Aplikasi Metode *Sillhouette Coefficient*, Metode *Elbow* dan Metode *Gap Staticstic* dalam Menentukan *K* Optimal pada Analisis *K-Medoids*

Hilda Lailatul Ramadhania[1], Widiarti[1,*], La Zakaria[1] dan Nusyirwan[1]

[1]Jurusan Matematika, Fakultas MIPA, Universitas Lampung
Jl. Sumantri Brojonegoro no 1, Bandar Lampung, Indonesia

widiarti.1980@fmipa.unila.ac.id

Abstrak

Metode *K-Medoids* merupakan analisis klaster metode non hierarki dimana diperlukan informasi jumlah klaster yang tepat. Data yang digunakan dalam penelitian ini menggunakan data simulasi dari referensi data persentase rumah tangga menurut sumber air minum. Data simulasi yang digunakan menggunakan distribusi normal multivariat, sehingga data simulasi memungkinkan adanya data yang negatif. Dalam penelitian ini dilakukan dua opsi pada hasil data yang negatif yaitu dijadikan nol dan dimutlakkan. Metode dalam penentuan jumlah klaster yang optimal menggunakan metode *Sillhouette Coefficient*, metode *Elbow* dan metode *Gap Statistic*. Rata-rata nilai *Dunn Index* dari data pada opsi yang di nolkan menghasilkan nilai *Dunn index* paling besar pada penentuan jumlah klaster menggunakan metode *Gap Statistic* yaitu sebesar 0,125734, sedangkan pada data opsi kedua rata-rata *Dunn Index* paling besar pada penentuan jumlah klaster optimal menggunakan metode *Sillhouette Coefficient* yaitu sebesar 0,113315.

Kata Kunci: K-Medoids; Sillhouette Coefficient, Elbow, Gap Statistic, Dunn index

Abstract

The *K-Medoids* method is a non-hierarchical cluster analysis method where information on the exact number of clusters is required. The data used in this study uses simulation data from reference data on the percentage of households according to drinking water sources. The simulation data used uses a multivariate normal distribution, so that the simulation data allows for negative data. In this study, two options were carried out on negative data results, namely being zero and absolute. The method in determining the optimal number of clusters used the Sillhouette Coefficient method, the Elbow method and the Gap Statistics method. The average Dunn Index value from the data on the zeroed option produces the largest Dunn Index value in determining the optimal number of clusters using the Gap Statistic method, which is 0,125734, while in the second option data, the Dunn Index average is greatest in determining the number of clusters optimally using the Sillhouette Coefficient method, which is 0,113315.

Keywords: K-Medoids; Sillhouette Coefficient, Elbow, Gap Statistic, Dunn index

1. Pendahuluan

Analaisis klaster adalah teknik analisis untuk membagi kelompok utama individu atau objek menjadi beberapa bagian. Secara khusus, analisis klaster adalah mengelompokan sample entitas (individu atau objek) kedalam sejumlah kecil kelompok berdasarkan kesamaan diantara entitas [1]. Proses pengklasteran dilakukan dengan dua metode, yaitu metode hierarki dan metode non hierarki [2]. Metode analisis klaster non hierarki yang paling umum diguanakan adalah algoritma *K-Means*. *K-Means* adalah metode pengklasteran berbasis jarak yang membagi data kedalam sejumlah klaster dan berlaku pada atribut numerik. Algoritma *K-Means* memiliki kelemahan yaitu, tidak robust terhadap pencilan karena menggunakan nilai rata-rata sebagai pusat kelompoknya. Maka dikembangkan metode *K-Medoids* yang merupakan varian umum dari metode *K-Means*. Perbedaan metode *K-Medoids* dari Metode *K-Means* yaitu pada pemilihan *medoid* atau nilai tengah sebagai pusatnya. Akan tetapi metode *K-Medois* memiliki permasalahan dalam penentuan jumlah kelompok sebelum dilakukan analisis [3].

Beberapa penelitian mengenai perbandingan penentuan jumlah *k* optimal diantaranya Utami & Saputro [3] melakukan penelitian pengelompokan data yang memuat pencilan dengan kriteria *Elbow* dan koefisien *Sillhouette* (algoritma *K-Medoids*) yang menghasilkan 3 kelompok dengan menggunakan evaluasi klaster dari nilai koefisien *Sillhouette* didapatkan hasil sebesar 0,6409981. Kemudian penelitian Dewi & Paramita than [4] mengenai analisis perbandingan metode *Elbow* dan *Sillhouette* pada algoritma *clustering K-Medoids* dalam pengelompokan produksi

kerajinan Bali menggunakan evaluasi klaster *Davies Bouldin Index* (DBI). Serta penelitian dari Widjaja & Oetama [5] mengenai *K-Means Clustering* Video *Trending* di Youtube Amerika Serikat dengan hasil penelitiannya menyarankan untuk membutuhkan penelitian selanjutnya menentukan *k* atau jumlah klaster sebelum menjalankan algoritmanya.

