JURNAL MATEMATIKA "MANTIK"
Edisi: Mei 2018. Vol. 04 No. 01
ISSN: 2527-3159 **E-ISSN: 2527-3167**

Analisis Cluster dengan Data Outlier Menggunakan *Centroid Linkage* dan K-Means Clustering untuk Pengelompokan Indikator HIV/AIDS di Indonesia

Rini Silvi
Prodi Magister Statistika Terapan,Universitas Padjajaran, rinisilvi@stis.ac.id

DOI:https://doi.org/10.15642/mantik.2018.4.1.22-31

Abstrak

Analisis klaster adalah salah satu metode yang digunakan untuk mengelompokkan obyek atau pengamatan yang didasarkan atas kemiripannya. Obyek yang berada dalam satu kelompok memiliki kemiripan satu sama lain. Pada penelitian ini analisis klaster yang digunakan adalah *K-means* dengan *Centroid Linkage. K-means* adalah salah satu metode *clustering* non-hierarki yang sederhana dan mudah diimplementasikan. Sedangkan, *Centroid Linkage* adalah metode klaster hierarki yang dapat digunakan pada data yang mengandung outlier, dimana outlier bisa membuat data yang diolah tidak mencerminkan gambaran sebenarnya. Untuk memudahkan, outlier seringkali dibuang, padahal acapkali outlier mengandung informasi penting. HIV/AIDS adalah salah satu tantangan serius terhadap kesehatan masyarakat dunia, karena HIV/AIDS merupakan penyakit menular yang menyerang sistem kekebalan tubuh sehingga penderita mengalami penurunan ketahanan tubuh secara terus-menerus yang berujung pada kematian. Data Indikator HIV/AIDS di Indonesia mengandung outlier. Penelitian ini menggunakan gap statistik untuk menentukan jumlah klaster ideal yang mengelompokkan propinsi berdasarkan indikator HIV/AIDS sedemikian hingga terbagi menjadi 7 klaster. Dari perbandingan rasio S_w/S_b, *Centroid Linkage* lebih homogen dibandingkan *K-means*. Dengan *custering*, diharapkan pemerintah dapat mengambil kebijakan berdasarkan indikator-indikator dominan yang terdapat pada masing-masing klaster.

Kata Kunci : *Clustering, Centroid Linkage, K-means*

Abstract

Cluster analysis is a method to group data (objects) or observations based on their similarities. Objects that become members of a group have similarities among them. Cluster analyses used in this research are K-means clustering and Centroid Linkage clustering. K-means clustering, which falls under non-hierarchical cluster analysis, is a simple and easy to implement method. On the other hand, Centroid Linkage clustering, which belongs to hierarchical cluster analysis, is useful in handling outliers by preventing them skewing the cluster analysis. To keep it simple, outliers are often removed even though outliers often contain important information. HIV/AIDS is a serious challenge for global public health since HIV/AIDS is an infectious disease attacking body's immune system that in turn lowering the ability to fight infections which in the end causing death. HIV/AIDS indicators data in Indonesia contain outliers. This research uses gap statistic to define the number of clusters based on HIV/AIDS indicators that groups Indonesia provinces into 7 clusters. By comparing S_w/S_b ratio, Centroid Linkage clustering is more homogenous than K-means clustering. Using clustering, the government shall be able to create a better policy for fighting HIV/AIDS based on the dominant indicators in each cluster.

Keyword : *Clustering, Centroid Linkage, K-means*

1. Pendahuluan

HIV/AIDS merupakan penyakit menular yang disebabkan oleh infeksi *Human Immunodeficiency Virus* yang menyerang sistem kekebalan tubuh sehingga tidak dapat melawan infeksi dan berujung pada kematian. Penyakit ini merupakan masalah dan tantangan serius terhadap kesehatan masyarakat di dunia. Di Indonesia, sebagian besar infeksi baru diperkirakan terjadi pada beberapa sub-populasi berisiko tinggi dengan prevalensi > 5% [1]. Risiko penularan HIV/AIDS tidak hanya terbatas pada sub-populasi yang berperilaku risiko tinggi, tetapi juga dapat menular pada pasangan atau bahkan anaknya.

Pada tahun 2016, kasus baru infeksi HIV meningkat 33,4% dibandingkan tahun sebelumnya. Terdapat 69,3% kasus baru infeksi HIV pada kelompok umur 25-49, sementara 63,3% penderita adalah laki-laki. Rasio HIV/AIDS antara laki-laki dan perempuan tercatat pada kisaran 2:1 [2]. Pada tahun 2016, jumlah kasus baru infeksi HIV terbanyak adalah Jawa Timur, kemudian diikuti oleh Jakarta, Jawa Barat, Jawa Tengah, dan Papua. Akan tetapi, persentase kasus baru infeksi HIV terbesar adalah Papua, jika dibagi dengan jumlah penduduknya. Sementara, Gorontalo menempati urutan terakhir propinsi dengan jumlah kasus baru infeksi HIV terbanyak di Indonesia.

