K-means Algorithm (KMA)

Kasus 1.

Diketahui angka kematian kasar (CDR) dan angka kelahiran kasar (CBR) 10 negara seperti terlihat pada Tabel 1. Negara-negara tersebut akan dikelompokkan berdasarkan CBR dan CDRnya menjadi tiga kelompok. Proses pengelompokkan menggunakan metode k-means.

Tabel 1. CDR dan CBR tahun 1994 (sumber: ESCAP Population Data Sheet 1996)
http://bankdata.depkes.go.id/Profil/Indo98/Annex/lvia5.htm

No	Negara	CBR	CDR
1	Brunei Darusalam	27	3
2	Kamboja	38	14
3	Indonesia	24	8
4	Laos	43	15
5	Malaysia	28	5
6	Myanmar	32	11
7	Filipina	30	7
8	Singapura	17	5
9	Thailand	20	6
10	Vietnam	29	8

1. Misalkan kita akan pengelompokkan data tersebut menjadi 3 cluster, K = 3. Misalkan pusat cluster kita tetapkan sembarang, c_1 = (20, 5); c_2 = (25,4); dan c_3 = (30,10).

2. Hitung jarak setiap data terhadap setiap pusat cluster. Misalkan untuk menghitung jarak data pertama (Brunei Darusalam) dengan pusat cluster pertama adalah:

   $$d_{11} = \sqrt{(27-20)^2 + (3-5)^2} = 7{,}2801$$

   Jarak data pertama (Brunei Darusalam) dengan pusat cluster kedua adalah:

   $$d_{12} = \sqrt{(27-25)^2 + (3-4)^2} = 2{,}2361$$

   Jarak data pertama (Brunei Darusalam) dengan pusat cluster ketiga adalah:

   $$d_{13} = \sqrt{(27-30)^2 + (3-10)^2} = 7{,}6158$$

   Hasil perhitungan jarak selengkapnya adalah;

No	Negara	CBR	CDR	Jarak		
				C1	C2	C3
1	Brunei Darusalam	27	3	7,2801	2,2361	7,6158
2	Kamboja	38	14	20,1246	16,4012	8,9443
3	Indonesia	24	8	5,0000	4,1231	6,3246
4	Laos	43	15	25,0799	21,0950	13,9284
5	Malaysia	28	5	8,0000	3,1623	5,3852
6	Myanmar	32	11	13,4164	9,8995	2,2361
7	Filipina	30	7	10,1980	5,8310	3,0000
8	Singapura	17	5	3,0000	8,0623	13,9284
9	Thailand	20	6	1,0000	5,3852	10,7703
10	Vietnam	29	8	9,4868	5,6569	2,2361

3. Suatu data akan menjadi anggota dari suatu cluster yang memiliki jarak terkecil dari pusat clusternya. Misalkan untuk data pertama, jarak terkecil diperoleh pada cluster kedua, sehingga data pertama akan menjadi anggota dari cluster pertama. Demikian juga untuk data kedua (Kamboja), jarak terkecil ada pada cluster ketiga, maka data tersebut akan masuk pada cluster ketiga. Posisi cluster selengkapnya adalah:

No	Negara	CBR	CDR	Anggota Cluster		
				C1	C2	C3
1	Brunei Darusalam	27	3		*	
2	Kamboja	38	14			*
3	Indonesia	24	8		*	
4	Laos	43	15			*
5	Malaysia	28	5		*	
6	Myanmar	32	11			*
7	Filipina	30	7			*
8	Singapura	17	5	*		
9	Thailand	20	6	*		
10	Vietnam	29	8			*

Berdasarkan hasil penggolongan tersebut, diperoleh anggota cluster pertama ada 2, cluster kedua ada 3, dan cluster keempat ada 5.

4. Hitung pusat cluster baru. Untuk cluster pertama, ada 2 data yaitu data ke-8 dan data ke-9, sehingga:

$$c_{11} = \frac{17 + 20}{2} = 18{,}5;$$

$$c_{12} = \frac{5 + 6}{2} = 5{,}5.$$

Untuk cluster kedua, ada 3 data yaitu data ke-1, data ke-3 dan data ke-5, sehingga:

$$c_{21} = \frac{27 + 24 + 28}{3} = 26{,}33\,;$$

$$c_{22} = \frac{3 + 8 + 5}{3} = 5{,}33$$

Untuk cluster ketiga, ada 5 data yaitu data ke-2, data ke-4, data ke-6, data ke-7 dan data ke-10, sehingga:

$$c_{31} = \frac{38 + 43 + 32 + 30 + 29}{5} = 34{,}4\,;$$

$$c_{32} = \frac{14 + 15 + 11 + 7 + 8}{5} = 11$$

5. Ulangi menghitung jarak setiap data terhadap setiap pusat cluster yang baru. Hasil perhitungan jarak selengkapnya terlihat pada Tabel 4.