Beberapa metode yang biasanya digunakan dalam penentuan jumlah klaster yang tepat, diantaranya yang paling umum yaitu, metode *Elbow*, metode koefisien *Sillhouette*, dan *Gap Statistic*. Setiap metode mempunyai kelebihan dan kekurangan, sehingga perlu ketepatan dalam memadukan metode *clustering* yang digunakan, metode untuk menentukan jumlah klaster yang tepat dan struktur data serta ukuran data [4].

2. Metodologi dan Materi

2.1 Data Penelitian

Data penelitian yang digunakan adalah data penelitian Andini, dkk. [6] yang diambil dari Badan Pusat Statistik Sumatera Utara, yaitu persentase rumah tangga Sumatra Utara pada tahun 2015 yang disimulasikan dengan menggunakan *software Rstudio* versi 4.0.3. Jumlah variabel yang digunakan sebanyak lima dengan keterangan setiap variabel sebagai berikut:

X1 = persentase rumah tangga menurut sumber air minum kemasan,
X2 = persentase rumah tangga menurut sumber air minum ledeng,
X3 = persentase rumah tangga menurut sumber air minum pompa,
X4 = persentase rumah tangga menurut sumber air minum sumur,
X5 = persentase rumah tangga menurut sumber air minum mata air.

Lalu membangkitkan pencilan sebesar 10% pada data yang dibangkitkan tanpa merubah jumlah objek pada data yang dibangkitkan. Banyaknya objek (*n*) yang dibangkitkan terdiri atas 30, 50, 100, 150, 210, 370, 500, 750, 885 dan 1000 data bangkitan.

Tahapan data simulasi yang digunakan sebagai berikut:

a) Membangkitkan data berdistribusi normal dengan lima variabel bebas yang masing-masing variabel bebas memiliki distribusi yaitu $X_1 \sim N(22.39, 18.31)$, $X_2 \sim N(12.70, 16.58)$, $X_3 \sim N(14.74, 16.24)$, $X_4 \sim N(22.12, 18.37)$, $X_5 \sim N(18.59, 18.38)$. Sehingga diperoleh: $\mu = (22.39, 12.70, 14.74, 22.12, 18.59)$
dan
$$\boldsymbol{\Sigma} = \begin{bmatrix} 267.913 & -4{,}533 & -37.348 & 136.347 & -86.482 \\ -4.533 & 245.135 & -23.734 & 77.280 & 8.643 \\ -37.348 & -23.734 & 317.808 & -50.920 & -100.754 \\ 136.347 & 77.280 & -50.920 & 493.681 & 8.593 \\ -86.482 & 8.643 & -100.754 & 8.593 & 389.153 \end{bmatrix} \quad (1)$$
b) Membangkitkan data berdistribusi normal multivariat dengan lima variabel bebas untuk setiap *n* sehingga $X \sim \boldsymbol{N}_5(n, \mu, \boldsymbol{\Sigma})$ dengan $\mu = (22.39, 12.70, 14.74, 22.12, 18.59)$ dan Σ seperti pada Persamaan (1).
c) Membangkitkan data pencilan sebesar 10% dari data yang dibangkitkan dengan distribusi normal (10, 1).
d) Memasukkan pencilan ke dalam data tanpa merubah jumlah data yang dibangkitkan. Ketentuan data ke-1 sampai ke-*p* sebagai pencilan dengan *p* = jumlah pencilan.

2.2 Uji Asumsi Klaster

Ada dua asumsi yang harus dipenuhi analisis klaster, yaitu sampel representatif dan tidak adanya multikolinieritas antar variabel [1].

a) Sample Representatif

Sampel representatif adalah sampel yang diambil dapat mewakili populasi yang ada. Karena keberadaan pencilan mengakibatkan terdapatnya sampel yang tidak mewakili populasi. Maka, perlunya metode yang *robust* terhadap pencilan agar sampel representatif. Metode yang dapat digunakan pada kasus multivariat untuk mendeteksi pencilan adalah pengukuran jarak Mahalanobis [7]. Pengukuran jarak kuadrat Mahalanobis objek ke-*i* dapat dihitung dengan rumus sebagai berikut:

$$d_{MD}^2(i) = (x_i - \bar{x})' \, \Sigma^{-1} (x_i - \bar{x}) \quad (2)$$

dengan,
$d_{MD}^2(i)$ = jarak kuadrat Mahalanobis objek pada pengamatan ke-i
x_i = vektor data objek pengamatan ke-i berukuran $p \times 1$
$\bar{x}$ = vektor rata-rata dari tiap variabel berukuran $p \times 1$
$\boldsymbol{\Sigma}$ = matrix kovarian berukuran p x p, dimana p banyaknya variabel.
Pengamatan ke-i terindifikasi pencilan jika,

$$d_{MD}^2(i) > x_{p,1-a}^2 \quad (3)$$

dimana $x_{p,1-a}^2$ merupakan batas pencilan dengan probabilitas $1 - \alpha$.

b) Tidak Terjadi Multikolinieritas

Nilai VIF (*Variance inflation factor*) dapat mengukur seberapa erat hubungan antar variabel. Jika nilai VIF < 10 artinya data tersebut tidak mengandung multikolinieritas [8]. Adapun rumus untuk mengetahui nilai VIF adalah sebagai berikut:

$$VIF_j = \frac{1}{1 - R_j^2} \quad (4)$$

dengan, R_j^2 adalah koefisien determinasi variabel dependen X_j dengan variabel bebas lainnya selain variabel ke-*j*.