Perbedaan jumlah kasus baru HIV tampak begitu nyata di beberapa propinsi. Contohnya DKI Jakarta dengan tingkat prevalensi 58,56 padahal dengan tingkat kemiskinan paling kecil yaitu sebesar 3,73. Meskipun tingkat pendidikan di Gorontalo termasuk rendah, akan tetapi tingkat prevalensi kasus baru HIV hanya sekitar 0.6. Tingkat prevalensi disini dihitung berdasarkan jumlah kasus baru infeksi HIV suatu daerah dibandingkan setiap seratus ribu penduduk.

Terdapat beberapa indikator yang mempengaruhi HIV/AIDS yaitu tingkat penggunaan kontrasepsi (kondom), jumlah dokter/tenaga medis, proporsi muslim, tingkat kesuburan remaja, dan rata-rata lama sekolah [3]. Sementara menurut [4], HIV dipengaruhi oleh indikator seperti kurangnya pendidikan, kemiskinan, seks bebas, dan kehidupan malam meskipun pada penelitiannya lebih difokuskan kepada perempuan.

Ada beberapa indikator yang dapat mempengaruhi prevalensi HIV yang digunakan disini. Pada indikator tersebut teridentifikasi beberapa data outlier. Hal ini bisa mempengaruhi hasil kesimpulan dari penelitian. Outlier dapat menghasilkan output yang tidak sesuai dengan gambaran yang sebenarnya, termasuk dalam hal pengklasteran indikator prevalensi HIV. Maka dari itu, penelitian ini perlu menggunakan metode klaster yang dapat menangani pengaruh keberadaan outlier tersebut. Metode yang dimaksud adalah K-Means Clustering.

Tujuan dari penelitian ini adalah untuk mengumpulkan propinsi ke dalam suatu kelompok sedemikian hingga dapat dibedakan menurut karakteristik indikator dari prevalensi HIV. Dengan adanya kelompok-kelompok berdasarkan indikator-indikator tersebut, diharapkan pemerintah dapat membuat kebijakan yang tepat untuk pencegahan dini menyebarnya HIV/AIDS.

2. Tinjauan Pustaka

2.1 Analisis Klaster

Analisis klaster adalah suatu teknik statistik yang bertujuan untuk mengelompokkan obyek ke dalam suatu kelompok sedemikian sehingga obyek yang berada dalam satu kelompok akan memiliki kesamaan yang tinggi dibandingkan dengan obyek yang berada di kelompok lain [5]. Dengan kata lain tujuan dari analisis cluster adalah pengklasifikasian obyek-obyek berdasar-kan similaritas diantaranya dan menghim-pun data menjadi beberapa kelompok [16]. Ada dua metode dalam analisis klaster yaitu metode hierarki dan metode non hierarki. Menurut [6], metode non hierarki umumnya digunakan jika jumlah satuan pengamatan besar dan jumlah klaster tidak ditentukan sebelumnya. Salah satu metode non hierarki adalah metode K-means. Ini adalah metode non hirarki yang paling banyak digunakan. Algoritma K-means mudah diimplementasikan dan juga mudah diadaptasi sehingga menjadikannya lebih populer dalam hal pengelompokan. Pada teknik K-means, biasanya peneliti sudah terlebih dahulu menentukan banyaknya klaster yang akan dibentuk.

Metode hierarki merupakan metode pengelompokkan yang terstruktur dan bertahap berdasarkan kemiripan sifat antar obyek. Kemiripan sifat tersebut dapat ditentukan dari kedekatan jarak Euclidean atau jarak Mahalanobis.

Jarak Euclidean digunakan jika tidak terjadi korelasi. Jarak Euclidean dirumuskan sebagai berikut:

$$d(y,x)=\sqrt{\sum_{k=1}^{l}(y_k-x_k)^2}\ ;\ l=1,2,3,...,n$$

dimana:

$d(y,x)$:kuadrat jarak Euclid antar obyek y dengan obyek pada x.

y_k : nilai dari obyek y pada variabel ke-k

x_k : nilai dari obyek x pada variabel ke-k

Jarak Mahalanobis digunakan jika data terjadi korelasi. Jarak Mahalanobis antara dua sampel X dan Y dari suatu variabel acak didefinisikan sebagai berikut [5] :

$$d_{Mahalanobis}(y,x)=\sqrt{(y-x)^T\sum{}^{-1}(y-x)}$$

Dengan $\sum$ adalah suatu matriks varians kovarians.

Dalam metode klaster hierarki terdapat beberapa metode penghitungan jarak yang dapat digunakan, antara lain metode pautan tunggal (*single linkage*), metode pautan lengkap (*complete linkage*), metode pautan rataan (*average linkage*), metode Ward, dan metode *Centroid Linkage*.

