PENGELOMPOKAN DATA YANG MEMUAT PENCILANDENGAN KRITERIA *ELBOW*DAN KOEFISIEN *SILHOUETTE* (ALGORITME *K-MEDOIDS*)

Dwi Sari Utami[1)], Dewi Retno Sari Saputro[2)]
[1,2)]Program Studi Matematika Universitas Negeri Sebelas Maret
dsutami12@gmail.com, dewiretnoss@staff.uns.ac.id

Abstrak

Analisis kelompok adalah metode statistika multivariat yang bertujuan untuk mengelompokkan objek pengamatan yang memiliki kemiripan (karakteristik sama). Terdapat dua metode pengelompokan dalam analisis kelompok yaitu metode pengelompokan hierarchical (hierarki) dan nonhierarchical (nonhierarki). K-medoids merupakan metode pengelompokan nonhierarki yang mempartisi n data ke dalam k kelompok yang memiliki karakteristik sama dan menggunakan medoid (median) sebagai pusat kelompoknya. Dengan demikian, k-medoids ini robust terhadap adanya data pencilan. Dalam proses pengelompokan digunakan algoritme k-medoidsdengan kriteria elbow dan validasinya dengan koefisien silhouette. Kriteria elbow digunakan dengan melihat plot jumlah kuadrat sesatan (JKS) dari beberapa jumlah kelompok(k). Jika terbentuk siku (elbow) untuk nilai JKS pada suatu nilai k, maka nilai tersebut menjadi banyaknya kelompok yang akan dibentuk. Koefisien silhouette berada antara-1dan 1. Pada artikel ini dilakukan kajian kriteria elbow dan koefisien silhouette dengan algoritme k-medoids untuk pengelompokan data yang memuat pencilan dan penerapannya pada kasus demam berdarah di Indonesia tahun 2016. Kajian menunjukkan bahwa pengelompokan kasus demam berdarah pada 34 provinsi di Indonesia tahun 2016 menghasilkan 3 kelompok dengan nilai koefisien silhouette sebesar 0.6409981.

***Kata kunci** : k-medoids;kriteria elbow;koefisien silhouette.*

1. PENDAHULUAN

Statistika merupakan pengetahuan yang berhubungan dengan cara mengumpulkan atau memperoleh data, menganalisis data, dan menarik kesimpulan berdasarkan kumpulan data. Data menjadi kebutuhan bagi masyarakat baik di kalangan akademis, perusahaan, kesehatan, maupun pemerintahan. Dalam statistika tidak hanya terdapat satu variabel melainkan juga bisa lebih dari satu variabel. Analisis yang membahas lebih dari satu variabel secara bersamaan disebut analisis multivariat. Salah satu teknik yang dibahas dalam analisis tersebut adalah *clustering*(pengelompokan).

Analisis kelompok merupakan suatu analisis multivariat yang digunakan untuk mengelompokkan objek pengamatan menjadi beberapa kelompok berdasarkan ukuran kemiripan antarobjek. Tujuan analisis kelompokmempartisi himpunan objek menjadi dua kelompok atau lebih berdasarkan kesamaan karakteristik khusus yang dimilikinya. Pengelompokan data diperlukan untuk menyederhanakan permasalahan dengan melakukan pengelompokan berdasarkan karakteristik variabel ke dalam sejumlah kelompok yang relatif lebih homogen untuk memudahkan analisis.

Menurut Johnson dan Wichern (2002), terdapat dua metode pengelompokan dalam analisis kelompok yaitu *hierarchical* (hierarki) dan *nonhierarchical* (nonhierarki). Metode pengelompokan hierarki adalah suatu metode pengelompokan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kemiripan terdekat. Kemudian proses dilanjutkan ke objek lain yang memiliki kedekatan kedua dan seterusnya hingga kelompok akan membentuk seperti pohon dimana terdapat hierarki (tingkatan) yang jelas antarobjek dari yang paling mirip sampai tidak mirip. Sedangkan metode pengelompokan nonhierarki dimulai dengan menentukan jumlah kelompok yang diinginkan terlebih dahulu. Kemudian proses pengelompokan baru dilakukan. Dalam analisis kelompok, syarat yang harus diperhatikan adalah sampel yang digunakan harus dapat mewakili populasi dan tidak adanya multikolinearitas (Santoso, 2010).