2.3 Metode Penentuan Jumlah *K* Optimal

2.3.1 Metode *Sillhouette Coefficient*

Metode *Sillhouette* merupakan metode yang digunakan untuk menentukan jumlah klaster dengan melakukan pendekatan nilai rata-rata metode *Sillhouette* untuk menduga kualitas klaster yang terbentuk [9]. Untuk menghitung nilai *Silhouette coefficient*, diperkulan perhitungan nilai *sillhouette index* dari sebuah data ke-*i*. Nilai *Silhouette coefficient* didapatkan dengan mencari nilai maksimal dari nilai *Sillhouette index* global dari jumlah klaster 2 sampai jumlah klaster *n-1* seperti pada persamaan berikut:

$$SC = maks_k SI\ (k) \quad (5)$$

dengan, *SC = Sillhouette Coefficient*, *SI = Sillhouette Index Global* dan *k* = jumlah klaster.

Nilai *SI* dari sebuah data ke-1, ada 2 komponen yaitu a_i dan b_i. Nilai $a_i$adalah rata-rata jarak ke-*i* terhadap semua data lainnya dalam satu klaster. Sedangkan b_i didapatkan dengan menghitung rata-rata jarak data ke-*i* terhadap semua data dari klaster lainnya yang tidak satu klaster dengan data ke-*i*, lalu diambil yang terkecil [10]. Berikut persamaan untuk menghitung nilai a_i.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C, i \neq j} d(i,j) \quad (6)$$

dengan, C_i = klaster ke-*i*, $d(i,j)$ = jarak objek ke-*i* dengan objek lainnya pada satu klaster *j*.
Kemudian menghitung nilai b_i dengan persamaan sebagai berikut :

$$b(i) = \min \frac{1}{|C_k|} \sum_{j \epsilon C_k} d(i,j) \quad (7)$$

dengan, C_k = jumlah data pada klaster *k*, $d(i,j)$ = jarak objek ke-*i* dengan objek *j* pda klaster *k*.
Nilai *Sillhouette Index* data ke-*i* dalam klaster *j*, SI_i^j, diperoleh dengan menggunakan persamaan

$$SI_i^j = \frac{b_i^j - a_i^j}{\max\{a_i^j, b_i^j\}} \quad (8)$$

dengan, b_i^j = rata-rata jarak ke-*i* terhadap semua data yang tidak dalam satu klaster dengan data ke-*i*, a_i^j = rata-rata jarak data ke-*i* terhadap semua data dalam satu klaster. Adapun nilai rata-rata *sillhouette Index cluster j*, SI_j dihitung menggunakan persamaan berikut:

$$SI_j = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_i^j \quad (9)$$

dengan, m_j = jumlah data dalam klaster ke-*j* dan *i* = *index* data (*i* = 1,2,... m_j). Lebih jauh nilai rata-rata *Sillhouette Index* dari dataset, *SI*, diberikan oleh persamaan

$$SI = \frac{1}{k}\sum_{j=1}^{k} SI_j \tag{10}$$

dengan *k* adalah jumlah klaster. Hasil nilai rata-rata *Sillhouette coeficient* yang paling besar yaitu $0.7 < SC \leq 1$ yang artinya terdapat ikatan yang sangat baik antara objek dan kelompok yang terbentuk.

2.3.2 Metode *Elbow*

Metode *Elbow* merupakan metode untuk menentukan jumlah klaster yang tepat melalui persentase hasil perbandingan antara jumlah klaster yang akan membentuk siku pada suatu titik. Jika nilai klaster pertama dengan nilai klaster kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar dapat menjadi jumlah nilai klaster yang tepat untuk digunakan [4]. Untuk mendapatkan perbandingannya adalah dengan menghitung *sum square error* (*SSE*) dari masing-masing nilai klaster *k* menggunakan persamaan berikut:

$$SSE = \sum_{k=1}^{n}\sum_{x_i} |x_i - c_k|^2 \tag{11}$$

Dengan x_i menyatakan objek data ke-*i* dan c_k adalah pusat klaster ke-*i*.

2.3.3 Metode *Gap Statistic*

Gap Statistic merupakan metode untuk menduga kelompok optimum pada analisis klaster. Teknik ini berdasar pada perubahan dispersi dalam klaster dengan peningkatan jumlah kelompok dari data [11]. Berikut adalah *Gap Statistic* untuk *k* tertentu:

$$Gap(k) = \left[\frac{1}{B}\right] \Sigma_b\{\log(W_{kb}^*) - \log(W_k) \tag{12}$$

dimana *B* adalah *resampling* (dari data simulasi) dengan pengambilan sebanyak *B* kali dengan distribusi *uniform*. Tahapan penentuan jumlah klaster optimal menggunakan metode *gap statistic* sebagai berikut [12]*:*

a. Mengelompokkan data dan mengubah-ubah banyaknya kelompok mulai dari $k = 1,2, \dots, n,$ dan hitung total variasi *intracluster* W_k, dengan $k = 1,2, \dots, n$.
b. Hasilkan kumpulan data referensi *B* dengan distribusi referensi *uniform*. Klasterkan masing-masing dari kumpulan data referensi ini dengan berbagai jumlah kelompok $k = 1, \dots, k_{max}$ dan menghitung total variasi *intracluster* $W_{kb.}$
c. Hitung estimasi *Gap Statistic* sebagai penyimpangan nilai W_k yang diamati dari W_{kb} dan juga hitung standar deviasinya.
d. Pilih jumlah klaster sebagai nilai terkecil dari *k* sehingga *Gap Statistic* berada dalam satu standar deviasi dari celah pada *k+1*.

