 701278 | 1618426 |
| | Kabupaten Tanggamus | -5.30274 | 104.56552 | 14831519 | 2175713 | 1911240 |
| | Kabupaten Tulang Bawang | -4.31765 | 105.50054 | 13723285 | 1574286 | 1730416 |
| | Kota Bandar Lampung | -5.39713 | 105.26678 | 60305690 | 7951584 | 18855425 |
| | Kota Metro | -5.11783 | 105.30726 | 17774258 | 753564 | 2565855 |

Dari data diatas saya coba untuk mengukur bagaimana performance dari clustering yang menggunakan K-Means dengan membandingkan 2 cara yaitu dengan Coding PhP dan Rapidminer.

Nama : Bhijanta Wyasa WM
NIM : 202420019
Kelas : MTI23

- **Berbasiskan Web**

## Stage Awal Data

Proses Iterasi Selanjutnya

| | Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rumah Tangga | Industri | Bisnis | 4035984 | 396909 | 14840900 | 396909 | 4035984 | 14840900 | 14840900 | 396909 | 4035984 | | | |
| | Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 14841505 | | | 16517975 | | | 7075721 | | | 0 | 0 | 1 |
| | Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 15083214 | | | 16907244 | | | 6872410 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 13881925 | | | 15600637 | | | 7431963 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 46235745 | | | 47110171 | | | 37776216 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 44222985 | | | 46371873 | | | 33450421 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 18170250 | | | 20521510 | | | 4949129 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 21952614 | | | 24765091 | | | 7788708 | | | 0 | 0 | 1 |
| | Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 18876749 | | | 21877595 | | | 3610102 | | | 0 | 0 | 1 |
| | Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 16937645 | | | 19467785 | | | 2771059 | | | 0 | 0 | 1 |
| | Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 16343647 | | | 18855689 | | | 2819738 | | | 0 | 0 | 1 |
| | Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 56916336 | | | 60170677 | | | 48412148 | | | 0 | 0 | 1 |
| | Kota Metro | 17774258 | 753564 | 2565855 | 18426722 | | | 21527268 | | | 3300465 | | | 0 | 0 | 1 |

## Iterasi 1

**Iterasi Ke-1**

| | Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rumah Tangga | Industri | Bisnis | 0 | 0 | 0 | 0 | 0 | 0 | 22670312 | 5861819 | 3896587 | | | |
| | Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 8798276 | | | 8798276 | | | 15036139 | | | 1 | 1 | 0 |
| | Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 8986880 | | | 8986880 | | | 14985829 | | | 1 | 1 | 0 |
| | Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 8065152 | | | 8065152 | | | 15788369 | | | 1 | 1 | 0 |
| | Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 49503456 | | | 49503456 | | | 28632421 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 47303153 | | | 47303153 | | | 23966456 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 18022653 | | | 18022653 | | | 5720564 | | | 0 | 0 | 1 |
| | Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 22092515 | | | 22092515 | | | 2979616 | | | 0 | 0 | 1 |
| | Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 17593299 | | | 17593299 | | | 7648868 | | | 0 | 0 | 1 |
| | Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 15111602 | | | 15111602 | | | 8886825 | | | 0 | 0 | 1 |
| | Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 13921252 | | | 13921252 | | | 10155025 | | | 0 | 0 | 1 |
| | Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 63683051 | | | 63683051 | | | 40553120 | | | 0 | 0 | 1 |
| | Kota Metro | 17774258 | 753564 | 2565855 | 17974307 | | | 17974307 | | | 7199754 | | | 0 | 0 | 1 |

Nama : Bhijanta Wyasa WM
NIM : 202420019
Kelas : MTI23

## Iterasi 2

**Iterasi Ke-2**

| Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rumah Tangga | Industri | Bisnis | 8480239 | 1079238 | 953548 | 8480239 | 1079238 | 953548 | 27400337 | 7456012 | 4877600 | | | |
| Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 424674 | | | 424674 | | | 20122005 | | | 1 | 1 | 0 |
| Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 604352 | | | 604352 | | | 20070365 | | | 1 | 1 | 0 |
| Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 842248 | | | 842248 | | | 20871868 | | | 1 | 1 | 0 |
| Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 42124082 | | | 42124082 | | | 24599103 | | | 0 | 0 | 1 |
| Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 39010289 | | | 39010289 | | | 19028794 | | | 0 | 0 | 1 |
| Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 9599374 | | | 9599374 | | | 10790042 | | | 1 | 1 | 0 |
| Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 13502579 | | | 13502579 | | | 7387296 | | | 0 | 0 | 1 |
| Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 9056769 | | | 9056769 | | | 12416646 | | | 1 | 1 | 0 |
| Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 6515994 | | | 6515994 | | | 13951919 | | | 1 | 1 | 0 |
| Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 5323356 | | | 5323356 | | | 15217136 | | | 1 | 1 | 0 |
| Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 55259240 | | | 55259240 | | | 35754544 | | | 0 | 0 | 1 |
| Kota Metro | 17774258 | 753564 | 2565855 | 9438452 | | | 9438452 | | | 11955265 | | | 1 | 1 | 0 |

## Iterasi 3

**Iterasi Ke-3**

| Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rumah Tangga | Industri | Bisnis | 13303145 | 1636702 | 1711578 | 13303145 | 1636702 | 1711578 | 41404647 | 14312052 | 8266606 | | | |
| Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 4753986 | | | 4753986 | | | 35989913 | | | 1 | 1 | 0 |
| Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 4583319 | | | 4583319 | | | 35971236 | | | 1 | 1 | 0 |
| Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 5484994 | | | 5484994 | | | 36740518 | | | 1 | 1 | 0 |
| Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 38165991 | | | 38165991 | | | 14435341 | | | 0 | 0 | 1 |
| Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 34365049 | | | 34365049 | | | 4845408 | | | 0 | 0 | 1 |
| Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 5056271 | | | 5056271 | | | 26631954 | | | 1 | 1 | 0 |
| Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 8625908 | | | 8625908 | | | 23238759 | | | 1 | 1 | 0 |
| Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 4305393 | | | 4305393 | | | 28295952 | | | 1 | 1 | 0 |
| Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 1632887 | | | 1632887 | | | 29896697 | | | 1 | 1 | 0 |
| Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 425168 | | | 425168 | | | 31164567 | | | 1 | 1 | 0 |
| Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 50428448 | | | 50428448 | | | 22579372 | | | 0 | 0 | 1 |
| Kota Metro | 17774258 | 753564 | 2565855 | 4636870 | | | 4636870 | | | 27833908 | | | 1 | 1 | 0 |

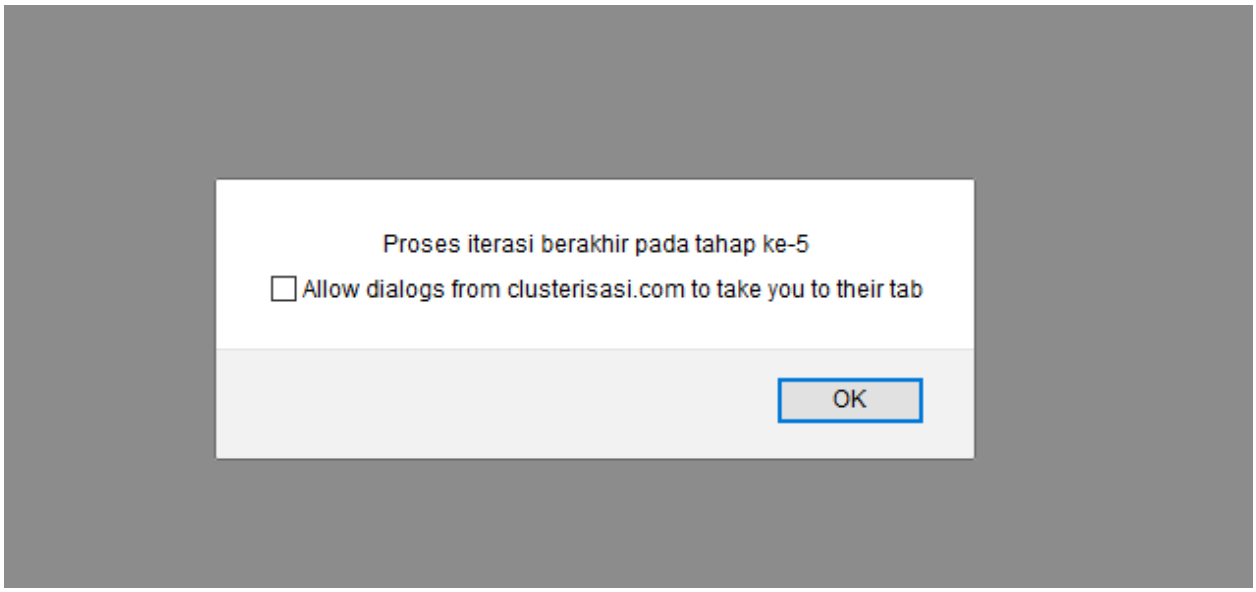
Nama : Bhijanta Wyasa WM
NIM : 202420019
Kelas : MTI23

## Iterasi 4

**Iterasi Ke-4**

| Kabupaten/Kota | Mei 2020 | | | Centroid 1 | | | Centroid 2 | | | Centroid 3 | | | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rumah Tangga | Industri | Bisnis | 14234152 | 1858091 | 1764582 | 14234152 | 1858091 | 1764582 | 47978793 | 17873003 | 10292604 | | | |
| Kabupaten Tulang Bawang Barat | 8647851 | 1424419 | 771600 | 5690417 | | | 5690417 | | | 43682130 | | | 1 | 1 | 0 |
| Kabupaten Way Kanan | 8922355 | 904997 | 580164 | 5525072 | | | 5525072 | | | 43676654 | | | 1 | 1 | 0 |
| Kabupaten Lampung Barat | 7870512 | 908299 | 1508881 | 6439208 | | | 6439208 | | | 44425547 | | | 1 | 1 | 0 |
| Kabupaten Lampung Selatan | 39811088 | 28596132 | 6923204 | 37359253 | | | 37359253 | | | 13894235 | | | 0 | 0 | 1 |
| Kabupaten Lampung Tengah | 43819603 | 17071293 | 5099183 | 33434412 | | | 33434412 | | | 6701733 | | | 0 | 0 | 1 |
| Kabupaten Lampung Timur | 17150724 | 4651063 | 3006046 | 4224726 | | | 4224726 | | | 34326134 | | | 1 | 1 | 0 |
| Kabupaten Lampung Utara | 21682207 | 3629202 | 2188612 | 7667473 | | | 7667473 | | | 30985012 | | | 1 | 1 | 0 |
| Kabupaten Pringsewu | 17504659 | 701278 | 1618426 | 3472145 | | | 3472145 | | | 36038623 | | | 1 | 1 | 0 |
| Kabupaten Tanggamus | 14831519 | 2175713 | 1911240 | 692271 | | | 692271 | | | 37621722 | | | 1 | 1 | 0 |
| Kabupaten Tulang Bawang | 13723285 | 1574286 | 1730416 | 585404 | | | 585404 | | | 38889575 | | | 1 | 1 | 0 |
| Kota Bandar Lampung | 60305690 | 7951584 | 18855425 | 49515797 | | | 49515797 | | | 17991910 | | | 0 | 0 | 1 |
| Kota Metro | 17774258 | 753564 | 2565855 | 3793990 | | | 3793990 | | | 35568129 | | | 1 | 1 | 0 |