Pada proses pengelompokan terdapat masalah terhadap hasil yang dicapai, salah satunya adalah kepekaan pada *outlier* (pencilan). Menurut Barnett dan Lewis (1994), pencilan merupakan pengamatan yang tidak mengikuti sebagian besar pola dan terletak jauh dari pusat data. Pencilan akan menimbulkan penyimpangan struktur hasil pengelompokan sehingga tidak merepresentasikan populasi dengan benar. *K-means* dan *k-medoids*merupakan metode pengelompokan nonhierarki yang mempartisi data ke dalam kelompok sehingga data yang memiliki karakteristik sama dikelompokkan ke dalam satu kelompok dan data yang mempunyai karakteristik berbeda dikelompokkan ke dalam kelompok yang lain. Menurut Han dan Kamber (2012), algoritme*k-means* sensitif terhadap pencilan. Algoritme *k-means* tidak *robust* terhadap pencilan karena menggunakan nilai rata-rata (*mean*) sebagai pusat kelompoknya.Kemudian menurut Kaufman dan Rousseeuw (1987), algoritme*k-medoids*dapat mengatasi kelemahan tersebut. *K-medoids* merupakan metode pengelompokan dengan menggunakan*medoid*sebagai pusat kelompok. *Medoid*adalah objek yang letaknya terpusat dalam suatu kelompoksehingga algoritme *k-medoids* lebih*robust* dibanding dengan algoritme *k-means*. Algoritme *k-medoids* memiliki permasalahan dalam penentuan jumlah kelompok sehingga dilakukan kriteria *elbow* untuk menentukan jumlah kelompok terbaik.

Masalah penting lainnya dalam pengelompokan adalahvalidasi hasil pengelompokan untuk memperoleh partisi yang paling sesuai dengan data dasar. Terdapat tiga pendekatan utama dalam melakukan validasi kelompok yaitu kriteria eksternal, kriteria internal, dan kriteria relatif. Validasi dengan pendekatan kriteria internal lebih sering digunakan karena lebih sederhana dan mudah. Salah satu contoh validasi dengan pendekatan kriteria internal adalahkoefisien *silhouette.* Pada artikel ini dilakukan kajian tentang pengelompokan data yang memuat pencilan dengan kriteria *elbow* dan koefisien *silhouette* pada algoritme *k-medoids*dan penerapannya pada kasus demam berdarah di Indonesia tahun 2016.

2. METODE PENELITIAN

Penelitian ini merupakan penelitian berdasarkan teori dan penerapan. Penelitian berdasarkan teori yaitu melakukan kajian tentang pengelompokan data yang memuat pencilan dengan kriteria *elbow* dan koefisien *silhouette* pada

algoritme *k-medoids*.Sementara penerapannya dengan data kasus demam berdarah di Indonesia tahun 2016. Metode yang digunakan adalah studi literatur dengan mempelajari dan menurunkan ulang beberapa materi terkait tentang pencilan dan metode penyelesaiannya. Literatur berupa jurnal referensi dari berbagai situs pendukung di internet dan *textbook*. Langkah-langkah yang dilakukan dalam penelitian ini diuraikan sebagai berikut.

a. Melakukan kajian beberapa metode pendeteksian pencilan dalam kasus data multivariat, dalam hal ini kriteria *elbow* dan koefisien *silhouette* (algoritme *k-medoids*).
b. Menuliskan hasil kajian pada (a) dan melakukan interpretasi.
c. Mengidentifikasipengaruh pencilan dalam pengelompokan.
d. Melakukan kajianalgoritme *k-medoids* untuk data pencilan, menuliskan ulang dalam bentuk flowchart, mengidentifikasi langkah dan melakukan interpretasi.
e. Mengidentifikasi validasi kelompok dengan metode koefisien*silhouette*.
f. Menuliskan hasil interpretasi (a)-(e) dan melakukan penarikan simpulan.
g. Menerapkan algoritme *k-medoids*dan validasinya dengan metode koefisien *silhouette* pada data yang telah disebutkan sebelumnya.
h. Membuat pengelompokannya dengan gambar.
i. Melakukan analisis hasil, interpretasi, dan simpulan.