2.4 Metode Pengklasteran

Metode *K-Medoids* dikenal juga sebagai PAM (*Partitioning Around Medoids*) merupakan salah satu metode yang digunakan untuk proses klastering. Dalam metode ini data yang terdiri dari *n* objek di partisi menjadi *k* klaster dimana jumlah $k \leq n$. Maka dari itu tujuannya adalah untuk menemukan *k* objek tersebut [13]. pengelompokan didasarkan pada kemiripan antar objek yang diukur dengan menggunakan ukuran jarak (*distance*). Ukuran jarak dengan nilai yang lebih besar menunjukan kesamaan yang lebih rendah [1]. Pada penelitian ini pengukuran kemiripan yang digunakan dalam proses pengelompokan adalah *Euclidean Distance*. Rumus jarak *Euclidean* dinyatakan sebagai berikut:

$$d_{ij} = \sqrt{\sum_{t=1}^{n}(x_{ik} - x_{jk})^2} \tag{13}$$

dengan, d_{ij} = jarak antara objek ke-*i* dan objek ke-*j*, *n* = jumlah variabel klaster, x_{ik} = data dari objek ke-*i* pada variabel ke-*k* dan x_{jk} = data dari objek ke-*j* pada variabel ke-*k*.

2.5 *Dunn Index*

Dunn Index adalah salah satu pengukur validitas klaster yang diajukan oleh J. C. Dunn. Luthfi & Wijayanto dalam [14] menjelaskan bahwa *index validitas Dunn* menghitung nilai minimum dari perbandingan antara nilai fungsi *dissimilaritas* antara dua klaster sebagai *separation* dan nilai maksimum dari diameter klaster sebagai *compacness*. Jumlah klaster terbaik ditunjukkan dengan semakin besarnya nilai *Dunn Index* berikut:

$$D = \left\{ \frac{\min\limits_{1\le i\le k} \left(d(c_i, c_j) \right)}{\max\limits_{1\le l\le q} diam(c_k)} \right\} \quad (14)$$

dengan, *D, q*, berturut-turut menyatakan *Dunn Index* dan jumlah klaster. Sedangkan $d(c_i, c_j)$ menyatakan jarak *Euclidean* antar pasangan objek pada klaster *i* dan klaster *j* (*intercluster*). Adapun $diam(c_k)$ adalah jarak *Euclidean* antar objek dengan nilai rata-rata klaster (*intracluster*). *Dunn Index* memiliki rentang nilai dari nol sampai tak hingga.

3. Hasil dan Pembahasan

3.1. Data Penelitian

Data penelitian yang digunakan menggunakan data simulasi dengan distribusi normal multivariat sehingga memungkankan hasilnya terdapat data yang negatif dengan menggunakan parameter dari data asli. Penelitian ini melakukan dua opsi pada data yang negatif, yaitu dengan di nolkan dan di mutlakkan. Berikut adalah data simulasi untuk *n* = 30 dengan μ dan matriks kovarian pada persamaan (1).

Tabel 1. Data dengan *n* = 30 dan proporsi pencilan 10% untuk nilai negatif yang di nol kan

Data Ke-	X1	X2	X3	X4	X5	Data Ke-	X1	X2	X3	X4	X5
1	10,346	0	41,310	5,903	29,303	16	43,243	15,727	0	22,216	0
2	9,670	10,706	22,185	36,682	37,970	17	52,722	0	0	32,270	1,121
3	9,712	12,445	23,099	79,697	14,751	18	40,982	18,465	30,719	47,386	0
4	8,735	13,928	24,454	9,888	16,192	19	54,303	3,893	0	14,449	5,390
5	0	0	16,519	0	36,105	20	12,010	21,949	28,415	10,874	0
6	46,172	6,585	22,122	25,632	0	21	56,350	10,179	0	26,334	0
7	30,571	34,932	15,207	26,039	17,980	22	7,730	0	2,787	12,209	26,051
8	36,042	0	22,826	41,207	9,257	23	9,338	0	44,594	0	14,459
9	1,873	9,639	35,558	6,904	4,461	24	15,592	39,994	1,179	40,063	43,539
10	32,473	18,961	15,139	48,756	18,277	25	52,230	23,051	0	33,259	24,088
11	6,197	0	35,261	21,767	3,330	26	64,421	37,851	24,809	41,301	0
12	0	0	47,804	0	15,244	27	15,869	0	14,781	0	15,706
13	0	14,212	23,199	0	46,885	28	44,622	9,958	0,967	25,781	2,863
14	51,488	0	44,174	42,488	0	29	26,625	21,332	24,343	16,161	6,283
15	28,133	19,586	0	53,881	60,555	30	39,807	4,029	31,130	72,710	37,684