## Iterasi 5




Pada saat iterasi ke 5 data sudah didapatkan sampai dengan semua nilainya stabil atau sama, sehingga didapat nilai cluster yang sesuai sebagai berikut :

Iterasi ke-1

| C1 | C2 | C3 |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |

Nama : Bhijanta Wyasa WM
NIM : 202420019
Kelas : MTI23

Iterasi ke-2

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |

Iterasi ke-3

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |

Nama : Bhijanta Wyasa WM
NIM : 202420019
Kelas : MTI23

Iterasi ke-4

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |

Iterasi ke-5

| C1 | C2 | C3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |

Nama : Bhijanta Wyasa WM
NIM : 202420019
Kelas : MTI23

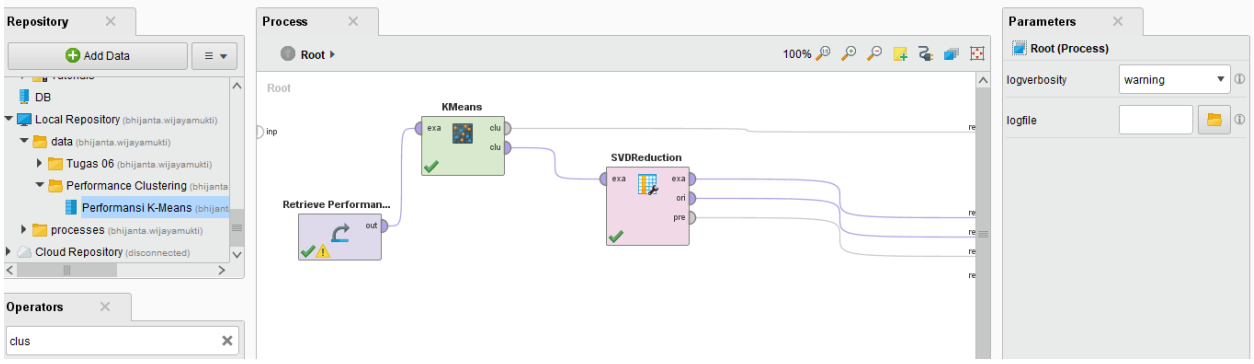
**Dari Hasil Iterasi ke 5 didaptkan kesimpulan sebagai berikut :**

| | Kabupaten/Kota | Bulan | Zona |
|---|---|---|---|
| | Kabupaten Tulang Bawang Barat | Mei | Kuning |
| | Kabupaten Way Kanan | Mei | Kuning |
| | Kabupaten Lampung Barat | Mei | Kuning |
| | Kabupaten Lampung Selatan | Mei | Hijau |
| | Kabupaten Lampung Tengah | Mei | Hijau |
| | Kabupaten Lampung Timur | Mei | Kuning |
| | Kabupaten Lampung Utara | Mei | Kuning |
| | Kabupaten Pringsewu | Mei | Kuning |
| | Kabupaten Tanggamus | Mei | Kuning |
| | Kabupaten Tulang Bawang | Mei | Kuning |
| | Kota Bandar Lampung | Mei | Hijau |
| | Kota Metro | Mei | Kuning |

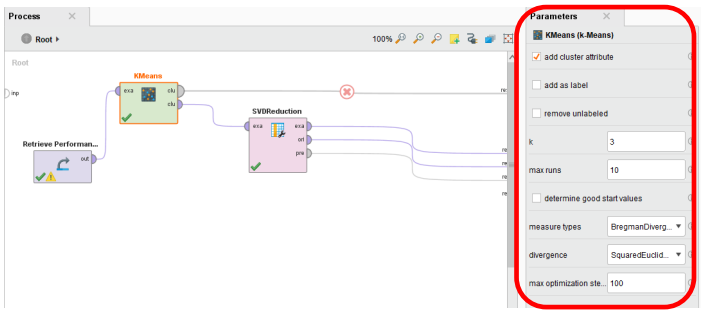
- **Berbasiskan Rapidminer**

Dengan data yang sama seperti diatas saya coba mengukur bagaimana performance dari Clustering K-Means menggunakan tool rapidminer, adapun hasilnya sebagai berikut :