2.2 Multikolinearitas

Multikolinearitas adalah adanya hubungan linear yang sempurna atau pasti di antara beberapa atau semua variabel [7]. Multikolinearitas berkenaan dengan terdapatnya lebih dari satu hubungan linear pasti. Untuk mengetahui adanya multikolinearitas salah satunya adalah dengan menghitung nilai *Variance Inflation Factor* (VIF) dengan rumus:

$$VIF_j=\frac{1}{1-R_j^2}$$

Menurut [7], terjadinya multikolinearitas apabila nilai (VIF_j) ≥ 10. Jika terindikasi terjadi multikolinearitas maka harus dilakukan tindakan perbaikan multikolinearitas.

3. Metode Penelitian

3.1 Sumber Data dan Variabel Penelitian

Penelitian ini menggunakan data sekunder tahun 2016 yang diperoleh dari Badan Pusat Statistik Republik Indonesia (BPS RI) kemudian diolah dengan *sofware R* versi 3.4.3. Data prevalensi kasus baru infeksi HIV digunakan sebagai data pembanding dalam pengelompokkan klaster. Variabel indikator yang digunakan menjadi faktor risiko prevalensi HIV yaitu persentase penduduk miskin (X_1) [8], tingkat pengangguran terbuka (X_2) [9], jumlah puskesmas (X_3) [10] dan rasio penduduk 15 tahun ke atas yang masih berpendidikan rendah (X_4) [11], dan persentase pasangan usia subur yang memakai alat kontrasepsi berupa kondom (X_5) [2]. Pendidikan rendah diasumsikan untuk yang berumur 15 tahun keatas tapi tidak punya ijazah, atau hanya sampai tamatan SD dan SMP. Alat kontrasepsi yang digunakan difokuskan pada kondom karena alat kontrasepsi yang lainnya hanya digunakan untuk mencegah kehamilan tetapi tidak bisa mengurangi penularan HIV/AIDS. Unit observasi pada penelitian ini adalah seluruh propinsi di Indonesia.

3.2 Tahapan Penelitian

3.2.1 Standardisasi Data

Proses standardisasi dilakukan apabila di antara variabel-variabel yang diteliti terdapat perbedaan ukuran satuan yang besar. Perbedaan satuan yang mencolok dapat mengakibatkan perhitungan pada analisis klaster menjadi tidak valid. Oleh karena itu, perlu dilakukan proses standardisasi dengan melakukan transformasi pada data asli sebelum dianalisis lebih lanjut.

3.2.2 Deteksi Outlier dan Multikolinearitas

Analisis klaster pada hakekatnya adalah teknik algoritma, bukan alat inferensi statistik. Oleh sebab itu persyaratan seperti distribusi data yang harus normal (di analisis statistik lainnya) ataupun hubungan linier antar variabel tidak menjadi syarat dalam analisis klaster. Namun demikian, karena data yang diolah dalam analisis klaster biasanya hanya sebagian kecil dari populasi, maka agar hasilnya bisa digeneralisasi, data yang diolah sebaiknya mencerminkan gambaran umum atau bersifat representatif. Oleh sebab itu, *outliers* tetap harus dihilangkan dari sampel agar hasilnya tidak bias.

Deteksi outlier digunakan untuk mencari data yang berbeda dengan mayoritas data yang lain. Walaupun memiliki perilaku yang berbeda dengan mayoritas data yang lain dan sering dianggap *noise*, tetapi outlier sering kali mengandung informasi yang sangat berguna. Tidak semua data yang mengandung outlier bisa

ditransformasi karena kasus data yang berbeda-beda. Akan tetapi, dengan menggunakan metode *Centroid Linkage*, outlier tidak berpengaruh secara signifikan.

Selain itu, data yang digunakan seharusnya tidak berkorelasi, dengan kata lain sebaiknya tidak ada multikolinieritas. Alasannya adalah di dalam analisis klaster setiap variabel diberi bobot yang sama dalam perhitungan jarak. Manakala beberapa varibel saling berkorelasi, korelasi tersebut akan menyebabkan pembobotan yang tidak berimbang sehingga akan mempengaruhi hasil analisis [12].

3.2.3 Penentuan Jumlah Klaster Optimum

Penentuan jumlah *cluster* optimum dilakukan dengan menggunakan gap statistik pada R. Gap statistik bertujuan untuk menentukan jumlah klaster lebih konstan dibandingkan pengukuran lainnya. Jarak obyek berpasangan dalam klaster dihitung dengan rumus:

$$D_r = \sum_{i,i` \epsilon C_r} d_{ii`}$$

Dimana *d* adalah jarak euclidean kuadrat. Kemudian hitung jumlah kuadrat di dalam klaster menggunakan rumus:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

Nilai gap didapatkan dengan mengestimasi jumlah klaster optimum pendekatan standardisasi W_k:

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$$

Dimana E_n^* adalah nilai ekspektasi dari distribusi jumlah sampel.