Tabel 2. Data dengan n = 30 dan proporsi pencilan 10% untuk nilai negatif yang di mutlakan

Data Ke-	X1	X2	X3	X4	X5	Data Ke-	X1	X2	X3	X4	X5
1	10,346	19,538	41,310	5,903	29,303	16	43,243	15,727	2,670	22,216	3,698
2	9,670	10,706	22,185	36,682	37,970	17	52,722	2,991	19,346	32,270	1,121
3	9,712	12,445	23,099	79,697	14,751	18	40,982	18,465	30,719	47,386	2,654
4	8,735	13,928	24,454	9,888	16,192	19	54,303	3,893	10,401	14,449	5,390
5	6,028	1,841	16,519	6,018	36,105	20	12,010	21,949	28,415	10,874	18,258
6	46,172	6,585	22,122	25,632	4,827	21	56,350	10,179	1,432	26,334	19,535
7	30,571	34,932	15,207	26,039	17,980	22	7,730	9,177	2,787	12,209	26,051
8	36,042	4,860	22,826	41,207	9,257	23	9,338	16,917	44,594	24,104	14,459
9	1,873	9,639	35,558	6,904	4,461	24	15,592	39,994	1,179	40,063	43,539
10	32,473	18,961	15,139	48,756	18,277	25	52,230	23,051	13,300	33,259	24,088
11	6,197	28,528	35,261	21,767	3,330	26	64,421	37,851	24,809	41,301	6,912
12	1,278	7,208	47,804	12,512	15,244	27	15,869	1,441	14,781	5,870	15,706
13	8,680	14,212	23,199	13,537	46,885	28	44,622	9,958	0,967	25,781	2,863
14	51,488	5,566	44,174	42,488	6,910	29	26,625	21,332	24,343	16,161	6,283
15	28,133	19,586	24,855	53,881	60,555	30	39,807	4,029	31,130	72,710	37,684

3.2 Uji Asumsi Klaster

3.2.1 Uji Sample Representatif dengan Pendekatan Pencilan

Hasil uji sampel representatif dengan pendekatan pencilan menggunakan jarak Mahalanobis disajikan dalam Tabel 3 berikut

Tabel 3. Hasil perhitungan jarak Mahalanobis data dengan nilai negatif yang di nol kan

Objek Ke-	d_{MD}^2	Objek Ke-	d_{MD}^2	Objek Ke-	d_{MD}^2
1	4,8211	11	3,8114	21	3,6039
2	2,2238	12	4,4545	22	5,1245
3	15,2026	13	5,8921	23	3,9896
4	1,4125	14	7,6415	24	8,4768
5	4,2091	15	9,0752	25	3,6574
6	1,7273	16	3,2218	26	10,1589
7	4,3241	17	4,9800	27	2,5522
8	2,1380	18	3,0717	28	2,6727
9	3,6188	19	4,9026	29	1,8567
10	1,3636	20	4,6763	30	10,1395

Tabel 4. Hasil perhitungan jarak mahalanobis data dengan nilai negatif yang di mutlakan

Objek Ke-	d_{MD}^2	Objek Ke-	d_{MD}^2	Objek Ke-	d_{MD}^2
1	5,3331	11	5,7927	21	4,3662
2	2,5152	12	5,0354	22	5,0818
3	15,3387	13	4,8170	23	3,2312
4	1,4874	14	6,8702	24	10,2682
5	5,1087	15	9,5824	25	3,0987
6	1,9164	16	3,3873	26	9,5650
7	4,0513	17	3,2793	27	3,3975
8	1,8840	18	2,8378	28	4,1007
9	4,0797	19	4,4777	29	1,5347
10	1,7763	20	1,8469	30	8,9387

Pengamatan ke-i teridentifikasi pencilan jika $d_{MD}^2(i) > x_{p,1-\alpha}^2$ dengan p merupakan banyaknya variabel yang diteliti dan nilai α yang digunakan sebesar 10%. Berdasarkan kasus ini, banyaknya variabel yang diteliti yaitu 5 variabel dan nilai α yang digunakan sebesar 0.010. Sehingga diperoleh nilai x_p^2 = 5,(1 – 0.010) yaitu sebesar 9.236. Berdasarkan pendeteksian pencilan dengan membandingkan hasil jarak Mahalanobis tiap objek dan nilai $x_{5,0.90}^2$. Diketahui bahwa terdapat 3 objek yang merupakan pencilan, yaitu objek ke-3, objek ke-26 dan objek ke-30 pada data di Tabel 3, dan terdapat 4 objek yang merupakan pencilan, yaitu objek ke-3, objek-15, objek ke-24 dan objek ke-26 pada data Tabel 4.