**Desain Mapping Rapidminer**

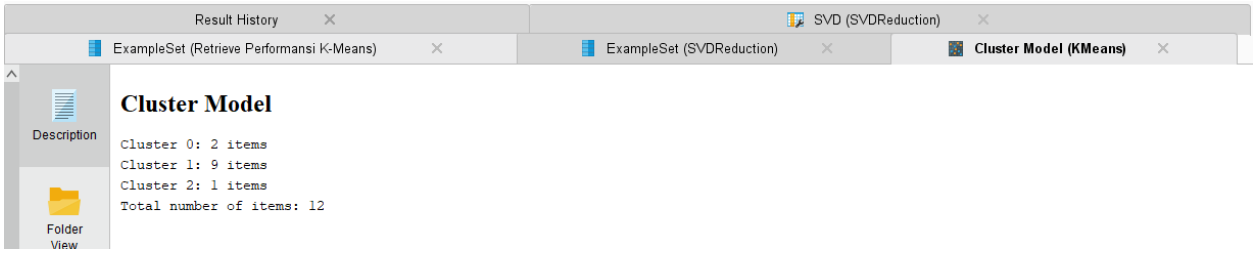


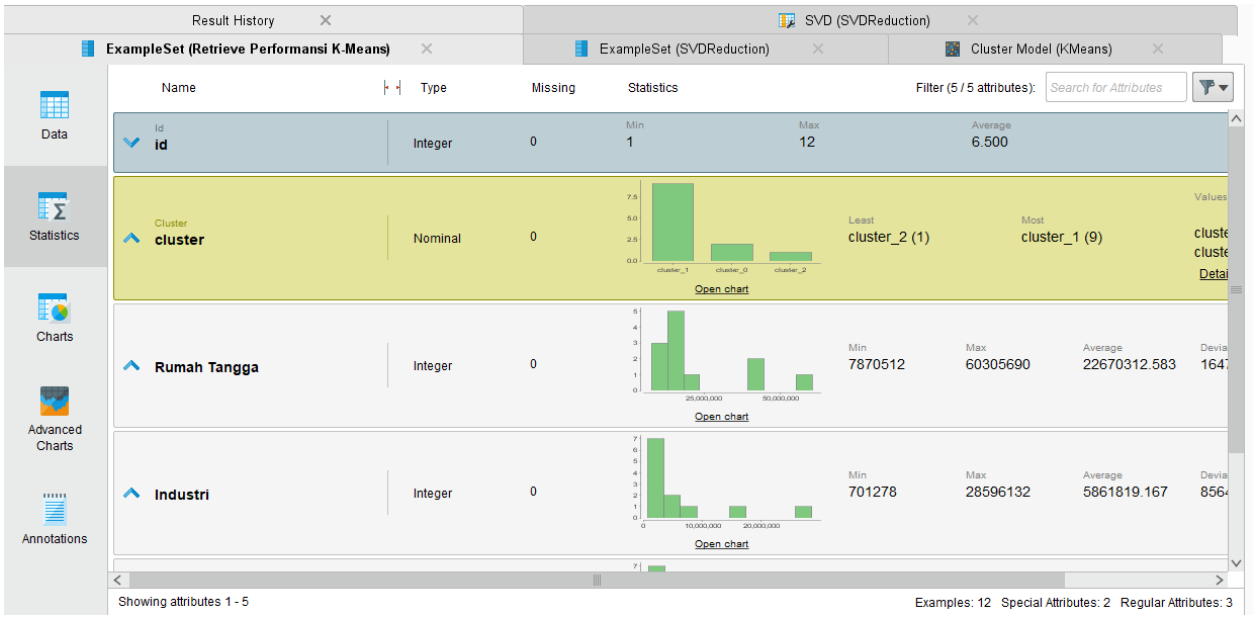

**Parameter K-Means**

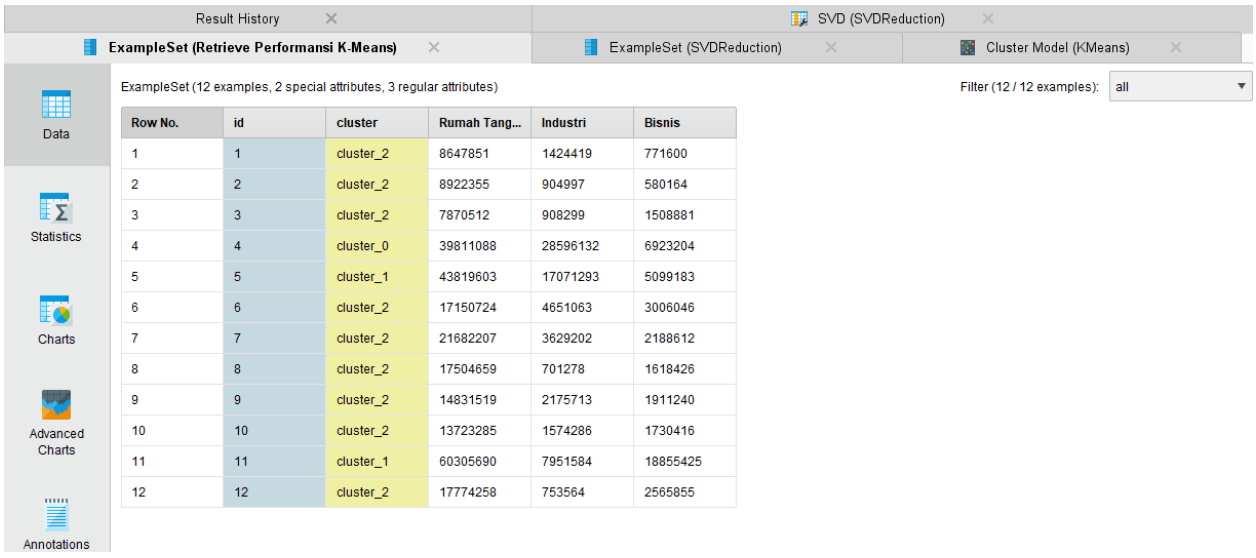


Nama : Bhijanta Wyasa WM
NIM : 202420019
Kelas : MTI23

## Hasil Clustering

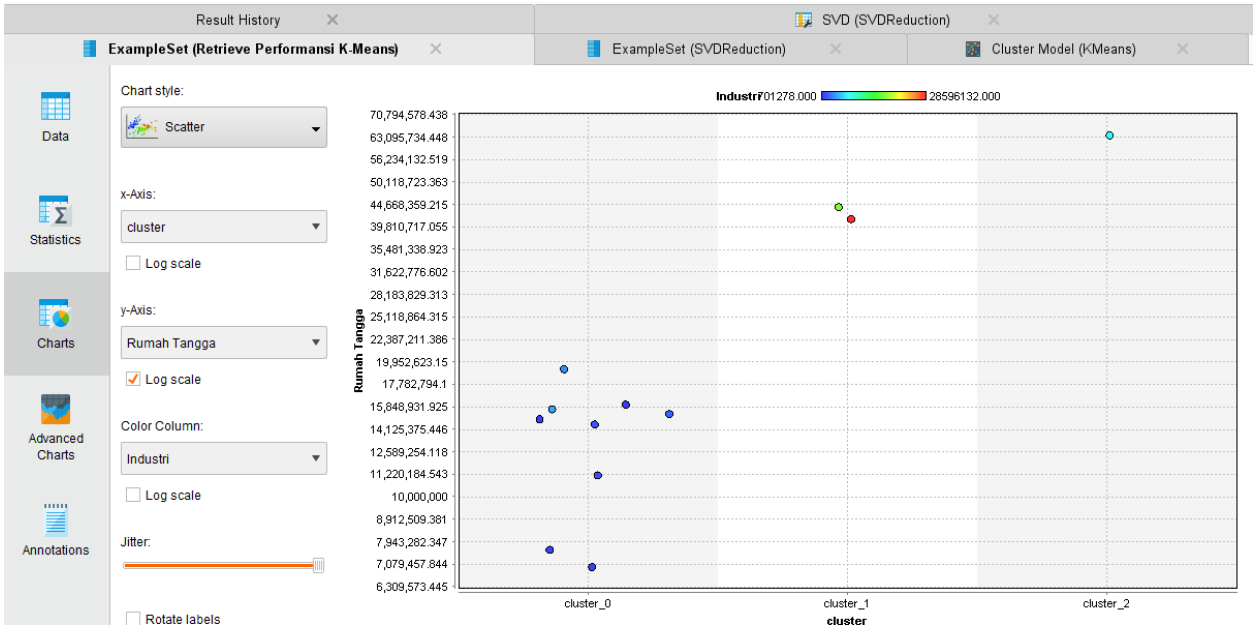



| Row No. | id | cluster | Rumah Tang... | Industri | Bisnis |
|---|---|---|---|---|---|
| 1 | 1 | cluster_2 | 8647851 | 1424419 | 771600 |
| 2 | 2 | cluster_2 | 8922355 | 904997 | 580164 |
| 3 | 3 | cluster_2 | 7870512 | 908299 | 1508881 |
| 4 | 4 | cluster_0 | 39811088 | 28596132 | 6923204 |
| 5 | 5 | cluster_1 | 43819603 | 17071293 | 5099183 |
| 6 | 6 | cluster_2 | 17150724 | 4651063 | 3006046 |
| 7 | 7 | cluster_2 | 21682207 | 3629202 | 2188612 |
| 8 | 8 | cluster_2 | 17504659 | 701278 | 1618426 |
| 9 | 9 | cluster_2 | 14831519 | 2175713 | 1911240 |
| 10 | 10 | cluster_2 | 13723285 | 1574286 | 1730416 |
| 11 | 11 | cluster_1 | 60305690 | 7951584 | 18855425 |
| 12 | 12 | cluster_2 | 17774258 | 753564 | 2565855 |

Nama : Bhijanta Wyasa WM
NIM : 202420019
Kelas : MTI23




## Kesimpulan

*Sumber : https://clusterisasi.com/k-means/admin/iterasi_kmeans_hasil*

1. K-means merupakan salah satu algoritma clustering. Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan supervised learning yang menerima masukan berupa vektor (-x-1 , y1) , (-x-2 , y2) , ..., (-x-i , yi), di mana xi merupakan data dari suatu data pelatihan dan yi merupakan label kelas untuk xi
2. Kelebihan pada algoritma k-means, yaitu [2]:
    - Mudah untuk diimplementasikan dan dijalankan.
    - Waktu yang dibutuhkan untuk menjalankan pembelajaran ini relatif cepat.
    - Mudah untuk diadaptasi.
    - Umum digunakan.
3. Secara performance dari Clustering dengan K-Means cukup dengan baik dapat menyelesaikan clustering baik secara metode Web dan Rapidminer.

**NAMA** : **CORNELIA TRI WAHYUNI**

**MATA KULIAH** : **ADVANCED DATABASE**

**NIM** : **202420044**

## TUGAS 5

Pelajari tentang cara mengukur performance sebuah clustering dan buatlah tutorialnya. Tulis dalam format word dan dilengkapi dengan gambar dan keterangan yang jelas untuk tiap langkah-langkahnya. Sertakan sumber-sumber rujukan yang anda gunakan

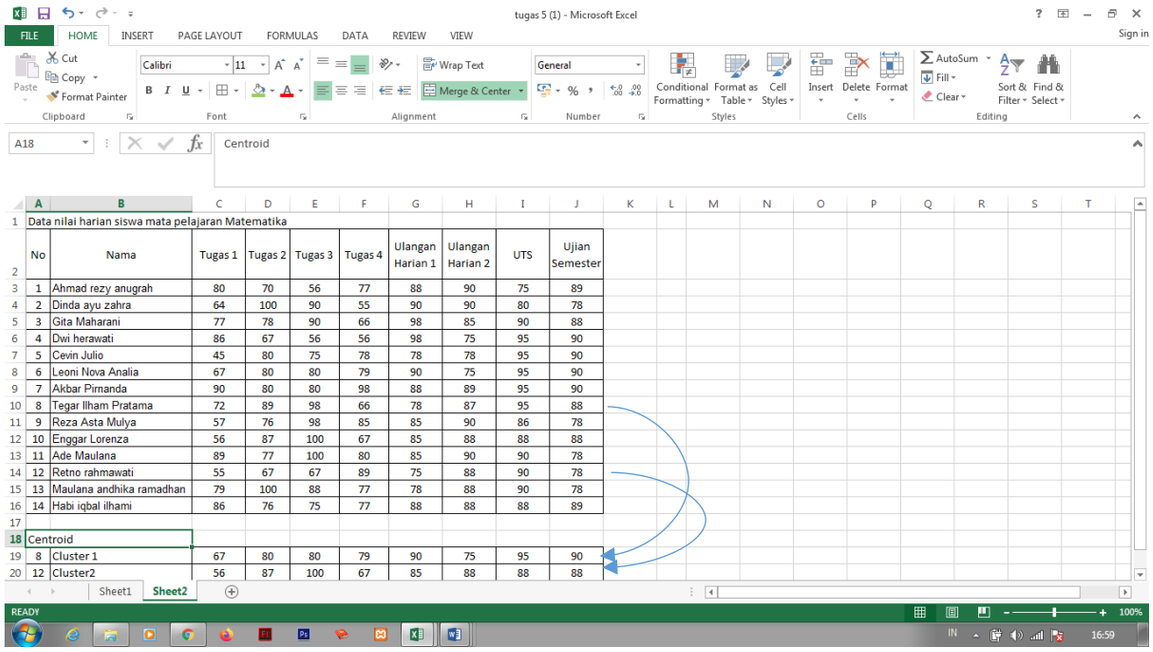
Langkah-langkah sebagai berikut :

1. Terlebih dahulu siapakan data yang akan diolah untuk mengukur clustering dari cluster 1, cluster 2, jarak terdekat dan kelompok data. Contoh data yang dipakai adalah data nilai siswa mata pelajaran matematika. Perhitungan menggunakan Ms.Ecxel

Data nilai harian siswa mata pelajaran Matematika

| No | Nama | Tugas 1 | Tugas 2 | Tugas 3 | Tugas 4 | Ulangan Harian 1 | Ulangan Harian 2 | UTS | Ujian Semester |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Ahmad rezy anugrah | 80 | 70 | 56 | 77 | 88 | 90 | 75 | 89 |
| 2 | Dinda ayu zahra | 64 | 100 | 90 | 55 | 90 | 90 | 80 | 78 |
| 3 | Gita Maharani | 77 | 78 | 90 | 66 | 98 | 85 | 90 | 88 |
| 4 | Dwi herawati | 86 | 67 | 56 | 56 | 98 | 75 | 95 | 90 |
| 5 | Cevin Julio | 45 | 80 | 75 | 78 | 78 | 78 | 95 | 90 |
| 6 | Leoni Nova Analia | 67 | 80 | 80 | 79 | 90 | 75 | 95 | 90 |
| 7 | Akbar Pirnanda | 90 | 80 | 80 | 98 | 88 | 89 | 95 | 90 |
| 8 | Tegar Ilham Pratama | 72 | 89 | 98 | 66 | 78 | 87 | 95 | 88 |
| 9 | Reza Asta Mulya | 57 | 76 | 98 | 85 | 85 | 90 | 86 | 78 |
| 10 | Enggar Lorenza | 56 | 87 | 100 | 67 | 85 | 88 | 88 | 88 |
| 11 | Ade Maulana | 89 | 77 | 100 | 80 | 85 | 90 | 90 | 78 |
| 12 | Retno rahmawati | 55 | 67 | 67 | 89 | 75 | 88 | 90 | 78 |
| 13 | Maulana andhika ramadhan | 79 | 100 | 88 | 77 | 78 | 88 | 90 | 78 |
| 14 | Habi iqbal ilhami | 86 | 76 | 75 | 77 | 88 | 88 | 88 | 89 |