No	Negara	CBR	CDR	Jarak		
				C1	C2	C3
1	Brunei Darusalam	27	3	8,8600	2,4267	10,8977
2	Kamboja	38	14	21,2720	14,5335	4,6861
3	Indonesia	24	8	6,0415	3,5434	10,8240
4	Laos	43	15	26,2774	19,2671	9,4847
5	Malaysia	28	5	9,5131	1,6997	8,7727
6	Myanmar	32	11	14,5774	8,0139	2,4000
7	Filipina	30	7	11,5974	4,0277	5,9464
8	Singapura	17	5	1,5811	9,3393	18,4054
9	Thailand	20	6	1,5811	6,3683	15,2434
10	Vietnam	29	8	10,7935	3,7712	6,1774

6. Posisi cluster selengkapnya terlihat pada Tabel 5.

No	Negara	CBR	CDR	Anggota Cluster		
				C1	C2	C3
1	Brunei Darusalam	27	3		*	
2	Kamboja	38	14			*
3	Indonesia	24	8		*	
4	Laos	43	15			*
5	Malaysia	28	5		*	
6	Myanmar	32	11			*
7	Filipina	30	7		**	
8	Singapura	17	5	*		

9	Thailand	20	6	*		
10	Vietnam	29	8		**	

Terlihat masih ada 2 data yang berubah posisi dari kondisi semula, yaitu data ke-7 dan ke-10. Sehingga perlu dihitung pusat cluster baru.

7. Hitung pusat cluster baru sebagaimana pada langkah ke-4, sehingga diperoleh:

   c_{11} = 18,5; c_{12} = 5,5;
   c_{21} = 27,6; c_{22} = 6,2;
   c_{31} = 37,67; c_{32} = 13,33;

8. Ulangi menghitung jarak setiap data terhadap setiap pusat cluster yang baru. Hasil perhitungan jarak selengkapnya terlihat pada Tabel 6.

No	Negara	CBR	CDR	Jarak		
				C1	C2	C3
1	Brunei Darusalam	27	3	8,8600	3,2558	14,8511
2	Kamboja	38	14	21,2720	13,0000	0,7454
3	Indonesia	24	8	6,0415	4,0249	14,6705
4	Laos	43	15	26,2774	17,7370	5,5877
5	Malaysia	28	5	9,5131	1,2649	12,7628
6	Myanmar	32	11	14,5774	6,5115	6,1283
7	Filipina	30	7	11,5974	2,5298	9,9443
8	Singapura	17	5	1,5811	10,6677	22,2835
9	Thailand	20	6	1,5811	7,6026	19,1282
10	Vietnam	29	8	10,7935	2,2804	10,1762

9. Posisi cluster selengkapnya terlihat pada Tabel 7.

No	Negara	CBR	CDR	Anggota Cluster		
				C1	C2	C3
1	Brunei Darusalam	27	3		*	
2	Kamboja	38	14			*
3	Indonesia	24	8		*	
4	Laos	43	15			*
5	Malaysia	28	5		*	
6	Myanmar	32	11			*
7	Filipina	30	7		*	
8	Singapura	17	5	*		
9	Thailand	20	6	*		
10	Vietnam	29	8		*	

Terlihat bahwa posisi data sudah tidak mengalami perubahan, sehingga proses iterasi sudah dapat dihentikan.

Hasil akhir yang diperoleh adalah 3 cluster, dengan:

- Cluster pertama memiliki pusat (18,5; 5,5) yang dapat diartikan sebagai kelompok negara-negara dengan dengan CBR rendah dan CDR rendah. Ada 2 negara yang termasuk dalam kelompok ini, yaitu Singapura dan Thailand.
- Cluster kedua memiliki pusat (27,6; 6,2) yang dapat diartikan sebagai kelompok negara-negara dengan dengan CBR sedang dan CDR sedang. Ada 5 negara yang termasuk dalam kelompok ini, yaitu Brunei Darusalam, Indonesia, Malaysia, Filipina dan Vietnam.
- Cluster ketiga memiliki pusat (37,67; 13,33) yang dapat diartikan sebagai kelompok negara-negara dengan dengan CBR tinggi dan CDR tinggi. Ada 3 negara yang termasuk dalam kelompok ini, yaitu Kamboja, Laos dan Myanmar.