3.3 Uji Asumsi Tidak Terjadi Multikolinieritas

Uji asumsi nonmultikolinieritas dilakukan untuk melihat apakah data yang digunakan terdapat hubungan linier antar variabel atau tidak. Berdasarkan data pada Tabel 5 dan Tabel 6 diperoleh nilai VIF menggunakan Persamaan (1) sebagai berikut:

Tabel 5. Nilai VIF dari setiap variabel untuk data negatif yang di nol kan

VIF				
X1	**X2**	**X3**	**X4**	**X5**
2,42275	1,175232	1,52099	1,52099	1,729737

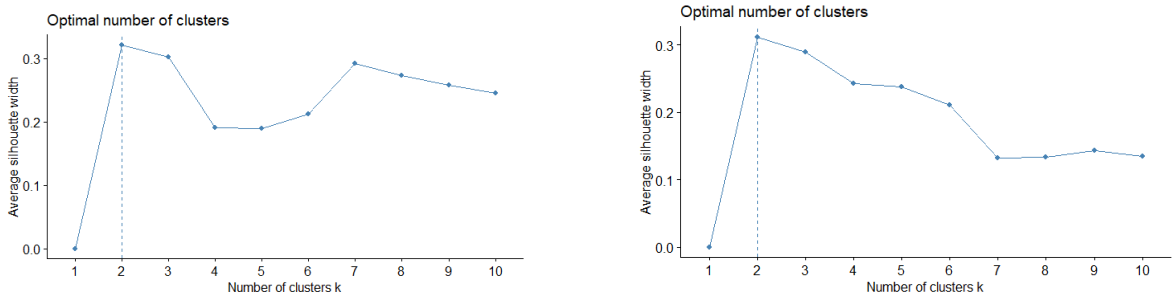
Tabel 6. Nilai VIF dari setiap variabel untuk data negatif yang di mutlakkan

VIF				
X1	**X2**	**X3**	**X4**	**X5**
1,708228	1,028202	1,240678	1,33434	1,371249

Berdasarkan Tabel 5 dan 6, diperoleh nilai VIF variabel X1, X2, X3, X4 dan X5 < 10, maka dapat disimpulkan tidak terjadi multikolinieritas pada data tersebut. Sehingga asumsi tidak adanya multikolinieritas terpenuhi.

3.4 Penentuan Jumlah Klaster Optimal

a) Metode Koefisien *Sillhouette*

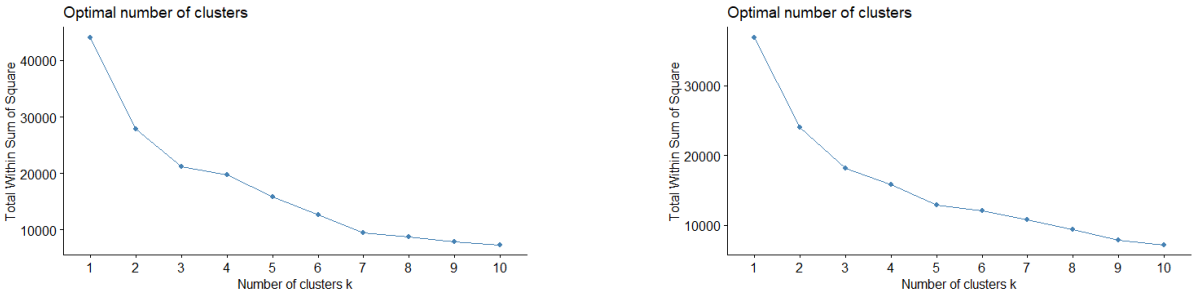


(a) (b)

Gambar 1. Plot hasil nilai *Sillhouette coefficient* (a) untuk nilai negatif yang di nol kan (b) untuk nilai negatif yang di mutlakkan.

Dihasilkan klaster optimal menggunakan metode *Sillhouette* sebanyak 2 klaster pada kedua data yang digunakan untuk jumlah *n* sebanyak 30 data.

b) Metode *Elbow*

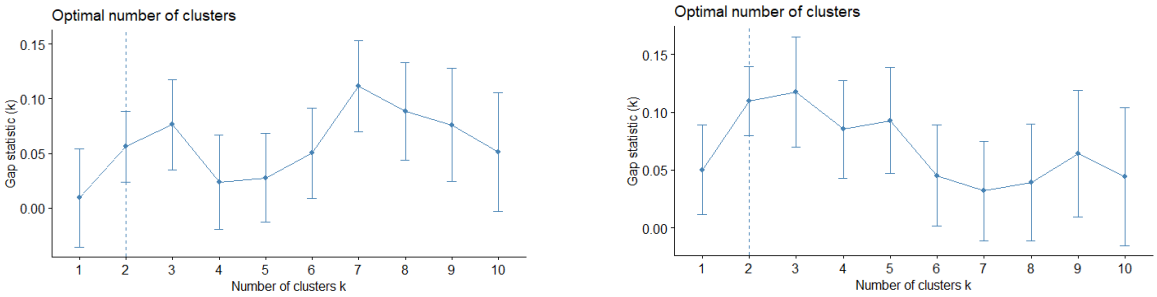


(c) (d)

Gambar 2. Plot hasil nilai *SSE* (c) untuk nilai negatif yang di nol kan (d) untuk nilai negatif yang di mutlakkan

Dari hasil grafik Metode *Elbow* dapat dilihat bahwa yang mengalami patahan atau grafik berbentuk siku pada saat klaster sebanyak 3. Sehingga jumlah klaster optimal sebanyak 3 klaster pada kedua data yang digunakan untuk jumlah *n* sebanyak 30 data.