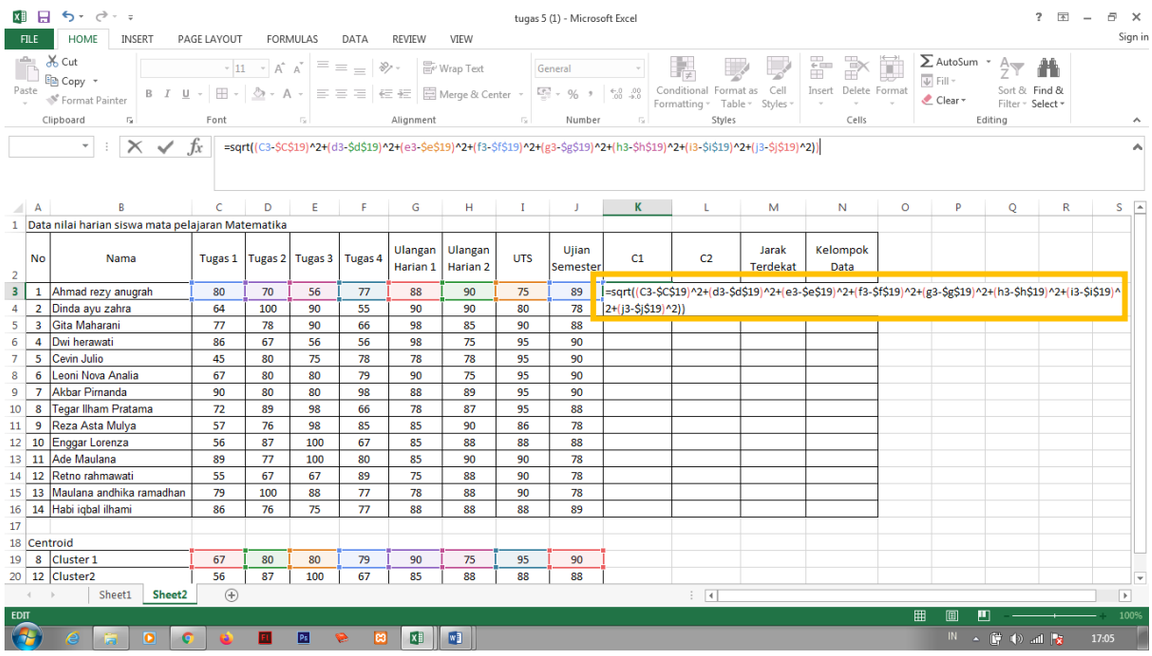
2. Langkah selanjuntnya adalah menentukan centroid atau data pusat pada data ini diambil dari dari data 8 untuk nilai cluster 1 dan data 12 untuk nilai cluster 2



Data nilai harian siswa mata pelajaran Matematika

| No | Nama | Tugas 1 | Tugas 2 | Tugas 3 | Tugas 4 | Ulangan Harian 1 | Ulangan Harian 2 | UTS | Ujian Semester |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Ahmad rezy anugrah | 80 | 70 | 56 | 77 | 88 | 90 | 75 | 89 |
| 2 | Dinda ayu zahra | 64 | 100 | 90 | 55 | 90 | 90 | 80 | 78 |
| 3 | Gita Maharani | 77 | 78 | 90 | 66 | 98 | 85 | 90 | 88 |
| 4 | Dwi herawati | 86 | 67 | 56 | 56 | 98 | 75 | 95 | 90 |
| 5 | Cevin Julio | 45 | 80 | 75 | 78 | 78 | 78 | 95 | 90 |
| 6 | Leoni Nova Analia | 67 | 80 | 80 | 79 | 90 | 75 | 95 | 90 |
| 7 | Akbar Pirnanda | 90 | 80 | 80 | 98 | 88 | 89 | 95 | 90 |
| 8 | Tegar Ilham Pratama | 72 | 89 | 98 | 66 | 78 | 87 | 95 | 88 |
| 9 | Reza Asta Mulya | 57 | 76 | 98 | 85 | 85 | 90 | 86 | 78 |
| 10 | Enggar Lorenza | 56 | 87 | 100 | 67 | 85 | 88 | 88 | 88 |
| 11 | Ade Maulana | 89 | 77 | 100 | 80 | 85 | 90 | 90 | 78 |
| 12 | Retno rahmawati | 55 | 67 | 67 | 89 | 75 | 88 | 90 | 78 |
| 13 | Maulana andhika ramadhan | 79 | 100 | 88 | 77 | 78 | 88 | 90 | 78 |
| 14 | Habi iqbal ilhami | 86 | 76 | 75 | 77 | 88 | 88 | 88 | 89 |
| | | | | | | | | | |
| Centroid | | | | | | | | | |
| 8 | Cluster 1 | 67 | 80 | 80 | 79 | 90 | 75 | 95 | 90 |
| 12 | Cluster2 | 56 | 87 | 100 | 67 | 85 | 88 | 88 | 88 |

3. Setelah menentukan centroid dilakukan menghitung nilai cluster 1 pada data dengan rumus seperti pada gambar, yaitu dari nilai-nilai pada data nilai siswa dijumlahkan dengan data centroid cluster 1

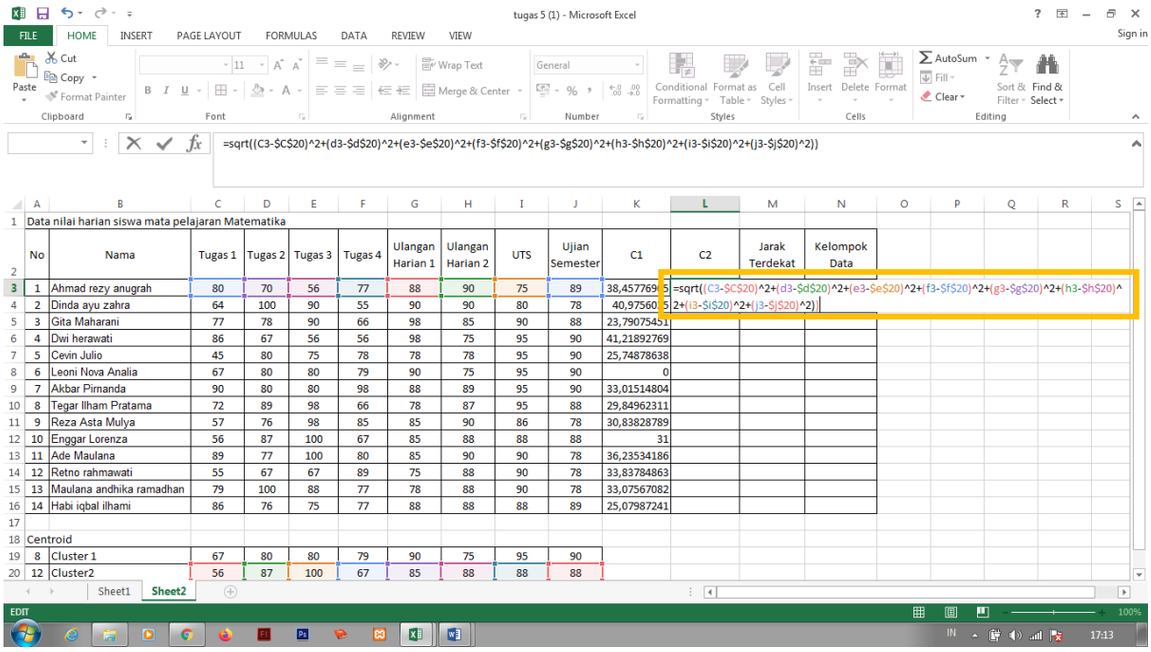


=sqrt((C3-$C$19)^2+(d3-$d$19)^2+(e3-$e$19)^2+(f3-$f$19)^2+(g3-$g$19)^2+(h3-$h$19)^2+(i3-$i$19)^2+(j3-$j$19)^2))

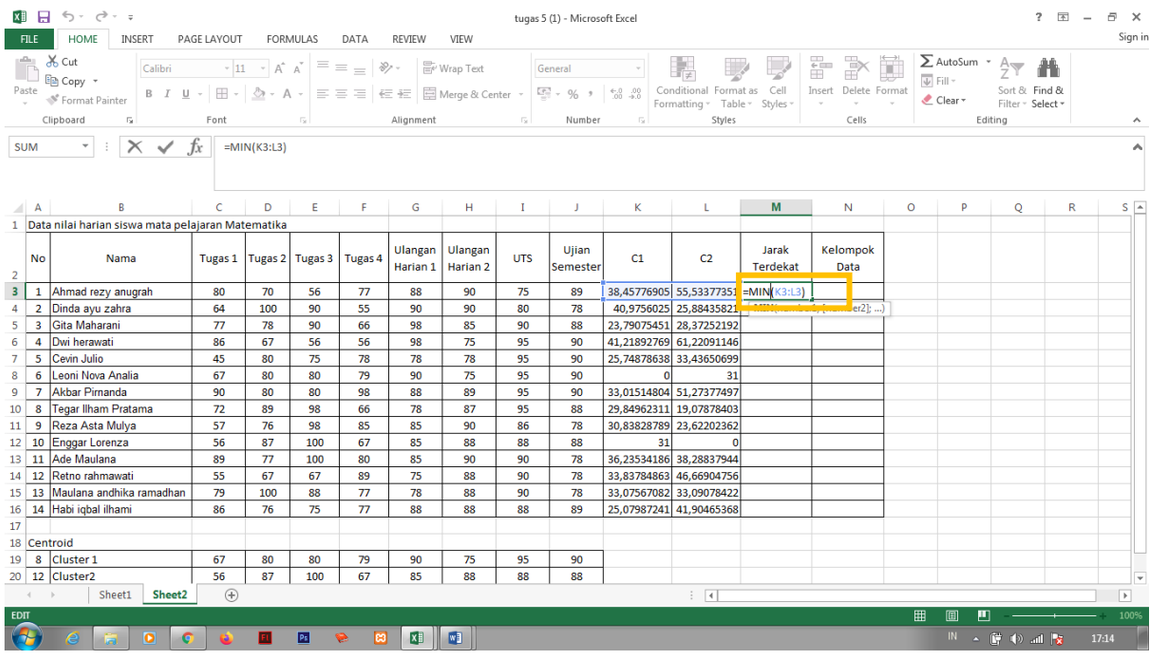
Data nilai harian siswa mata pelajaran Matematika