Untuk mendapatkan jumlah cluster optimal dengan metode *silhoutte measure*, perlu dilakukan proses clustering dengan jumlah cluster dari 2 sampai 9. Dengan menggunakan KMA, pusat cluster untuk jumlah cluster = 2 adalah:

C_1 = (25; 6) dan C_2 = (37,67; 13,33)

Posisi data pada cluster adalah sebagai berikut:

Data ke-	Berada pada Cluster ke-
1	1
2	2
3	1
4	2
5	1
6	2
7	1
8	1
9	1
10	1

Jarak antar data adalah:

Data ke-	Jarak									
	1	2	3	4	5	6	7	8	9	10
1	-	15,56	5,83	20	2,24	9,43	5	10,2	7,62	5,39
2	15,56	-	15,23	5,1	13,45	6,71	10,63	22,85	19,7	10,82
3	5,83	15,23	-	20,25	5	8,54	6,08	7,62	4,47	5
4	20	5,1	20,25	-	18,03	11,7	15,26	27,86	24,7	15,65
5	2,24	13,45	5	18,03	-	7,21	2,83	11	8,06	3,16
6	9,43	6,71	8,54	11,7	7,21	-	4,47	16,16	13	4,24
7	5	10,63	6,08	15,26	2,83	4,47	-	13,15	10,05	1,41
8	10,2	22,85	7,62	27,86	11	16,16	13,15	-	3,16	12,37
9	7,62	19,7	4,47	24,7	8,06	13	10,05	3,16	-	9,22
10	5,39	10,82	5	15,65	3,16	4,24	1,41	12,37	9,22	-

Untuk data pertama, dihitung rata-rata jarak antara X_1 dengan semua data yang ada dicluster pertama (karena data pertama terletak di cluster pertama), yaitu jarak terhadap data ke-3, 5, 7 – 10. Nilai a_1 merupakan rata-rata jarak tersebut, yaitu: a_1 = (5,83 + 2,24 + 5 + 10,2 + 7,62 + 5,39)/6 = 6,047. Demikian seterusnya,diperoleh nilai a_i sebagai berikut.

i	a_i
1	6,047
2	5,904
3	5,667

4	8,402
5	5,382
6	9,207
7	6,422
8	9,583
9	7,098
10	6,092

Demikian pula, untuk data pertama, dihitung rata-rata jarak antara X_1 dengan semua data yang ada dicluster kedua (hanya ada cluster kedua selain cluster pertama), yaitu jarak terhadap data ke-2, 4, dan 6. Nilai rata-rata jarak tersebut = (15,56 + 20 + 9,43)/3 = 14,997. Karena hanya ada 1 cluster diluar cluster pertama, maka diperoleh nilai b_1 = 14,997. Demikian seterusnya sehingga diperoleh b_i sebagai berikut.

i	b_i
1	14,997
2	15,462
3	14,675
4	20,250
5	12,898
6	9,009
7	10,122
8	22,287
9	19,132
10	10,237

Dengan menggunakan persamaan (3), diperoleh nilai s_i sebagai berikut.

i	s_i
1	0,597
2	0,618
3	0,614
4	0,585
5	0,583
6	-0,022
7	0,366
8	0,570
9	0,629
10	0,405
Rata-rata	***0,495***

Sebagai contoh, pada data pertama, nilai $a_1 < b_1$, sehingga diperoleh $s_1 = 1 - \frac{6{,}047}{14{,}997} = 1 - 0{,}403 = 0{,}597.$

Sehingga nilai rata-rata s_i untuk 2 cluster adalah $\tilde{s}_2 = 0{,}495.$ Demikian seterusnya, proses tersebut dilakukan untuk jumlah cluster 3 sampai 9, hasilnya dapat dilihat pada Tabel.

Jumlah cluster	$\tilde{s}_k$
2	0,495
3	0,418
4	0,299
5	0,131
6	0,159
7	0,185
8	0,193
9	0,105

Terlihat bahwa nilai $\tilde{s}_2$ memiliki angka terbesar, yaitu 0,495, sehingga jumlah cluster yang optimal untuk kasus tersebut adalah 2.