3. HASIL PENELITIAN DAN PEMBAHASAN

Pada penelitian ini dibahas pencilan dan cara mendeteksinya, pengelompokan data yang mengandung pencilan dengan algoritme *k-medoids*, serta validasinya dengan koefisien *silhouette*.

a. Pencilan dan Cara Mendeteksinya

Pencilan adalah objek yang menyimpang dari objek lainnya dalam suatu himpunan data. Pada proses pengelompokan, adanya pencilan dapat memengaruhi hasil analisis yang dicapai. Oleh karena itu, pencilan pada data perlu dideteksi. Beberapa metode yang dapat digunakan untuk mendeteksi pencilan yaitu metode grafis, *boxplot*, standardisasi, dan jarak kuadrat Mahalanobis.Pada kasus multivariat, metode yang dapat digunakan untuk mendeteksi pencilan adalah pengukuran jarak kuadrat Mahalanobis (Folzmiser, 2005).Jarak kuadrat Mahalanobis merupakan ukuran jarak yang melibatkan kovariansi atau korelasi antar variabel. Pengukuran jarak kuadrat Mahalanobis dinyatakan dengan

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

dengan $d_{MD}^2(i)$ adalah jarak kuadrat Mahalanobis objek pada pengamatanke-i, $\boldsymbol{x_i}$ adalah vektor data objek pada pengamatan ke-i berukuran $p \times 1$, $\bar{\boldsymbol{x}}$ adalah vektor rata-rata berukuran $p \times 1$, dan Σ adalah matriks kovariansiberukuran $p \times p$ dimana p banyaknya variabel. Pengamatan ke-i disebut pencilan jika

$$d_{MD}^2(i) > \chi_{p,1-\alpha}^2,$$

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

b. Pengelompokan Data yang Memuat Pencilan

Jika pencilan dikelompokkan ke dalam suatu kelompok, maka pencilan tersebut dapat mendistorsi nilai rata-rata dari kelompok (Han dan Kamber, 2012).Algoritme *k-medoids* dapat mengatasi ketika adanya pencilan dalam data. *K-medoids* bekerja dengan *medoid* yaitu memilih suatu titik sebagai titik tengah. *Medoid* dapat diartikan sebagai sebuah objek yang mempunyai rata-rata jarak terkecil ke objek lainnya, dengan kata lain yaitu objek yang terletak di tengah kelompok data (Flowrensia, 2010). Berikut adalah algoritme *k-medoids*.

1) Menentukan jumlah kelompok (k) yang ingin dibentuk.
Metode penentuan jumlah kelompok yang akan dibentuk menggunakan kriteria *elbow*. Kriteria *elbow* adalah suatu metode untuk menghasilkan informasi dalam menentukan jumlah kelompok terbaik dengan cara melihat persentase hasil perbandingan antara jumlah kelompok yang akan membentuk siku pada suatu titik (Madulatha, 2012). Hasil persentase yang berbeda dari setiap nilai ditunjukkan dengan grafik. Untuk memperoleh perbandingannya diukur dengan menghitung jumlah kuadrat sesatan (JKS). Semakin besar jumlah kelompok, maka nilai JKS semakin kecil. JKS ditulis sebagai

$$\text{JKS} = \sum_{i=1}^{n} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})'(\boldsymbol{x_i} - \bar{\boldsymbol{x}}),$$

dengan $\boldsymbol{x_i}$ adalah vektor data objek ke-i berukuran $p \times 1$, $\bar{\boldsymbol{x}}$ adalah vektor rata-rata berukuran $p \times 1$, dan n adalah banyak objek pengamatan.Pada kriteria *elbow,* jumlah kelompok terbaik diambil dari nilai JKS yang mengalami penurunan signifikan berbentuk siku (Kokasih, 2016).Nilai JKS yang awalnya tinggi akan mengalami penurunan secara drastis, kemudian turun secara perlahan sampai nilai JKS tersebut stabil. Jika terlihat penurunan drastis dan terbentuk siku untuk nilai JKS pada suatu nilai k, maka nilai tersebut menjadi banyaknya kelompok yang akan dibentuk.