| No | Nama | Tugas 1 | Tugas 2 | Tugas 3 | Tugas 4 | Ulangan Harian 1 | Ulangan Harian 2 | UTS | Ujian Semester | C1 | C2 | Jarak Terdekat | Kelompok Data |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Ahmad rezy anugrah | 80 | 70 | 56 | 77 | 88 | 90 | 75 | 89 | =sqrt((C3-$C$19)^2+(d3-$d$19)^2+(e3-$e$19)^2+(f3-$f$19)^2+(g3-$g$19)^2+(h3-$h$19)^2+(i3-$i$19)^2+(j3-$j$19)^2)) | | | |
| 2 | Dinda ayu zahra | 64 | 100 | 90 | 55 | 90 | 90 | 80 | 78 | | | | |
| 3 | Gita Maharani | 77 | 78 | 90 | 66 | 98 | 85 | 90 | 88 | | | | |
| 4 | Dwi herawati | 86 | 67 | 56 | 56 | 98 | 75 | 95 | 90 | | | | |
| 5 | Cevin Julio | 45 | 80 | 75 | 78 | 78 | 78 | 95 | 90 | | | | |
| 6 | Leoni Nova Analia | 67 | 80 | 80 | 79 | 90 | 75 | 95 | 90 | | | | |
| 7 | Akbar Pirnanda | 90 | 80 | 80 | 98 | 88 | 89 | 95 | 90 | | | | |
| 8 | Tegar Ilham Pratama | 72 | 89 | 98 | 66 | 78 | 87 | 95 | 88 | | | | |
| 9 | Reza Asta Mulya | 57 | 76 | 98 | 85 | 85 | 90 | 86 | 78 | | | | |
| 10 | Enggar Lorenza | 56 | 87 | 100 | 67 | 85 | 88 | 88 | 88 | | | | |
| 11 | Ade Maulana | 89 | 77 | 100 | 80 | 85 | 90 | 90 | 78 | | | | |
| 12 | Retno rahmawati | 55 | 67 | 67 | 89 | 75 | 88 | 90 | 78 | | | | |
| 13 | Maulana andhika ramadhan | 79 | 100 | 88 | 77 | 78 | 88 | 90 | 78 | | | | |
| 14 | Habi iqbal ilhami | 86 | 76 | 75 | 77 | 88 | 88 | 88 | 89 | | | | |
| | | | | | | | | | | | | | |
| Centroid | | | | | | | | | | | | | |
| 8 | Cluster 1 | 67 | 80 | 80 | 79 | 90 | 75 | 95 | 90 | | | | |
| 12 | Cluster2 | 56 | 87 | 100 | 67 | 85 | 88 | 88 | 88 | | | | |

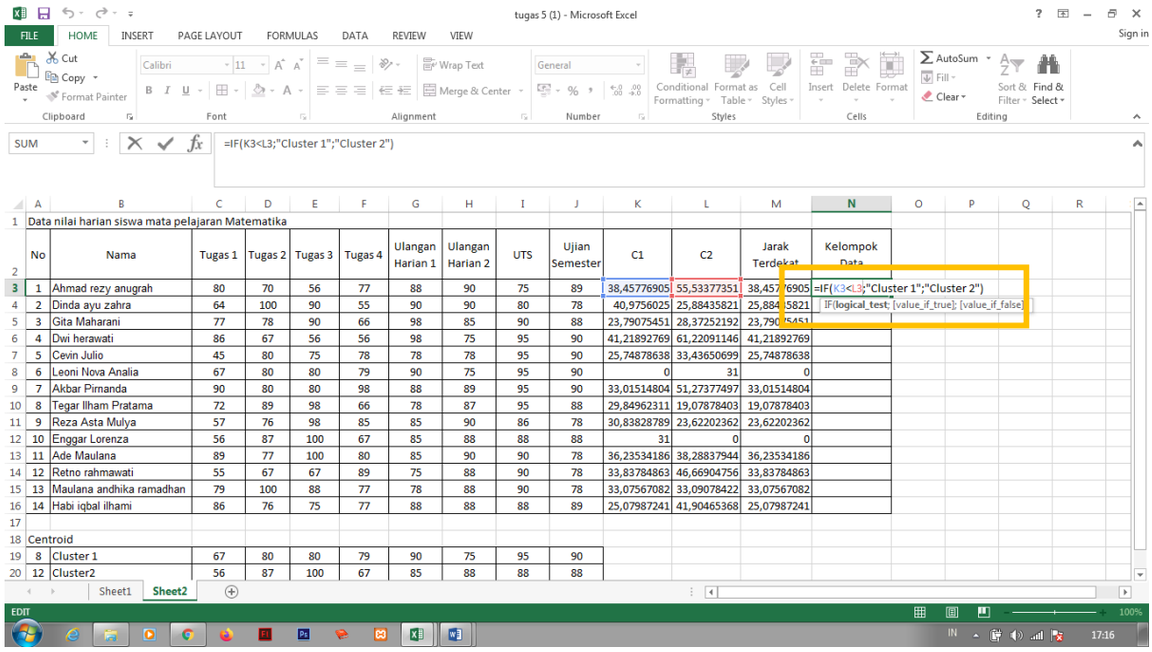
4. Selanjunya lakukan cara yang sama untuk menghitung cluster 2 tapi dengan menggunakan centroid cluster ke 2




5. Berikutnya menghitung jarak terdekat dari cluster dengan rumus seperti pada gambar yaitu mencari nila minimum dari nilai cluster 1 dan cluster 2 lalu didapat jarak terdekat

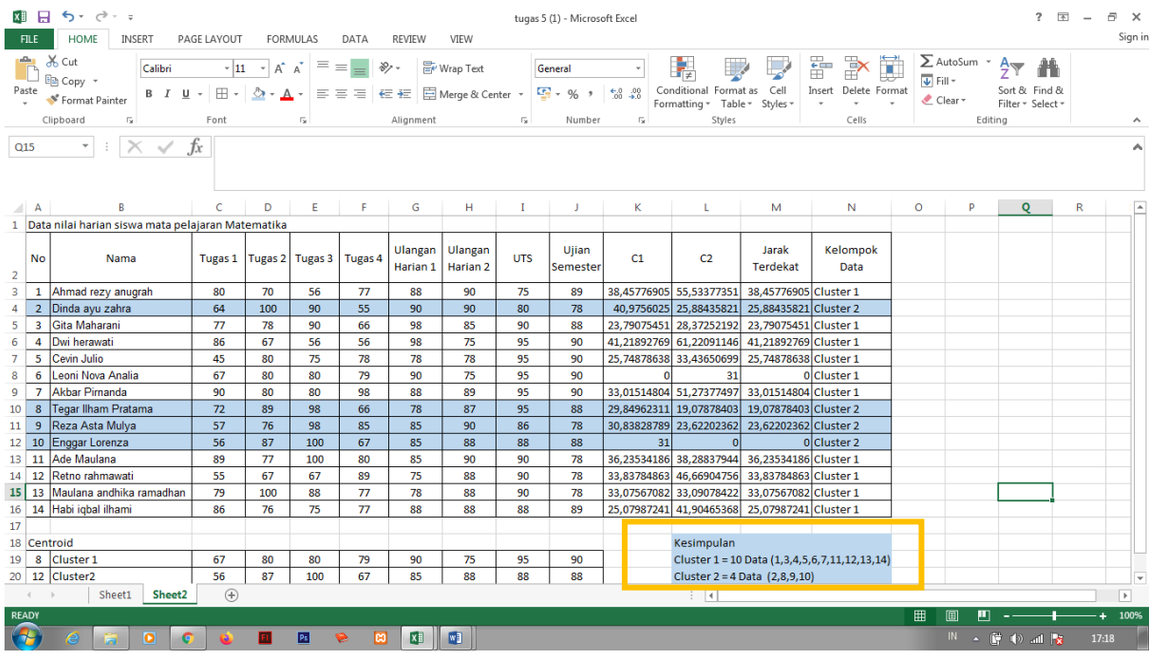


6. Setelah didapatkan jarak terdekat, selanjutnya adalah mengelompokkan data menjadi cluster 1 atau cluster 2 dari C1 dan C2




| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Data nilai harian siswa mata pelajaran Matematika | | | | | | | | | | | | | |
| 2 | No | Nama | Tugas 1 | Tugas 2 | Tugas 3 | Tugas 4 | Ulangan Harian 1 | Ulangan Harian 2 | UTS | Ujian Semester | C1 | C2 | Jarak Terdekat | Kelompok Data |
| 3 | 1 | Ahmad rezy anugrah | 80 | 70 | 56 | 77 | 88 | 90 | 75 | 89 | 38,45776905 | 55,53377351 | 38,45776905 | =IF(K3<L3;"Cluster 1";"Cluster 2") |
| 4 | 2 | Dinda ayu zahra | 64 | 100 | 90 | 55 | 90 | 90 | 80 | 78 | 40,9756025 | 25,88435821 | 25,88435821 | |
| 5 | 3 | Gita Maharani | 77 | 78 | 90 | 66 | 98 | 85 | 90 | 88 | 23,79075451 | 28,37252192 | 23,79075451 | |
| 6 | 4 | Dwi herawati | 86 | 67 | 56 | 56 | 98 | 75 | 95 | 90 | 41,21892769 | 61,22091146 | 41,21892769 | |
| 7 | 5 | Cevin Julio | 45 | 80 | 75 | 78 | 78 | 78 | 95 | 90 | 25,74878638 | 33,43650699 | 25,74878638 | |
| 8 | 6 | Leoni Nova Analia | 67 | 80 | 80 | 79 | 90 | 75 | 95 | 90 | 0 | 31 | 0 | |
| 9 | 7 | Akbar Pirnanda | 90 | 80 | 80 | 98 | 88 | 89 | 95 | 90 | 33,01514804 | 51,27377497 | 33,01514804 | |
| 10 | 8 | Tegar Ilham Pratama | 72 | 89 | 98 | 66 | 78 | 87 | 95 | 88 | 29,84962311 | 19,07878403 | 19,07878403 | |
| 11 | 9 | Reza Asta Mulya | 57 | 76 | 98 | 85 | 85 | 90 | 86 | 78 | 30,83828789 | 23,62202362 | 23,62202362 | |
| 12 | 10 | Enggar Lorenza | 56 | 87 | 100 | 67 | 85 | 88 | 88 | 88 | 31 | 0 | 0 | |
| 13 | 11 | Ade Maulana | 89 | 77 | 100 | 80 | 85 | 90 | 90 | 78 | 36,23534186 | 38,28837944 | 36,23534186 | |
| 14 | 12 | Retno rahmawati | 55 | 67 | 67 | 89 | 75 | 88 | 90 | 78 | 33,83784863 | 46,66904756 | 33,83784863 | |
| 15 | 13 | Maulana andhika ramadhan | 79 | 100 | 88 | 77 | 78 | 88 | 90 | 78 | 33,07567082 | 33,09078422 | 33,07567082 | |
| 16 | 14 | Habi iqbal ilhami | 86 | 76 | 75 | 77 | 88 | 88 | 88 | 89 | 25,07987241 | 41,90465368 | 25,07987241 | |
| 17 | | | | | | | | | | | | | | |
| 18 | Centroid | | | | | | | | | | | | | |
| 19 | 8 | Cluster 1 | 67 | 80 | 80 | 79 | 90 | 75 | 95 | 90 | | | | |
| 20 | 12 | Cluster2 | 56 | 87 | 100 | 67 | 85 | 88 | 88 | 88 | | | | |