Kriteria banyak *cluster* optimum diberikan oleh nilai gap statistik (k) yang paling tinggi, atau yang pertama kali mengindikasi kenaikan gap yang minimum jika gap selalu naik [13]. Setelah penentuan klaster optimum, maka akan dibandingkan pengklasteran dengan menggunakan dua metode, yaitu metode *Centroid Linkage* dan metode K-means.

3.2.4 Metode Klaster K-means

Metode K-means adalah metode non hierarki yang paling banyak digunakan dalam pengklasteran. Algoritma K-means mudah diimplementasikan. Pada metode ini, peneliti menentukan sendiri jumlah klaster yang akan dibentuk. Peneliti mengelompokkan entitas ke dalam k-kelompok, biasanya dilakukan secara acak. Pada masing-masing kelompok dihitung rata-ratanya. Hitung jarak setiap entitas terhadap pusat masing-masing kelompok (rata-rata kelompoknya).

Masing-masing objek dialokasikan ke klaster terdekat dengan pusatnya. *Update* keanggotaan setiap entitas berdasarkan jarak terdekat dengan pusat kelompok dan ditentukan kembali pusat klaster yang baru. Proses pangalokasian obyek kembali dilakukan. Suatu obyek dapat berpindah ke klaster lain bila obyek tersebut lebih dekat ke pusat klaster tersebut. Proses ini dilakukan secara berulang sampai tidak ada lagi entitas yang berpindah kelompok.

3.2.5 Metode Klaster *Centroid Linkage*

Centroid Linkage adalah rata-rata semua obyek dalam klaster. Jarak antara dua klaster adalah jarak antar *centroid* klaster tersebut. Klaster *centroid* adalah nilai tengah observasi pada variabel dalam suatu set variabel cluster. Dengan metode ini, setiap terjadi klaster baru segera terjadi perhitungan ulang *centroid* sampai terbentuk klaster yang tetap [5].

Keuntungan dari metode ini adalah outlier tidak berpengaruh secara signifikan, jika dibandingkan dengan metode lain. Jarak antara dua klaster didefinisikan sebagai berikut:

$$d_{(UV)W} = d\left(\bar{x}_1, \bar{x}_2\right)$$

Centorid klaster baru yang terbentuk didapat dengan rumus:

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2}$$

Dimana:

$N_1 = N_2$: banyaknya obyek.

Centroid adalah rata-rata dari semua anggota dalam klaster tersebut. Pada saat obyek digabungkan maka *centroid* baru dihitung, sehingga setiap kali ada penambahan anggota, *centroid* akan berubah pula [14].

3.2.6 Langkah-langkah K-means *Clustering*

Langkah-langkah K-means:

1. Menentukan k sebagai jumlah klaster yang diinginkan
2. Mengalokasikan data ke dalam cluster secara acak
3. Menentukan pusat klaster dari data yang ada pada masing-masing klaster dengan persamaan:

$$C_{kj} = \frac{x_{1j} + x_{2j} + \cdots + x_{nj}}{n}$$

dimana:

C_{kj} : pusat c*luster* ke-k pada variabel ke j (j=1, 2, …, p)

n : banyak data pada *cluster* ke-k

4. Menentukan jarak setiap obyek dengan setiap *centroid* dengan perhitungan jarak setiap obyek dengan setiap *centroid* menggunakan jarak *Euclidean*
5. Menghitung fungsi obyektif dengan formula: $l = \sum_{i=1}^{n}\sum_{j=1}^{k} a_{ij} d\left(x_i, C_{kj}\right)^2$
6. Mengalokasikan masing-masing data ke *centroid*/rata-rata terdekat.
7. Mengulangi kembali langkah 3-6 sampai tidak ada lagi perpindahan obyek atau tidak ada perubahan pada fungsi obyektifnya.

3.2.7 Langkah-langkah *Centroid Linkage Clustering*

Langkah-langkah *Centroid Linkage Clustering*:

1. Membuat *k* klaster. Masing-masing individu atau unit observasi jadi kelompok. Kemudian dibuat matrik jaraknya (dari *i* ke kelompoknya), dengan rumus:

$$D = \{d_{ik}\}$$

2. Mencari jarak terkecil dari pasangan klaster, yaitu d_{uv} (jarak klaster u dan klaster v)
3. Menggabungkan klaster u dan klaster v. kemudian update matriks jaraknya.
4. Ulangi langkah 2 & 3 sebanyak N-1 kali. Catat nilai jarak untuk setiap terjadi penggabungan klaster
5. Tentukan nilai *cut off* untuk menentukan klaster terbentuk. Lakukan dengan membuat dendogram, dan tentukan *cut off* jumlah klaster
6. Pemberian nama klaster berdasarkan *profiling*, yaitu melihat karakteristik klaster terbentuk secara rata-rata.