2) Memilih*medoid* awal secara acak dari objek-objek yang akan dikelompokkan.

3) Menentukan jarak objek *non-medoid* dengan *medoid*pada tiap kelompok dengan perhitungan jarak menggunakan Euclidean*distance.* Euclidean*distance*adalah ukuran kemiripan yang biasa digunakan dalam analisis kelompok. Euclidean*distance*merupakan jarak terpendek (*straight line*) antara dua titik.Euclidean*distance*ditulis sebagai

$$d_{euc}(x_{ij}, c_{kj}) = \sqrt{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - c_{kj})^2},$$

dengan $d_{euc}(x_{ij}, c_{kj})$adalah jarak Euclidean antara pengamatanke-$i$variabel ke-j ke pusat klaster ke-k pada variabel ke-j, x_{ij} adalah objek pada pengamatan ke-i pada variabel ke-j, c_{kj} adalah pusat kelompok ke-k pada variabel ke-j, p adalah banyak variabel yang diamati, dan n adalah banyak pengamatan yang diamati.

4) Mengalokasikan objek *non-medoid*berdasarkan jarak terdekat dengan *medoid*dan menghitung total jarak yang diperoleh.
5) Memilih secara acakobjek *non-medoid* pada masing-masing kelompok sebagai kandidat *medoid* baru dan menghitung jarak setiap objek *non-medoid* dengan kandidat *medoid* baru.
6) Mengalokasikan objek berdasarkan jarak terdekatnya dengan kandidat *medoid* baru. Kemudian menghitung kembali total jaraknya.
7) Menghitung total simpangan(S) dengan nilai total jarak baru dikurangi total jarak lama. Jika diperoleh $S < 0$, maka kandidat *medoid* baru tersebut menjadi *medoid* baru.
8) Mengulangi langkah (e) sampai dengan (h) hingga tidak terjadi perubahan *medoid*. Proses iterasi akan berhenti apabila diperoleh $S > 0$. Dan pada langkah ini diperoleh kelompok beserta anggota kelompoknya masing-masing.

c. Validasi Hasil Pengelompokan

Validasi hasil pengelompokan dilakukan untuk memperoleh partisi yang paling sesuai dengan data. Jika kelompok tidak divalidasi, maka akan berpengaruh pada hasil analisis. Pada artikel inimetode validasi yang digunakan adalah koefisien *silhouette*. Metode koefisien *silhouette* merupakan gabungan dua metode yaitu metode *cohesion*yang berfungsi untuk mengukur kedekatan data yang berada pada satu kelompok dan metode *separation*yang berfungsi untuk mengukur kedekatan antar kelompok yang terbentuk.

Plot*silhouette* menunjukkan *silhouette* semua kelompok sehingga kualitas kelompok dapat dibandingkan berdasarkan tingkat kelebaran (gelap) *silhouette.* Semakin lebar *silhouette* akan semakin baik kualitas suatu kelompok. Nilai k yang menghasilkan rata-rata lebar *silhouette* tertinggi disebut plot koefisien *silhouette*atau *silhouette coefficient* (SC) (Kaufman dan Rousseeuw, 1990). Berikut adalah langkah metode koefisien *silhouette* menurut Struyf, *et al.*(1997).

1) Menghitung rata-rata jarak objek ke-i dengan semua objek yang berada di dalam satu kelompok Adengan persamaan
$$a(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i,j),$$
dengan j adalah objek lain dalam satu kelompok A dan $d(i,j)$ adalah jarak antara objek i dan j.
2) Menghitung rata-rata jarak objek ke-i dengan semua objek yang berada pada kelompok lain dengan persamaan
$$d(i,C) = \frac{1}{|C|} \sum_{j \in C} d(i,j),$$
dengan $d(i,C)$ adalah jarak rata-rata objek i dengan semua objek pada kelompok lain C dimana $A \neq C$.
3) Menentukan nilai minimumnya yaitu $b(i)$ yang menunjukkan perbedaan rata-rata objek i untuk kelompok yang terdekat dengan tetangganya dapat dituliskan dengan persamaan

$C \neq A$

$$b(i) = min\ d(i, C).$$

4) Menghitung nilai *silhouette* dengan persamaan

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}.$$

Hasil perhitungan $s(i)$berada pada kisaran -1 hingga 1. Menurut Kaufman dan Rousseeuw (1990), nilai $s(i)$ dapat diartikan sebagai

a) $s(i) \approx 1$ artinya objek i terletak di kelompok yang tepat (dalam A),
b) $s(i) \approx 0$artinya objek i terletak di antara dua kelompok (A dan B),
c) $s(i) \approx -1$artinya objek i terletak di kelompok yang tidak tepat (lebih dekat ke B daripada A).