7. Dari pengelompokkan data diperoleh kesimpulan untuk cluster 1 memili 10 data dan cluster 2 memiliki 4 data




| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Data nilai harian siswa mata pelajaran Matematika | | | | | | | | | | | | | |
| 2 | No | Nama | Tugas 1 | Tugas 2 | Tugas 3 | Tugas 4 | Ulangan Harian 1 | Ulangan Harian 2 | UTS | Ujian Semester | C1 | C2 | Jarak Terdekat | Kelompok Data |
| 3 | 1 | Ahmad rezy anugrah | 80 | 70 | 56 | 77 | 88 | 90 | 75 | 89 | 38,45776905 | 55,53377351 | 38,45776905 | Cluster 1 |
| 4 | 2 | Dinda ayu zahra | 64 | 100 | 90 | 55 | 90 | 90 | 80 | 78 | 40,9756025 | 25,88435821 | 25,88435821 | Cluster 2 |
| 5 | 3 | Gita Maharani | 77 | 78 | 90 | 66 | 98 | 85 | 90 | 88 | 23,79075451 | 28,37252192 | 23,79075451 | Cluster 1 |
| 6 | 4 | Dwi herawati | 86 | 67 | 56 | 56 | 98 | 75 | 95 | 90 | 41,21892769 | 61,22091146 | 41,21892769 | Cluster 1 |
| 7 | 5 | Cevin Julio | 45 | 80 | 75 | 78 | 78 | 78 | 95 | 90 | 25,74878638 | 33,43650699 | 25,74878638 | Cluster 1 |
| 8 | 6 | Leoni Nova Analia | 67 | 80 | 80 | 79 | 90 | 75 | 95 | 90 | 0 | 31 | 0 | Cluster 1 |
| 9 | 7 | Akbar Pirnanda | 90 | 80 | 80 | 98 | 88 | 89 | 95 | 90 | 33,01514804 | 51,27377497 | 33,01514804 | Cluster 1 |
| 10 | 8 | Tegar Ilham Pratama | 72 | 89 | 98 | 66 | 78 | 87 | 95 | 88 | 29,84962311 | 19,07878403 | 19,07878403 | Cluster 2 |
| 11 | 9 | Reza Asta Mulya | 57 | 76 | 98 | 85 | 85 | 90 | 86 | 78 | 30,83828789 | 23,62202362 | 23,62202362 | Cluster 2 |
| 12 | 10 | Enggar Lorenza | 56 | 87 | 100 | 67 | 85 | 88 | 88 | 88 | 31 | 0 | 0 | Cluster 2 |
| 13 | 11 | Ade Maulana | 89 | 77 | 100 | 80 | 85 | 90 | 90 | 78 | 36,23534186 | 38,28837944 | 36,23534186 | Cluster 1 |
| 14 | 12 | Retno rahmawati | 55 | 67 | 67 | 89 | 75 | 88 | 90 | 78 | 33,83784863 | 46,66904756 | 33,83784863 | Cluster 1 |
| 15 | 13 | Maulana andhika ramadhan | 79 | 100 | 88 | 77 | 78 | 88 | 90 | 78 | 33,07567082 | 33,09078422 | 33,07567082 | Cluster 1 |
| 16 | 14 | Habi iqbal ilhami | 86 | 76 | 75 | 77 | 88 | 88 | 88 | 89 | 25,07987241 | 41,90465368 | 25,07987241 | Cluster 1 |
| 17 | | | | | | | | | | | | | | |
| 18 | Centroid | | | | | | | | | | | | | |
| 19 | 8 | Cluster 1 | 67 | 80 | 80 | 79 | 90 | 75 | 95 | 90 | | | | |
| 20 | 12 | Cluster2 | 56 | 87 | 100 | 67 | 85 | 88 | 88 | 88 | | | | |

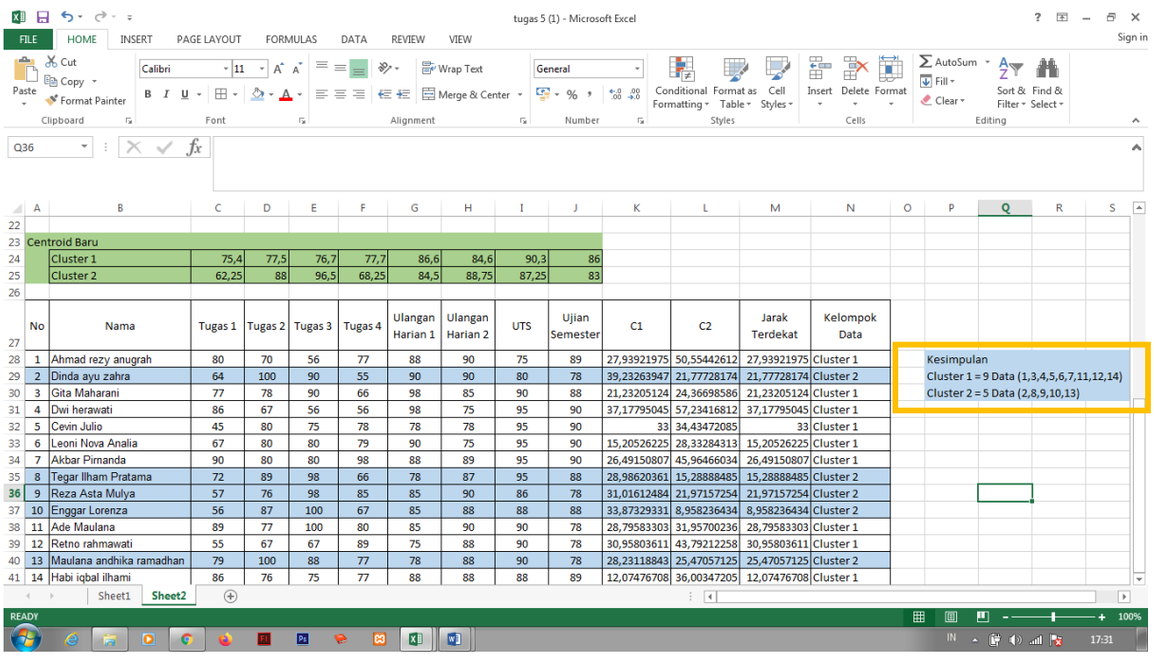
8. Setelah mendapat kesimpulan selanjunya adalah membuat interasi baru tapi sebelum itu kita harus menentukan centroid yang baru terlebih dahulu dengan menghitung jumlah data pada kesimpulan cluster 1 seluruh nilai yang menjadi cluster 1 kan dibagi jumlah data yaitu 10, begitu pula cluster 4 dicari dengan cara yang sama

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 2 | Dinda ayu zahra | 64 | 100 | 90 | 55 | 90 | 90 | 80 | 78 | 40,9756025 | 25,88435821 | 25,88435821 | Cluster 2 |
| 5 | 3 | Gita Maharani | 77 | 78 | 90 | 66 | 98 | 85 | 90 | 88 | 23,79075451 | 28,37252192 | 23,79075451 | Cluster 1 |
| 6 | 4 | Dwi herawati | 86 | 67 | 56 | 56 | 98 | 75 | 95 | 90 | 41,21892769 | 61,22091146 | 41,21892769 | Cluster 1 |
| 7 | 5 | Cevin Julio | 45 | 80 | 75 | 78 | 78 | 78 | 95 | 90 | 25,74878638 | 33,43650699 | 25,74878638 | Cluster 1 |
| 8 | 6 | Leoni Nova Analia | 67 | 80 | 80 | 79 | 90 | 75 | 95 | 90 | 0 | 31 | 0 | Cluster 1 |
| 9 | 7 | Akbar Pirnanda | 90 | 80 | 80 | 98 | 88 | 89 | 95 | 90 | 33,01514804 | 51,27377497 | 33,01514804 | Cluster 1 |
| 10 | 8 | Tegar Ilham Pratama | 72 | 89 | 98 | 66 | 78 | 87 | 95 | 88 | 29,84962311 | 19,07878403 | 19,07878403 | Cluster 2 |
| 11 | 9 | Reza Asta Mulya | 57 | 76 | 98 | 85 | 85 | 90 | 86 | 78 | 30,83828789 | 23,62202362 | 23,62202362 | Cluster 2 |
| 12 | 10 | Enggar Lorenza | 56 | 87 | 100 | 67 | 85 | 88 | 88 | 88 | 31 | 0 | 0 | Cluster 2 |
| 13 | 11 | Ade Maulana | 89 | 77 | 100 | 80 | 85 | 90 | 90 | 78 | 36,23534186 | 38,28837944 | 36,23534186 | Cluster 1 |
| 14 | 12 | Retno rahmawati | 55 | 67 | 67 | 89 | 75 | 88 | 90 | 78 | 33,83784863 | 46,66904756 | 33,83784863 | Cluster 1 |
| 15 | 13 | Maulana andhika ramadhan | 79 | 100 | 88 | 77 | 78 | 88 | 90 | 78 | 33,07567082 | 33,09078422 | 33,07567082 | Cluster 1 |
| 16 | 14 | Habi iqbal ilhami | 86 | 76 | 75 | 77 | 88 | 88 | 88 | 89 | 25,07987241 | 41,90465368 | 25,07987241 | Cluster 1 |
| 17 | | | | | | | | | | | | | | |
| 18 | Centroid | | | | | | | | | | | Kesimpulan | | |
| 19 | 8 | Cluster 1 | 67 | 80 | 80 | 79 | 90 | 75 | 95 | 90 | | Cluster 1 = 10 Data (1,3,4,5,6,7,11,12,13,14) | | |
| 20 | 12 | Cluster2 | 56 | 87 | 100 | 67 | 85 | 88 | 88 | 88 | | Cluster 2 = 4 Data (2,8,9,10) | | |
| 21 | | | | | | | | | | | | | | |
| 22 | | | | | | | | | | | | | | |
| 23 | Centroid Baru | | | | | | | | | | | | | |
| 24 | | Cluster 1 | =sum(C3;C5:C9;C13:C16)/10 | | | | | | | | | | | |
| 25 | | Cluster 2 | | | | | | | | | | | | |