3.2.8 Penentuan Metode Terbaik

Tahap Evaluasi dapat dilakukan dengan melakukan analisis klaster dengan ukuran jarak atau metode klaster yang berbeda kemudian dibandingkan hasilnya [7]. Pemilihan metode yang menghasilkan kualitas pengelompokan terbaik dilakukan dengan memperhatikan nilai rasio rata-rata simpangan baku dalam klaster terhadap simpangan baku antar klaster [15]. Rata-rata simpangan baku di dalam klaster (Sw) dinyatakan dengan:

$$s_w = \frac{1}{c}\sum_{k=1}^{c} s_k$$

Simpangan baku antar klaster (s_b) dinyatakan sebagai:

$$s_b = \left[\frac{1}{c-1}\sum_{k=1}^{c}(\bar{X}_k - \bar{X})^2\right]^{\frac{1}{2}}$$

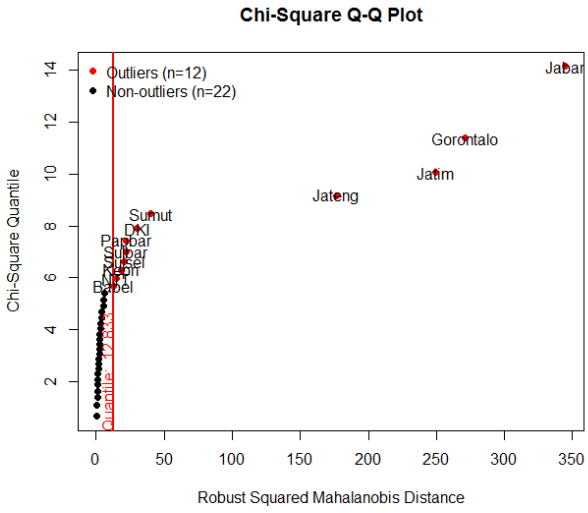
Dimana c adalah jumlah klaster, s_k merupakan simpangan baku di dalam klaster ke-k. $\bar{X}_k$ sebagai rata-rata klaster ke-kdan $\bar{X}$ adalah rata-rata dari semua klaster. Semakin kecil nilai s_w dan semakin besar nilai s_b maka metode tersebut memiliki kinerja yang baik, artinya memiliki homogenitas yang tinggi. Metode yang dipilih adalah yang memberikan nilai rasio s_w/s_b terkecil.

4. Hasil dan Pembahasan

4.1 Jumlah Klaster Optimum

Sebelum melakukan analisis lebih lanjut terlebih dahulu dilakukan standardisasi data, karena terdapat perbedaan satuan yang mencolok sehingga dapat mengakibatkan perhitungan pada analisis klaster menjadi tidak valid. Transformasi dilakukan menggunakan R GUI 3.4.3.

Pemeriksaan awal yang dilakukan adalah melihat apakah variabel yang digunakan terdapat outlier atau tidak. Pada data ini terdapat beberapa provinsi yang outlier, diantaranya yaitu Jawa Barat, Gorontalo, dan Jawa Timur, Jawa Tengah, serta beberapa daerah lainnya. Pada Gambar 1, terdapat 12 titik outlier. Untuk menangani masalah outlier, tanpa melakukan perubahan pada data, perlu dibandingkan metode terbaik dari beberapa metode yang ada.

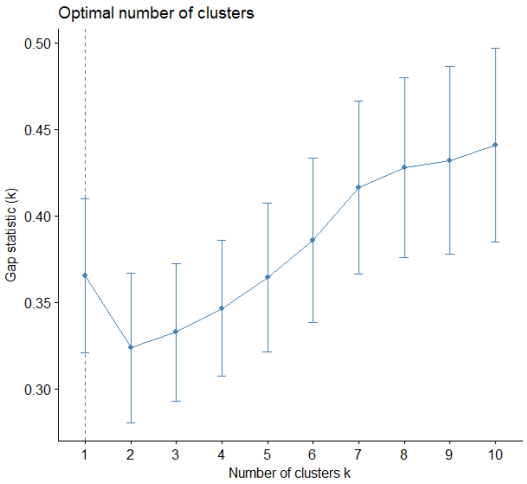


Gambar 1. Pendeteksian Outlier

Mendeteksi multikolinearitas dengan menggunakan VIF dengan hasil masing-masing

yaitu 1,28 (X1), 1,38 (X2), 1,32 (X3), 1,52 (X4), dan 1,18 (X5). Semua variabel menunjukkan nilai VIF sangat kecil (dibawah 2) yang artinya variabel yang digunakan tidak mengandung multikolinearitas.