5) Menghitung koefisien *silhouette* yang didefinisikan sebagai rata-rata $s(i)$ yaitu

$$SC = \frac{1}{n}\sum_{i=1}^{n} s(i),$$

dengan n adalah banyak pengamatan.

Pengelompokan terbaik dicapai jika SC maksimal artinya meminimalkan jarak dalam kelompok $(a(i))$ sekaligus memaksimalkan jarak antarkelompok $(b(i))$ (Vendramin,*et al.,* 2009). Ukuran nilai koefisien *silhouette* menurut Kaufman dan Rousseeuw (1990) adalah seperti ditunjukkan pada Tabel 1.

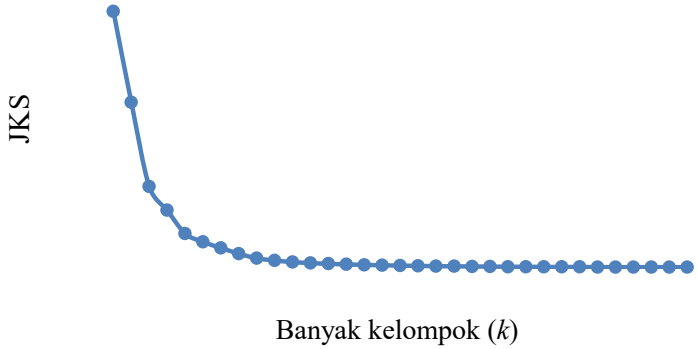
Tabel 1. Kategori Nilai Koefisien*Silhouette* dan Interpretasinya

Koefisien *Silhouette*	Interpretasi
$0.7 < SC \leq 1$	Terdapat ikatan yang sangat baik (*strong structure*) antara objek dan kelompok yang terbentuk.
$0.5 < SC \leq 0.7$	Terdapat ikatan yang cukup baik (*mediumstructure*) antara objek dan kelompok yang terbentuk.
$0.25 < SC \leq 0.5$	Terdapat ikatan yang lemah (*weak structure*) antara objek dan kelompok yang terbentuk.
$SC \leq 0.25$	Tidak terdapat ikatan antara objek dan kelompok yang terbentuk.

d. Penerapan

Pada penerapan ini digunakan data kasus demam berdarah (DBD) di Indonesia tahun 2016 dengan jumlah amatan 34 dengan variabel angka kesakitan (*Incident Rate* (IR)) dan kasus kematian (*Case Fatality Rate* (CFR)) sebagai indikator yang menunjukkan tingginya permasalahan DBD di suatu wilayah. Data diperoleh dari Kementrian Kesehatan Republik Indonesia. Sebelum dilakukan pengelompokan, data dinormalisasi terlebih dahulu agar tidak ada parameter yang mendominasi dalam perhitungan proses pengelompokan.Kemudian dilakukan deteksi pencilan pada data dan ditemukan data pencilan yaitu provinsi Kalimantan Timur, Bali, dan Maluku. Selanjutnya menentukan kelompok beserta anggota masing-masing kelompok dengan algoritme *k-medoids.*

Algoritme *k-medoids* diawali dengan menentukan jumlah kelompok yang ingin dibentuk menggunakan kriteria *elbow*. Kemudian diperoleh grafik kriteria *elbow* yang ditunjukkan pada Gambar 1.