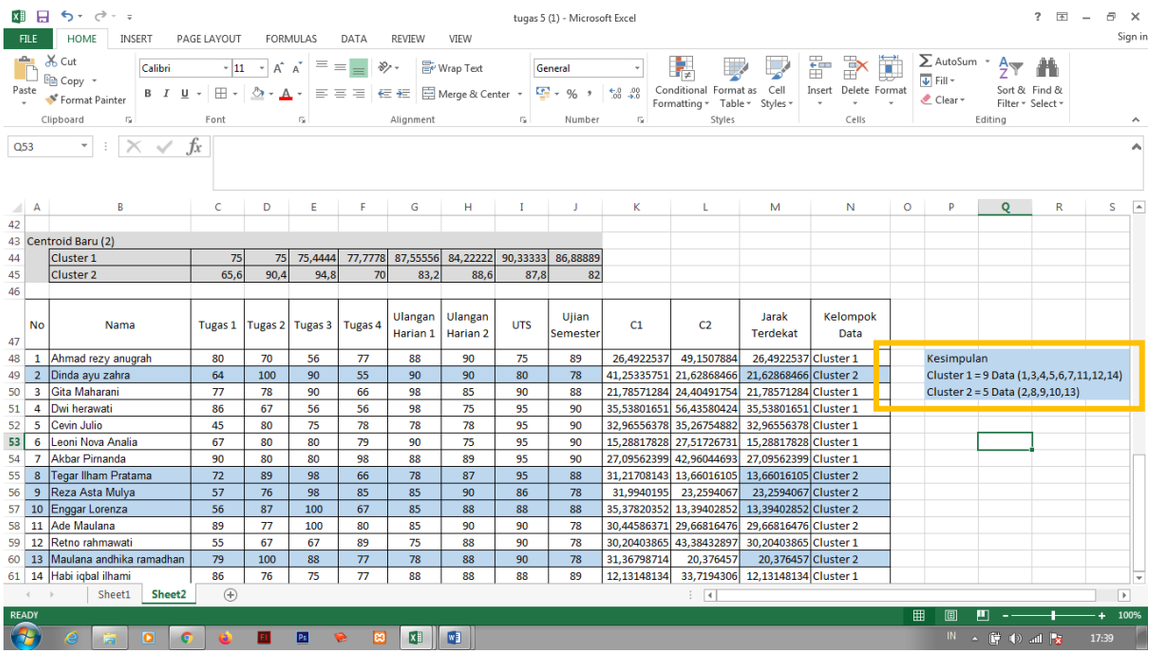
9. Setelah dilakukan menghitung nilai centroid baru maka akan didapat nilai centroid cluster 1 dan cluster 2

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 43 | Centroid Baru (2) | | | | | | | | |
| 44 | | Cluster 1 | 75 | 75 | 75,4444 | 77,7778 | 87,55556 | 84,22222 | 90,33333 | 86,88889 |
| 45 | | Cluster 2 | 65,6 | 90,4 | 94,8 | 70 | 83,2 | 88,6 | 87,8 | 82 |

10. Dari data centroid yang baru akan dilakukan kembali perhitungan untuk iterasi ke 2 dengan mengulang langkah yang sama seperti langka k3 hingga langkah ke 7 dengan hasil seperti pada gambar, dari hasil yang didapat adalah kelompok data cluster interasi 2 berbeda dengan interasi pertama yaitu pada interasi pertama cluster 1 memiliki 10 data dan cluster 2 memiliki 5 data sementara pada interasi 2 diperoleh data cluster 1 yaitu 9 data dan cluster 2 5 data maka pengkuran masih harus dilanjutkan



11. Pada tahap ini dilakukan kembali perhitungan untuk memperoleh centroid baru dan menghitung nilai cluster 1, cluster 2, jaraj terdekat dan pengelompokkan data. Setelah dilakukan pengkuran clustering pada interasi ke 3 ini didapat hasil yang sama pada interasi kedua yaitu cluster 1 berjumlah 9 dan cluster 2 berjumlah 5 makan perhitungan atau pengukuran dapat distop/dihentikan karena telah memperoleh data yang sama.



tugas 5 (1) - Microsoft Excel

Centroid Baru (2)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Cluster 1 | 75 | 75 | 75,4444 | 77,7778 | 87,55556 | 84,22222 | 90,33333 | 86,88889 |
| Cluster 2 | 65,6 | 90,4 | 94,8 | 70 | 83,2 | 88,6 | 87,8 | 82 |

| No | Nama | Tugas 1 | Tugas 2 | Tugas 3 | Tugas 4 | Ulangan Harian 1 | Ulangan Harian 2 | UTS | Ujian Semester | C1 | C2 | Jarak Terdekat | Kelompok Data |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Ahmad rezy anugrah | 80 | 70 | 56 | 77 | 88 | 90 | 75 | 89 | 26,4922537 | 49,1507884 | 26,4922537 | Cluster 1 |
| 2 | Dinda ayu zahra | 64 | 100 | 90 | 55 | 90 | 90 | 80 | 78 | 41,25335751 | 21,62868466 | 21,62868466 | Cluster 2 |
| 3 | Gita Maharani | 77 | 78 | 90 | 66 | 98 | 85 | 90 | 88 | 21,78571284 | 24,40491754 | 21,78571284 | Cluster 1 |
| 4 | Dwi herawati | 86 | 67 | 56 | 56 | 98 | 75 | 95 | 90 | 35,53801651 | 56,43580424 | 35,53801651 | Cluster 1 |
| 5 | Cevin Julio | 45 | 80 | 75 | 78 | 78 | 78 | 95 | 90 | 32,96556378 | 35,26754882 | 32,96556378 | Cluster 1 |
| 6 | Leoni Nova Analia | 67 | 80 | 80 | 79 | 90 | 75 | 95 | 90 | 15,28817828 | 27,51726731 | 15,28817828 | Cluster 1 |
| 7 | Akbar Pirnanda | 90 | 80 | 80 | 98 | 88 | 89 | 95 | 90 | 27,09562399 | 42,96044693 | 27,09562399 | Cluster 1 |
| 8 | Tegar Ilham Pratama | 72 | 89 | 98 | 66 | 78 | 87 | 95 | 88 | 31,21708143 | 13,66016105 | 13,66016105 | Cluster 2 |
| 9 | Reza Asta Mulya | 57 | 76 | 98 | 85 | 85 | 90 | 86 | 78 | 31,9940195 | 23,2594067 | 23,2594067 | Cluster 2 |
| 10 | Enggar Lorenza | 56 | 87 | 100 | 67 | 85 | 88 | 88 | 88 | 35,37820352 | 13,39402852 | 13,39402852 | Cluster 2 |
| 11 | Ade Maulana | 89 | 77 | 100 | 80 | 85 | 90 | 90 | 78 | 30,44586371 | 29,66816476 | 29,66816476 | Cluster 2 |
| 12 | Retno rahmawati | 55 | 67 | 67 | 89 | 75 | 88 | 90 | 78 | 30,20403865 | 43,38432897 | 30,20403865 | Cluster 1 |
| 13 | Maulana andhika ramadhan | 79 | 100 | 88 | 77 | 78 | 88 | 90 | 78 | 31,36798714 | 20,376457 | 20,376457 | Cluster 2 |
| 14 | Habi iqbal ilhami | 86 | 76 | 75 | 77 | 88 | 88 | 88 | 89 | 12,13148134 | 33,7194306 | 12,13148134 | Cluster 1 |

Kesimpulan
Cluster 1 = 9 Data (1,3,4,5,6,7,11,12,14)
Cluster 2 = 5 Data (2,8,9,10,13)

Sumber :

1. https://youtu.be/8mXL2z3lg_o
2. https://www.youtube.com/watch?v=5o_JQ6AHA0Y
3. https://www.youtube.com/watch?v=MayK_kwJgv4

Nama : Cynthia Anisa Agatha

NIM : 202420022

Clustering adalah pengelompokkan data dengan sifat yang mirip. Data untuk proses clustering tidak memiliki label (kelas). Pada umumnya, algoritma clustering dapat dikategorikan menjadi dua macam berdasarkan hasil yang diinginkan yaitu

1. *Partitional*, yaitu menentukan partisi sebanyak *K*

2. *Hierarchical*, yaitu mengelompokan data berdasarkan struktur taksonomi.

- ***Partitional Clustering***

*Partitional clustering* salah satu algoritma nya yaitu *K-means* dimana algoritma ini mengelompokkan data menjadi sebanyak *K* kelompok sesuai yang kita deifinisikan. Algoritma ini juga sering disebut sebagai flat clulstering. Algoritma ini terdiri dari 3 fase Pemilihan centroid dan mencari kemiripan data.