Tahap berikutnya adalah pemilihan jumlah klaster optimum dengan menggunakan pendekatan gap statistik. Pada penelitian ini, gap statistik menunjukkan jumlah klaster optimum adalah sebanyak 7, yaitu pada titik yang pertama kali perbedaan gap minimum karena secara praktis tidak ideal memilih titik tertinggi mengingat hasil gap statistik yang semakin meningkat hingga jumlah klaster maksimal.

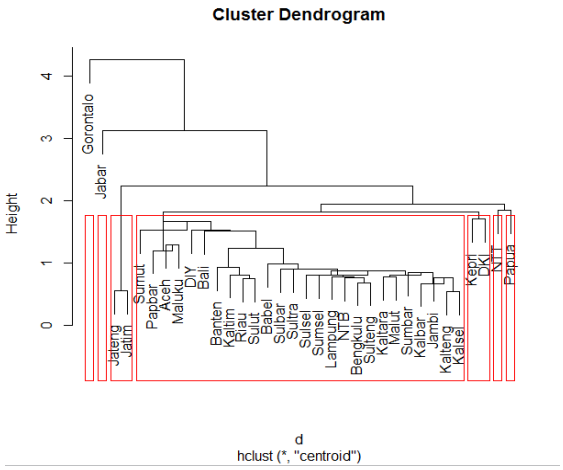


Gambar 2. Jumlah Klaster (k) Berdasarkan Gap Statistik

Pada Gambar 2 terlihat bahwa terdapat penurunan yang curam dari klaster 1 ke klaster 2. Kemudian terjadi kenaikan hingga ujung klaster yang dimungkinkan. Dapat diperhatikan bawah dari klaster 7 ke 8 terdapat kenaikan yang landai, tidak seperti kenaikan sebelumnya yang curam. Sehingga diambil jumlah klaster sebanyak 7, yaitu pada saat pertama kali kenaikan menjadi lebih landai.

4.2 Metode *Hierarchical Centroid*

Hasil dendogram dengan menggunakan 7 klaster terbentuk dapat dilihat pada Gambar 3 di bawah ini.



Gambar 3. Dendogram 7 Klaster Dengan *Centroid Linkage*

Kelompok yang terbentuk:

Klaster 1: Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Bangka Belitung, DIY, Banten, Bali, NTB, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat.

Klaster 2: Kepulauan Riau, dan DKI Jakarta.

Klaster 3: Jawa Barat.

Klaster 4: Jawa Tengah dan Jawa Timur.

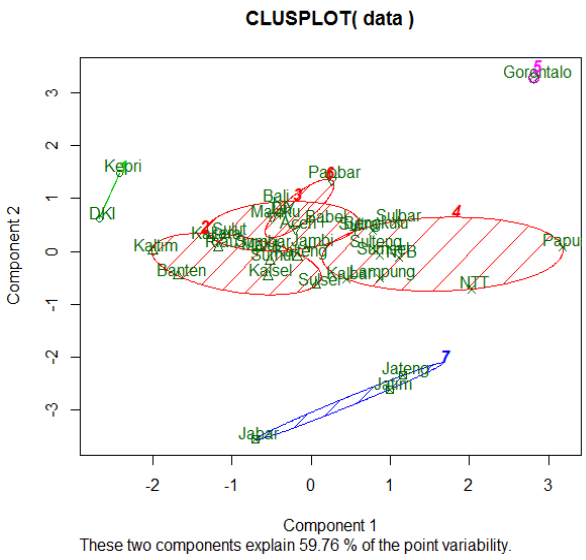
Klaster 5: Nusa Tenggara Timur

Klaster 6: Gorontalo

Klaster 7: Papua

4.3 Metode K-means

Dari gambar terlihat dengan jelas anggota dari masing-masing klaster dari warna dan bentuk titik masing-masing klaster.



Gambar 4. Clusplot 7 Klaster Dengan K-means

Hasil gambar 4, diketahui Gorontalo membentuk klaster sendiri terpisah dari kelompok besar seperti halnya klaster Daerah Khusus Ibukota (DKI) Jakarta dan Kepulauan Riau (Kepri). Klaster lain yang terpisah dari kelompok besar yaitu, klaster Jawa Barat, Jawa Tengah dan Jawa Timur. Pembagian klaster 7 tertuang dalam tabel 1 dibawah ini, dengan anggota sebagai berikut:

Tabel 1. Anggota Klaster K-means

klaster 1	klaster 2	klaster 3	klaster 4	klaster 5	klaster 6	klaster 7
Aceh	Gorontalo	Jabar	Sumsel	Sumut	Sumbar	DIY
Maluku		Jateng	Bengkulu	Riau	Jambi	Bali
Papbar		Jatim	Lampung	Kepri	Babel	Sultra
			NTB	DKI	Kalbar	
			NTT	Banten	Kalteng	
			Sulteng	Kaltim	Kalsel	
			Sulbar	Sulut	Kaltara	
			Papua		Sulsel	
					Malut	

4.4 Penentuan Metode Terbaik

Dengan membandingkan rasio S_w/S_b dapat diketahui metode mana yang memiliki kualitas ketepatan kelompok yang lebih baik. Hasil proses pengelompokkan dengan *Centroid Linkage* dan K-means disajikan pada tabel 2.