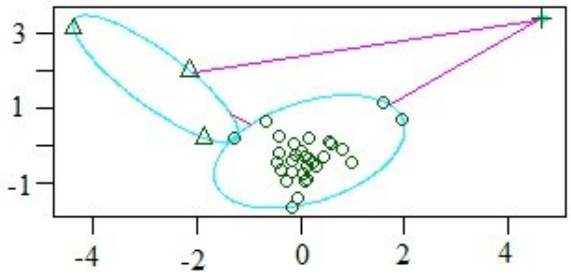
Gambar 1. Grafik Kriteria *Elbow* Kasus DBD di Indonesia Tahun 2016

Dari Gambar 1 tampak bahwa nilai JKS mengalami penurunan dari nilai kelompok $k = 2$ ke $k = 3$, kemudian dari $k = 3$ ke $k = 4$ terlihat penurunan drastis membentuk siku pada titik $k = 3$. Hal ini ditunjukkan juga pada Tabel 2terdapat penurunan nilai JKS paling besar pada $k = 3$sehingga jumlah kelompok yang akan dibentuk sebanyak 3.

Tabel 2. Hasil JKS dari tiap kelompok untuk 34 provinsi

k	JKS	Selisih	*k*	JKS	Selisih	*k*	JKS	Selisih
2	66,000	66,000	13	1,053	0,273	24	0,118	0,032
3	42,512	23,488	14	0,872	0,182	25	0,095	0,023
4	20,765	21,747	15	0,715	0,157	26	0,074	0,021
5	14,656	6,109	16	0,599	0,116	27	0,055	0,019
6	8,685	5,971	17	0,490	0,109	28	0,039	0,016
7	6,503	2,182	18	0,406	0,084	29	0,025	0,014
8	4,972	1,531	19	0,330	0,077	30	0,014	0,011
9	3,458	1,514	20	0,271	0,059	31	0,009	0,005
10	2,287	1,171	21	0,220	0,051	32	0,005	0,004
11	1,724	0,562	22	0,183	0,037	33	0,002	0,003
12	1,326	0,398	23	0,150	0,033	34	0,001	0,001

Kemudian dengan menggunakan *software* R, pengujian algoritme *k-medoids* diperoleh hasil pengelompokan seperti ditunjukkan pada Gambar 2.



Gambar 2. Plot *Scatter*Pengelompokan Data

Berdasarkan Gambar 2 diuraikan bahwa objek diplot ke dalam 3 kelompok yaitu:

1) Kelompok I terdiri atas30 provinsi yaituAceh, Sumatra Utara, Sumatra Barat, Riau, Kepulauan Riau, Jambi, Sumatra Selatan, Bangka Belitung, Bengkulu, Lampung, Banten, Jawa Barat, Jawa Tengah, D.I.Yogyakarta, Jawa Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Sulawesi Utara, Gorontalo, Sulawesi Tengah, Sulawesi Barat, Sulawesi Selatan, Sulawesi Tenggara, NTB, NTT, Maluku Utara, Papua Barat, Papua, dan Kalimantan Tenggaradengan Sulawesi Barat sebagai *medoid*.
2) Kelompok IIterdiri atas3 provinsi yaitu DKI Jakarta, Kalimantan Timur, dan Balidengan Kalimantan Timur sebagai*medoid*.
3) Kelompok III terdiri atas1 provinsi yaitu Maluku.

Dari hasil pengelompokan, analisis hasil pengelompokan dan karakteristiknya dapat ditunjukkan pada Tabel 3.

Tabel 3. Analisis tiap kelompok dan karakteristiknya

Kelompok	Koefisien *Silhouette* (SC)	
	CFR	IR
I	Sedang	Sedang
II	Rendah	Tinggi
III	Tinggi	Rendah

Berdasarkan Tabel 3 dapat diketahui karakteristik dari kelompok I adalah provinsi dengan IR sedang dan CFR sedang. Kemudian kelompok II merupakan provinsi dengan IR tinggi namun CFR-nya rendah.Sedangkan kelompok III merupakan provinsi dengan IR rendah namun CFR-nya tinggi. IR yang tinggi dapat dipengaruhi faktor curah hujan yang tinggi sepanjang tahun dan faktor kebersihan lingkungan. Provinsi dengan IR tinggi belum tentu mempunyai CFR tinggi. Provinsi yang umumnya mempunyai IR yang tinggi namun CFR-nya rendah dikarenakan kemungkinan pelayanan medis dan akses kesehatan di provinsi tersebut sudah maju. Maluku sebagai provinsi yang mempunyai IR rendah namun CFR tertinggi perlu meningkatkan akses dan pelayanan medis dalam upaya promosi kesehatan.