a) Metode *Gap Statistic*



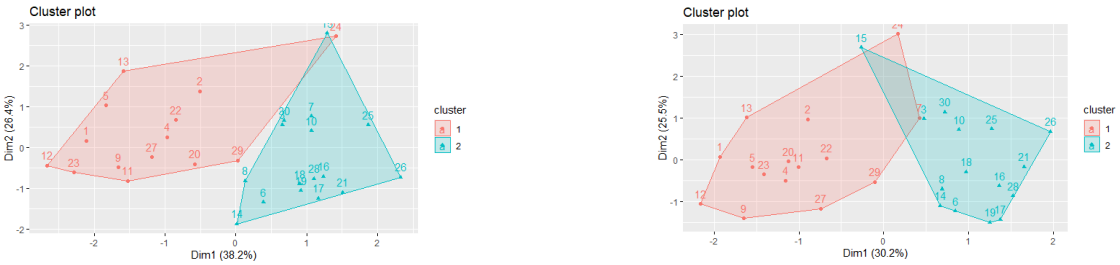
(e) (f)

Gambar 3. Plot hasil *Gap Statistic*(e) untuk nilai negatif yang di nol kan (f) untuk nilai negatif yang di mutlakkan

Dihasilkan jumlah klastet optimal pada grafik metode *Gap Statistic* sebanyak 2 klaster pada kedua data yang digunakan untuk jumlah *n* sebanyak 30 data.

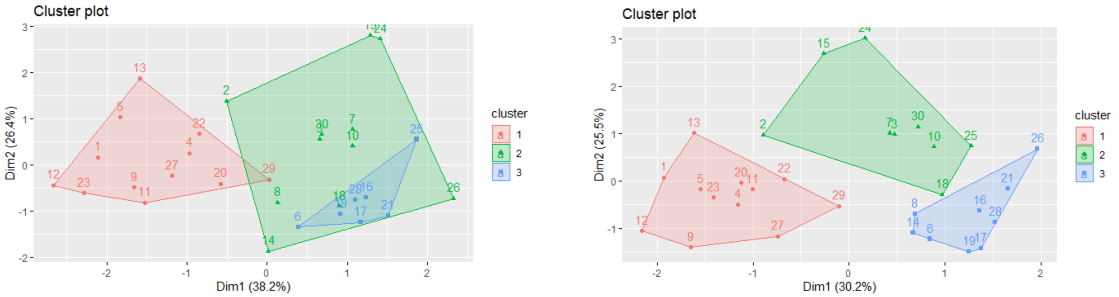
3.5 Analisis Klaster Mengguakan Metode *K-Medoids*

Hasil pengklasteran dengan pengelompokkan optimal dari setiap metode pada data simulasi dengan jumlah *n* 30 data disajikan dalam Gambar 4 dan Gambar 5.



(g) (h)

Gambar 4. Hasil Pengelompokan menggunakan metode *K-Medoids* pada saat klaster optimal sebanyak 2 klaster (g) untuk nilai negatif yang di nol kan (h) untuk nilai negatif yang di mutlakkan.



Gambar 5. Hasil Pengelompokan menggunakan metode *K-Medoids* pada saat klaster optimal sebanyak 3 klaster (i) untuk nilai negatif yang di nol kan (j) untuk nilai negatif yang di mutlakkan.

3.6 Evaluasi Klaster Berdasarkan Metode Penentuan Jumlah Klaster Optimal

Evaluasi klaster akan dilihat berdasarkan rata-rata nilai *Dunn Index* dari semua data simulasi yang digunakan.

Tabel 7. Hasil nilai *Dunn Index* pada setiap jumlah n data simulasi yang digunakan untuk nilai negatif yang di nol kan.

n	Jumlah Klaster Optimal Berdasarkan Metode					
	Sillhouette	***Dunn index***	***Elbow***	***Dunn index***	***Gap statistic***	***Dunn index***
30	2	0,273304	3	0,266645	2	0,273304
50	10	0,369837	4	0,247971	10	0,369837
100	2	0,149962	2	0,149962	2	0,149962
150	2	0,111326	4	0,081944	2	0,111326
210	4	0,069082	4	0,069082	4	0,069082
370	2	0,075955	3	0,093911	3	0,093911
500	5	0,069625	5	0,069625	5	0,069625
750	2	0,052415	4	0,046965	4	0,046965
885	3	0,027048	3	0,027048	3	0,027048
1000	2	0,046279	3	0,015217	3	0,046279
Rata-rata		**0,124483**		**0,106837**		**0,125734**

Tabel 8. Hasil nilai *Dunn Index* pada setiap jumlah n data simulasi yang digunakan untuk nilai negatif yang di mutlakkan.