Tabel 2. Rasio Sw/Sb untuk Metode Terbaik

Banyak Cluster	*Centroid Linkage*	K-means
(1)	(2)	(3)
7	0,067307	0,112232

Berdasarkan tabel 2, terlihat bahwa nilai rasio S_w/S_b menggunakan metode K-means adalah sebesar 0,112232. Rasio Sw/Sb menggunakan metode *Centroid Linkage* menghasilkan angka 0,067307 dan lebih kecil dari K-means. *Centroid Linkage* menghasilkan kelompok yang lebih homogen sehingga nilai rasio yang dihasilkan lebih kecil. Artinya metode *Centroid Linkage* memiliki kualitas ketepatan kelompok yang lebih baik dibandingkan K-means.

4.5 *Profiling Cluster*

Kelima variabel yang digunakan dalam penelitian ini menunjukkan indikator yang digunakan menjadi faktor risiko HIV/AIDS di Indonesia.

Tabel 3. Profiling Variabel Dominan dan Tidak Dominan Terhadap HIV/AIDS

Group	Penduduk Miskin	tingkat pengangguran terbuka	jml puskesmas	15thn++ Berpendidikan Rendah	PUS dengan kondom
1	10,685009	4,951923	215,8077	63,40231	2,840634
2	4,837749	6,905	206,5	42,69	4,588743
3	8,9159	8,89	1050	69,71	0,135491
4	12,642325	4,42	917,5	72,815	1,714192
5	22,098587	3,25	371	74,14	1,121557
6	17,657282	2,76	93	71,95	17,556
7	28,412421	3,35	394	76,87	5,711383

Klaster pertama, tidak ada indikator dominan yang mempengaruhi HIV/AIDS di daerah tersebut. Hal ini mungkin disebabkan oleh adanya indikator lain yang belum digunakan dalam penelitian ini, seperti indikator kehidupan malam, jumlah pekerja seks komersial (PSK), jumlah panti pijat masing-masing daerah, tingkat religius seperti yang dituliskan oleh [3], dan indikator lainnya.

Klaster kedua, adalah kelompok daerah dengan tingkat kemiskinan rendah dan proporsi umur 15 tahun ke atas dengan pendidikan diatas jenjang SMP (tidak berpendidikan rendah). Klaster ini terdiri dari DKI Jakarta dan Kepulauan Riau.

Klaster ketiga, adalah daerah proporsi pasangan usia subur (PUS) dengan kontrasepsi kondom paling kecil, akan tetapi jumlah puskesmas yang terdapat di daerah ini juga paling banyak. Daerah ini juga memiliki tingkat pengangguran terbuka (TPT) terbesar. Klaster ini terdiri dari 1 propinsi saja yaitu Jawa Barat.

Klaster keempat, adalah kelompok proporsi pasangan usia subur (PUS) dengan kontrasepsi kondom relatif kecil, akan tetapi jumlah puskesmas yang terdapat di daerah ini juga termasuk banyak. Klaster ini terdiri dari propinsi Jawa Tengah dan Jawa Timur.

Klaster kelima, adalah daerah dengan proporsi pasangan usia subur (PUS) dengan kontrasepsi kondom relatif kecil, akan tetapi jumlah puskesmas yang terdapat di daerah ini termasuk sedikit. Klaster ini adalah daerah Nusa Tenggara Timur.

Klaster keenam, adalah daerah dengan proporsi pasangan usia subur (PUS) dengan kontrasepsi kondom paling banyak, sehingga klaster ini termasuk daerah dengan jumlah HIV/AIDS terkecil, yaitu daerah Gorontalo.

Klaster ketujuh, adalah daerah dengan persentase penduduk miskin paling besar dan pendidikan paling rendah. Yaitu daerah Papua.

5. Kesimpulan

Hasil penelitian ini memberikan kesimpulan bahwa untuk data yang memiliki outlier, metode pengklasteran menggunakan *Centroid Linkage* lebih memberikan hasil yang sesuai dengan keadaan dibandingkan dengan metode K-means. Metode K-means lebih heterogen dalam hal ini. Dengan metode *Centroid Linkage*, outlier tidak mempengaruhi klaster analisis dan tidak mengubah hasil dari interpretasi data.

Dari 34 propinsi yang ada di Indonesia, terdapat 7 klaster berdasarkan indikator yang menyebabkan terjadinya HIV/AIDS.