Untuk hasil validasi menggunakan koefisien *silhouette* diperoleh hasil seperti ditunjukkan pada Tabel 4.

Tabel 4. Koefisien *Silhouette* untuk ketiga kelompok yang terbentuk

Kelompok	Koefisien *Silhouette*(SC)
I	0.7083344
II	0.1813012
III	0.0000000

Berdasarkan kategori nilai koefisien *silhouette* pada Tabel 1 diperoleh koefisien *silhouette* kelompok I sebesar 0.7083344 termasuk dalam *strong structure*, sedangkan kelompok II sebesar 0.1813012 termasuk dalam *weak structure*, dan kelompok III sebesar nol artinya tidak ada ikatan antara objek dan kelompok yang terbentuk. Nilai koefisien *silhouette* keseluruhan

sebesar0.6409981termasuk dalam *medium structure*artinya terdapat ikatan yang cukup baik antara objek dan kelompok yang terbentuk.

4. SIMPULAN

Berdasarkan hasil dan pembahasan diperoleh simpulan sebagai berikut.

a. Kriteria *elbow* dapat menentukan jumlah kelompok (k)terbaik untuk pengelompokan data yang memuat pencilan dengan algoritme *k-medoids*.
b. Koefisien *silhouette*sebagai metode validasi hasil pengelompokan, menghasilkan nilai berada diantara -1 dan 1. Jika koefisien *silhouette*mendekati 1, maka ikatan antara objek dan kelompok yang terbentuk sangat baik.
c. Pengelompokan data DBD Indonesia tahun 2016 menghasilkan 3 kelompok dengan kelompok I terdiri atas 30 provinsi, kelompok II terdiri atas 3 provinsi, dan kelompok III terdiri atas 1 provinsi.
d. Hasil validasi kelompok menunjukkan nilai koefisien *silhouette* keseluruhan sebesar0.6409981artinya terdapat ikatan yang cukup baik antara objek dan kelompok yang terbentuk.

5. DAFTAR PUSTAKA

Barnett, V. dan T. Lewis (1994). *Outliers in Statistical Data.* New York: John Wiley & Sons.

Flowrensia, Y. (2010). *Perbandingan Penggerombolan K-Means dan K-Medoids Pada Data Yang Mengandung Pencilan* [Skripsi].Institut Pertanian Bogor. Bogor.

Folzmiser, P. (2005). Identification of Multivariate Outliers: A Performance Study. *Australian Journal of Statistics*, **34**(2), 127-138.

Han, J., dan M. Kamber(2012). *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann Publisher.

Johnson, R.A., andD.W. Wichern(2002).*Applied Multivariate Analysis 5th Edition*. New Jersey: Prentice Hall.

Kaufman L., and P.J. Rousseeuw. (1987). *Clustering By Means of Medoids*. New York: John Wiley & Sons.

Kaufman L., and P.J. Rousseeuw.(1990). *Finding Groups in Data*. New York: John Wiley & Sons.

Kokasih, V. (2016). *Clustering Penggunaan Bandwith Menggunakan Metode K-Means Algorithm pada Penerapan Single Sign On (SSO) Universitas Sebelas Maret* [Skripsi]. Universitas Sebelas Maret. Surakarta.

Madulatha, T.S. (2012).An Overview On Clustering Methods. *IOSR Journal of Engineering,* **II**(4), 719-725.

Santoso, S. (2010). *Statistik Multivariat*. Jakarta: Elex Media Komputindo.

Struyf, A., M. Hubert, P.J. Rousseeuw. (1997). Integrating Robust Clustering Techniques in S-PLUS. *Journal of Computational Statistics and Data Analysis,***26**(1),17-37.

Vendramin,L., R.J.G.B. Campello, and E.R. Hruschka.(2009). *On the Comparison of Relative Clustering Validity Criteria*. Proceedings of the SIAM International Conference on Data Mining, **3**(4), 733-744.