Dalam pemilihan *centroid* paling sederhana yaitu dilakukan pemilihan secara acak. Pada kenyataannya, inisiasi *centroid* untuk setaip kelompok *cluster* dapat dilakukan secara acak, tetapi pada tahap selanjutnya pada umumnya *centroid* dipilih menggunakan nilai rata-rata/*mean*. dari *centroid* yang telah dipilih secara acak

Misalkan diberikan sebuah kumpulan data **D** = $\{\boldsymbol{d_1}, \boldsymbol{d_2}, \ldots, \boldsymbol{d_n}\}$; maka *centroid c* untuk *cluster* dihitung denga persamaan.

$$c = \frac{1}{N}\sum_{i=1}^{N}\sum_{e=1}^{F} d_{i[e]}$$

Dimana nilai rata-rata setiap elemen *feature vector* untuk seluruh anggota *cluster* tersebut, dimana *N* adalah banyaknya anggota *cluster, F* adalah dimensi vektor, $d_i$ adalah anggota ke-*I* dalam reperesentasi *feature vector* dan $d_{i[e]}$ melambangkan elemen ke –*e* pada vektor $d_i$. Maka

dari itu *centorid* umumnya bisa jadi bukan merupakan anggota elemen *cluster* (*centorid* bukan sebuah *instance* data).

- ***Hierarchical Clustering***

*Hierarchical clustering* adalah teknik untuk membentuk pembagian bersarang (*nested partitiion*). Berbeda dengan *K-means* yang hasil *clustering*-nya berbentuk *flat* atau rata, *hierarchical clustering* memiliki satu *cluster* paling atas yang mencakup konsep seluruh *cluster* dibawahnya. Terdapat dua cara untuk membentuk *hierarchical clustering*, yaitu:

1. **Agglomerative**. Dimulai dari beberapa *flat clusters*; pada setiap langkah iterasi, kita menggabungkan dua *clusters* paling identik. Artinya, kita harus mendeskripsikan arti "kedekatan" dua *clusters*.
2. **Divisive**. Dimulai dari satu *cluster* (seluruh data), kemudian kita memecah belah *cluster.* Artinya, kita harus mendefinisikan *cluster* mana yang harus dipecah dan bagaimana cara memecahnya.

- **Validasi Cluster**

Validasi cluster merupakan prosedur yang mengevaluasi hasil dari analisis clustering dengan cara yang objektif dan kuantitatif. Indeks validasi cluster untuk mengukur ketepatan struktur yang diperoleh dari analisis *clustering*, dalam hal dapat diterjemahkan secara objektif, adalah probabilitas. Ketepatan struktrur *cluster* menyatakan bahwa struktur menyediakan informasi yang benar mengenai data, atau kemampuan struktrur yang diperoleh untuk merefleksikan karakter intrinsik data. Tiga jenis struktur tersebut adalah *hierarchies*, *partitions*, dan *clusters*. *Hierarchies* adalah sekumpulan sekuens partisi. *Partitions* adalah hasil dari algoritma *clustering partitional*/non hierarkis, dan cluster adalah himpunan bagian individual dari *pattern*.

Terdapat tiga kriteria prosedur untuk validasi struktur *clustering* yaitu kriteria eksternal, internal, dan relatif.

Kriteria eksternal mengukur performansi dengan membandingkan struktur clustering dengan informasi yang sudah ada sebelumnya (*priori* information). Misalnya, kriteria eksternal mengukur derajat koresponden antara jumlah cluster yang diperoleh dari sebuah algoritma clustering dengan kategori label sebagai informasi priori.

Kriteria internal memperkirakan kecocokan antara struktur dan data, menggunakan data itu sendiri. Misalnya, kriteria internal akan mengukur derajat sebuah partisi yang diperoleh dari algoritma clustering akan ditempatkan oleh proximity matrix yang diberikan.

Kriteria relatif menentukan struktur yang lebih baik bila dibandingkan dengan struktur yang lain, seperti lebih stabil atau lebih cocok untuk data tertentu. Misalnya, kriteria relatif akan mengukur secara kuatitatif kesesuaian data dengan penerapan single-link atau complete-link pada hierarkis.

Untuk validasi hasil clustering, kriteria eksternal memiliki keunggulan karena melakukan penilaian kualitas cluster yang terpisah dan tidak mengandung perkiraaan-perkiraan atau prasangka. Namun, kriteria eksternal juga memiliki kerugian karena untuk pengukuran data ekspresi gen, "gold standard" eksternal jarang sekali ada.

Kriteria internal menghindari penggunaan standar, tetapi memiliki alternatif yaitu melakukan validasi terhadap cluster dengan menggunakan informasi yang sama dengan informasi untuk clustering.

Pendekatan untuk menggunakan kriteria eksternal adalah dengan mengasumsikan bahwa tidak dibutuhkan standar eksternal. Evaluasi penilaian kualitas cluster menggunakan satu kondisi (kolom) saja dari raw data yang akan digunakan sebagai masukan validasi. Contoh penerapan kriteria eksternal adalah *Figure Of Merit* (FOM).

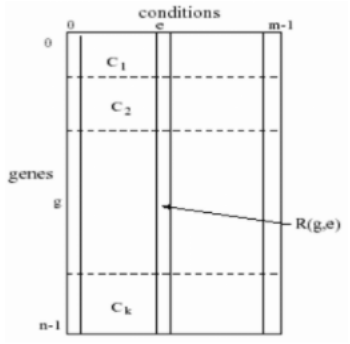
- ***Figure of Merit***

*Figure Of Merit* adalah pengukuran kemampuan prediktif suatu algoritma *clustering*. Himpunan data ekspresi gen terdiri dari ukuran level ekspresi dari n gen yang diukur terhadap m kondisi. Misalkan sebuah algoritma *clustering* diaplikasikan terhadap data dari kondisi 1, …., (e-1), (e+1),

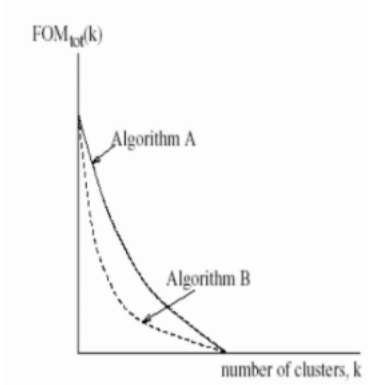
…., m, dan kondisi e digunakan untuk memperkirakan kemampuan prediktif algoritma. Misalkan terdapat cluster sebanyak k, C1, C2, ……, Ck. R(g,e) adalah level ekspresi gen g untuk kondisi e pada raw data matriks (gambar 3-10). μC1(e) adalah nilai rata-rata level ekspresi pada kondisi e dari gen pada cluster C1. Normalisasi ke-2 dari FOM adalah deviasi akar kuadrat nilai tengah pada kondisi e dari level ekspresi gen relatif terhadap nilai tengah cluster dirumuskan,

$$FOM(e,k) = (R(x,e) - \mu C_i(e))^2.$$

Setiap m kondisi dapat digunakan sebagai kondisi eksperimen yang dibuang (*left-out*). *Aggregate* FOM, FOMtot(k) = $\sum_{e=0}^{m-1} FOM\ (e,k)$ , adalah estimasi dari total kemampuan prediktif algoritma terhadap semua kondisi untuk sebanyak k cluster pada himpunan data.




Perhitungan terhadap data *raw matrix*




Perbandingan Algoritma A dan B

Kualitas hasil *clustering* dapat dihitung dengan tiga kriteria, yaitu:

1. *Intra-Cluster Similiarity,* yaitu menghitung rata-rata kedekatan antara suatu anggota dan anggota *cluster* lainnya.

$$I = \frac{1}{N^2} \sum_{\mathbf{d}_i \mathbf{d}_j, i \neq j} \text{cosSim}(\mathbf{d}_i, \mathbf{d}_j)$$

Perhitungan kedekatan antar tiap pasang anggota *cluster* sama halnya dengan menghitung *norm* dari *centroid cluster* tersebut, ketika *centroid* dihitung menggunakan *mean.*

$$I = \frac{1}{N^2} \sum_{\mathbf{d}_i \mathbf{d}_j, i \neq j} \text{cosSim}(\mathbf{d}_i, \mathbf{d}_j) = \|\mathbf{c}\|^2$$

Perhitungan ini dapat dinormalisasi sesuai dengan banyaknya anggota *cluster*

$$I' = \frac{\|\mathbf{c}\|^2}{N}$$

Semakin tinggi kemiripan anggota pada suatu *cluster,* semakin baik kualitas *cluster* tersebut.

2. *Inter-cluster similarity,* yaitu menghitung bagaimana perbedaan antar suatu *cluster* dan *cluster* lainnya. Hal tersebut dihitung dengan *cosine similiarity* antara *centroid* suatu *cluster* dan *centroid* dari seluruh data, dimana $c_k$adalah centroid *cluster* ke- *k,* $c^D$adalah centroid (*mean*) dari seluruh data, $N_k$adalah banyaknya anggota *cluster* ke-*k*, dan *K* adalah banyaknya *clusters*. Semakin kecil nilai *inter-cluster similarity,* maka semakin baik kualitas *clustering*.

$$E = \sum_{k=1}^{K} N_k \frac{\mathbf{c}_k \mathbf{c}^{\mathbf{D}}}{\|\mathbf{c}_k\|}$$

3. *Hybrid*. Perhitungan *intra-cluster* dan *inter-cluster* mengoptimalkan satu hal sementara tidak memperdulikan hal lainnya. *Intra-cluster* menghitung keeratan anggota *cluster*, sementara *Inter-cluster* menghitung separasi antar *clusters.* Kita dapat menggabungkan keduanya sebagai *hybrid* (gabungan), dihitung dengan:

$$H = \frac{\sum_{k=1}^{K} I'_k}{E} = \frac{\sum_{k=1}^{K} \frac{\|\mathbf{c}_k\|^2}{N_k}}{\sum_{k=1}^{K} N_k \frac{\mathbf{c}_k \mathbf{c}^{\mathbf{D}}}{\|\mathbf{c}_k\|}} = \sum_{k=1}^{K} \frac{\|\mathbf{c}_k\|}{N_k^2 \mathbf{c}_k \mathbf{c}^{\mathbf{D}}}$$

Semakin besar nilai perhitungan *hybrid*, semakin bagus kualitas *clusters*.Apabila terdapat informasi label/kelas untuk setiap data, kita juga dapat menghitung kualitas algoritma *clustering* dengan *Entropy* dan *Purity*.

Sumber: - Lumbantoruan Rosni, *Pengukuran Kemampuan Prediktif Teknik Clustering dengan Figure of Merit*, Sumatera Utara.

- Introduction to Machine Learnig Chapter 10