Klaster 1: Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Bangka Belitung, DIY, Banten, Bali, NTB, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat.
Klaster 2: Kepulauan Riau, dan DKI Jakarta.
Klaster 3: Jawa Barat.
Klaster 4: Jawa Tengah dan Jawa Timur.
Klaster 5: Nusa Tenggara Timur
Klaster 6: Gorontalo
Klaster 7: Papua

Referensi

[1] Kementerian Kesehatan RI, Survei Terpadu Biologis dan Perilaku, Jakarta: Kementerian Kesehatan, (2011)

[2] Kementrian Kesehatan RI, Profil Kesehatan Indonesia Tahun 2016, Jakarta: Kementrian Kesehatan RI, (2017)

[3] Mondal, M., & Shitan, M., Factors Affecting The HIV/AIDS Epidemic: An Ecological Analysis Of Global Data, African Health Sciences, 13(2) pp 301–310, (2013)

[4] Singh, R. K., & Patra, S., What Factors Are Responsible For Higher Prevalence Of HIV Infection Among Urban Women Than Rural Women In Tanzania?, Ethiopian Journal of Health Sciences, 25(4), pp 321–328, (2015)

[5] Rahmawati, L., Analisis Kelompok Dengan Menggunakan Metode Hierarki Untuk Pengelompokan Kabupaten/Kota Di Jawa Timur Berdasar Indikator Kesehatan, Jurnal Matematika Vol.1 No.2 Universitas Negeri Malang, (2012)

[6] Anderberg, M., Cluster Analysis For Applications, Academic Press.Inc., (1973)

[7] Ningrat, D.R., Analisis Cluster Dengan Algoritma K-Means dan Fuzzy C-Means Clustering Untuk Pengelompokan Data Obligasi Korporasi, Jurnal Gaussian Vol.5 No.4 Universitas Diponegoro, 2016.

[8] Badan Pusat Statistik (BPS), Indikator Pembangunan Berkelanjutan 2017: Jumlah Penduduk Miskin, (2017)

[9] Badan Pusat Statistik (BPS), Statistik Indonesia 2017: Tingkat Pengangguran Terbuka, (2017)

[10] Badan Pusat Statistik (BPS), Statistik Indonesia 2017: Jumlah Puskesmas, (2017)

[11] Badan Pusat Statistik (BPS), Statistik Kesejahteraan Rakyat 2017: Status Pendidikan Tertinggi, (2017)

[12] Puspitasari, M., Pengelompokan Kabupaten / Kota Berdasarkan Faktor-Faktor Yang Mempengaruhi Kemiskinan Di Jawa Tengah Menggunakan Metode Ward Dan Average Linkage, Jurnal Matematika Vol. 5 No. 6 Universitas Negeri Yogyakarta, (2016)

[13] Tibshirani, R., Walther, G., & Hastie, T., Estimating The Number Of Clusters In A Data Set Via The Gap Statistic, Journal of Royal Statistical Society Vol. 63 Issue 2., (2001)

[14] Laeli, S., Analisis Cluster Dengan Average Linkage Method Dan Ward's Method Untuk Data Responden Nasabah Asuransi Jiwa Unit Link, S1 Thesis, Universitas Negeri Yogyakarta, Indonesia, (2014)

[15] Purnamasari, S.B., Pemilihan Cluster Optimum Pada Fuzzy C-Means (Studi Kasus: Pengelompokan Kabupaten/Kota di

Jawa Tengah berdasarkan Indikator Indeks Pembangunan Manusia), Jurnal Gaussian Vol.3 No.3 Universitas Diponegoro, (2014)

[16] Lailiyah, S. dan Hafiyusholeh, M., Perbandingan antara Metode K-Means Clustering dengan Gath-Geva Clustering, Jurnal Matematika MANTIK, 1(2), Mei 2016. pp. 26-37

Lampiran

Tabel 3. Prevalensi Penderita HIV per100.000 Penduduk Menurut Propinsi Tahun 2016

Provinsi	Rasio Penderita HIV*100.000	Provinsi	Rasio Penderita HIV*100.000
Aceh	1.37	Nusa Tenggara Barat	3.59
Sumatera Utara	13.41	Nusa Tenggara Timur	9.36
Sumatera Barat	7.53	Kalimantan Barat	10.80
Riau	12.64	Kalimantan Tengah	5.53
Jambi	6.22	Kalimantan Selatan	11.19
Sumatera Selatan	4.24	Kalimantan Timur	23.22
Bengkulu	6.04	Kalimantan Utara	24.46
Lampung	4.64	Sulawesi Utara	16.78
Kep. Bangka Belitung	9.63	Sulawesi Tengah	5.37
Kepulauan Riau	51.13	Sulawesi Selatan	11.54
DKI Jakarta	58.56	Sulawesi Tenggara	5.25
Jawa Barat	11.54	Gorontalo	0.61
Jawa tengah	11.85	Sulawesi Barat	1.68
DI Yogyakarta	19.78	Maluku	36.20
Jawa Timur	16.67	Maluku Utara	10.12
Banten	8.95	Papua Barat	59.32
Bali	56.36	Papua	120.53