n	Jumlah Klaster Optimal Berdasarkan Metode					
	Sillhouette	***Dunn index***	***Elbow***	***Dunn index***	***Gap statistic***	***Dunn index***
30	2	0,325613	3	0,288103	2	0,325613
50	5	0,250183	3	0,181038	5	0,250183
100	2	0,106790	2	0,106790	2	0,106790
150	2	0,078453	3	0,076062	2	0,078453
210	3	0,079231	3	0,079231	3	0,079231
370	3	0,097225	3	0,097225	4	0,06851
500	3	0,056986	3	0,056986	3	0,056986
750	2	0,049924	4	0,050769	4	0,050769
885	2	0,039693	3	0,043595	3	0,043595
1000	2	0,049054	3	0,041246	3	0,041246
Rata-rata		**0,113315**		**0,102105**		**0,110138**

Dihasilkan nilai rata-rata *Dunn Index* dari data simulasi untuk data negatif yang di nol kan menghasilkan nilai *Dunn Index* paling besar menggunakan metode *Gap Statistic*. Sedangkan untuk data negatif yang di mutlakkan menghasilkan nilai rata-rata *Dunn Index* paling besar menggunakan metode *Sillhouette coefficient*.

4. Kesimpulan

Berdasarkan hasil dan pembahasan yang telah dilakukan, dapat disimpulan bahwa penentuan jumlah klaster yang optimal pada masing-masing metode dipengaruhi oleh setiap objek dan jumlah data. Selain itu, berdasarkan nilai rata-rata *Dunn Index* dari keseluruhan data yang dibangkitkan didapatkan nilai rata-rata paling besar menggunakan metode *Gap Statistic* yaitu senilai 0,125734 untuk data negatif yang di nol kan, sedangkan untuk nilai rata-rata *Dunn Index* paling besar menggunakan metode *Sillhouette coefficient* yaitu senilai 0,113315. Nilai *Dunn Index* pada jumlah klaster yang sama akan menghasilkan nilai *Dunn Index* paling besar saat jumlah data lebih kecil. Untuk pengembangan penelitian lebih lanjut, disarankan untuk mencoba beberapa persentase pencilan yang berbeda-beda. Hal ini untuk melihat pengaruh metode dalam penentuan jumlah klaster.

Daftar Pustaka:

[1] Hair, JR. J.F., Black, W.C., Babin, B.J., & Anderson, R.E. 2010. *Multivariate Data Analysis* 7th Edition. Pearson Education Limited, England.

[2] Widyadhana, D., Hastuti, R.B., Kharisudin, I., & Fauzi, F. 2021. Perbandingan Analisis Klaster K-means dan Average Linkage untuk Pengklasteran Kemiskinan di Provinsi Jawa Tengah. Prosiding Seminar Nasional Matematika. Semarang.

[3] Utami, D.S., & Saputro, D.R.S. 2018. Pengelompokan data yang Memuat Pencilan dengan Kriteria Elbow dan Koefisien Sillhouette (Algoritma K-medoids). Prosiding KNPMP III. Surakarta.

[4] Dewi, D.A.I.C. & Paramita, D.A.K. 2019. Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-medoids Dalam Pengelompokan Produksi Kerajinan Bali. *Jurnal Matrix*. **9**(03): 102-109.

[5] Widjaja, K., & Oetama, R.S. 2020. K-Means Clustering Video Trending di Youtube Amerika Serikat. ULTIMA InfoSys. **XI**(02): 78-84.

[6] Andini, N., Ramadhani, T., Rahayu, S.P., & Lindasari, D. 2019. Pengujian Normal Multivariat dan Vektor Mean pada Data Prosentase Rumah Tangga Menurut Sumber Mata Air Minum Provinsi Aceh dan Sumatera Utara Tahun 2015.

[7] Filzmoser, P. 2005. Identification of Multivariate Outliers: a Performance Study. *Austrian Journal of Statistics*. **34**(02): 127-138.

[8] Hidayat, F.P., & Hakim, R.B.F. 2021. Implementasi Metode Clustering K-medoids Dalam Mengelompokan Jumlah Aduan di Kabupaten Sleman, hlm 106-114. Prosiding Seminar Matematika dan Pendidikan Matematika. Yogyakarta.

[9] Dewa, F.A. & Jatipaningrum, M.T. 2019. Segmentasi E-comerace dengan Cluster K-Means dan Fuzzy C-Means. *Jurnal Statistika Industri dan Komputasi*. **04**(01): 53-67.

[10] Petrovic, S. 2006. A Comparison Betweem the Sillhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters. *In 11th Nordic Workshop on Secure IT-systems*.

[11] Arima, C., Hakamada, K., Okamoto, M., & Hanai, T. 2005. Validity Index for Fuzzy K-means Clustering Using the Gap Statistic Method. *Sixteenth International Conference on Genome Informatics*.

[12] Tibshirani, R., Walther, G., & Hastie, T. 2001. Estiming the Number of Cluster in a Data Set Via the Gap Statistic. *J. R. Statist*. **63**(02): 411-423.

[13] Kaufman, L. & Rousseeuw, P.J. 1989. *Finding Groups in Data an Introduction to Cluster Analysis*. Willey, J. & Sons. New Jersey.

[14] Luthfi, E. & Wijayanto, A.W. 2021. Analisis Perbandingan Metode Hierarchical, K-Means, dan K-Medoids Clustering dalam Pengelompokan Index Pembangunan Manusia Indonesia. *Jurnal Inovasi*. **17**(4): 761-773.