Simulasi Pemilihan Metode Analisis *Cluster* Hirarki *Agglomerative* Terbaik Antara *Average Linkage* Dan Ward Pada Data Yang Mengandung Masalah Multikolinearitas

Rizki Agung Prabowo[1*], Khoirin Nisa[1], Ahmad Faisol[1] dan Eri Setiawan[1]

[1]Jurusan Matematika, Fakultas MIPA, Universitas Lampung
Jl. Soemantri Brojonegoro 1 Bandar Lampung

***Email korespondensi: agungrizki794@gmail.com**

Abstrak

Multikolinearitas adalah hubungan linear yang ada di antara variabel bebas, pada analisis klaster efek yang ditimbulkan oleh multikolinearitas berbeda, dikarenakan pada dasarnya multikolinearitas adalah bentuk pembobotan implisit. Analisis komponen utama dapat digunakan untuk mereduksi jumlah himpunan peubah yang banyak dan saling berkorelasi menjadi peubah-peubah baru yang tidak berkorelasi dengan mempertahankan sebanyak mungkin keragaman data tersebut, dengan menggunakan hasil analisis komponen utama dilakukan analisis klaster menggunakan metode average linkage dan Ward, yang kemudian akan dipilih metode terbaiknya berdasarkan nilai indeks Dunn dan indeks RS, didapat kesimpulan bahwa metode Ward adalah metode terbaik dibandingkan average linkage yang ditinjau berdasarkan indeks RS, sedangkan dengan menggunakan indeks Dunn didapatkan kesimpulan bahwa metode average linkage adalah metode terbaik dibandingkan Ward.

Kata kunci: multikolinearitas, metode average linkage, metode ward, analisis komponen utama, indeks Dunn, indeks RS.

Abstract

Multicollinearity is a linear relation (collinearity) that exists between independent variables. In cluster analysis, the effect is different because multicollinearity is a form of implicit weighting. Principal component analysis can be used to reduce the number of variables that correlated into the number of new variables that uncorrelated by maintaining as much variety of data, by using the result of principal component analysis, we can do cluster analysis by average linkage and Ward's methods, then the best method will be chosen based on Dunn and RS indices, it was concluded that Ward's method is better than average linkage based on RS index which means that cluster formed using Ward's has more different characteristics than average linkage method while using Dunn index it can be concluded that average linkage is better than Ward's method which means that cluster formed using average linkage has more compactness than Ward's method.

Keywords: ***Multicollinearity, Average Linkage Method, Ward's Method, Principal Component Analysis, Dunn Index, RS Index*.**

1. Pendahuluan

Analisis klaster meneliti seluruh hubungan interdependensi, tidak ada pembedaan variabel bebas, dan tak bebas. Tujuan utama analisis klaster adalah mengelompokkan objek (kasus atau elemen) ke dalam kelompok-kelompok yang relatif homogen didasarkan pada suatu set variabel yang dipertimbangkan untuk diteliti. Pada analisis klaster menggunakan jarak euclid sebagai alat ukur kedekatan, semakin kecil besaran jarak suatu objek terhadap objek lain maka semakin besar kemiripan individu tersebut [1]. Ada dua asumsi yang harus dipenuhi pada analisis klaster, yaitu sampel representatif dan tidak boleh mengandung masalah multikolinearitas antara tiap variabel [2]. Multikolinearitas adalah hubungan linear yang ada di antara variabel independen. Multikoliniearitas dapat dilihat dari nilai *Variance Inflation Factor* (VIF), jika nilai VIF melebihi angka 10 maka dapat disimpulkan ada multikolinearitas [3].

Pada analisis klaster efek yang ditimbulkan oleh multikolinearitas berbeda dengan analisis multivariat yang lainnya, dikarenakan pada dasarnya multikolinearitas adalah bentuk pembobotan implisit pada tiap variabelnya sedangkan pada analisis klaster setiap variabel diberikan bobot yang sama [2]. Masalah multikolinearitas ini dapat diatasi dengan menggunakan analisis komponen utama (AKU) dengan cara mereduksi dimensi suatu data kedalam suatu dimensi seminimal mungkin dengan tetap mempertahankan informasi yang tekandung didalamnya. Pada penelitian ini akan dilakukan pemilihan metode terbaik antara dua metode klaster hirarki

agglomerative average linkage dan Ward dengan menggunakan indeks Dunn dan indeks RS pada data yang mengandung masalah multikolinearitas.

Hal yang paling penting di dalam masalah analisis klaster adalah pemilihan variabel-variabel yang akan dipergunakan untuk pengklasteran. Pada dasarnya set variabel yang akan dipilih harus menguraikan kemiripan (*similarity*) antara objek. Variabel harus dipilih berdasarkan penelitian sebelumnya, teori atau suatu pertimbangan berkenaan dengan hipotesis yang akan diuji [4]. Analisis klaster adalah metode untuk mengukur karakteristik struktural dari serangkaian pengamatan. Pada analisis klaster asumsi seperti normalitas, linearitas dan homoskedastisitas tidak banyak berpengaruh. Ada dua asumsi yang harus dipenuhi pada analisis klaster, yaitu sampel representatif dan multikolinearitas antara tiap variabel [2]. Pada penelitian ini hanya difokuskan pada asumsi multikolinearitas.

Multikolinearitas adalah hubungan linear yang ada di antara variabel independen. Multikoliniearitas dapat dilihat dari nilai *Variance Inflation Factor* (VIF). Rumus untuk menghitung VIF yaitu sebagai berikut:

$$VIF_i = \frac{1}{(1 - R_i^2)} \tag{1}$$

dengan R_i^2 menyatakan koefisien determinasi pada variabel i. Jika nilai VIF melebihi angka 10 maka dapat disimpulkan ada multikolinearitas [3].

Analisis klaster berupaya mengidentifikasi dari vektor-vektor pengamatan yang serupa dan mengelompokkannya menjadi kelompok-kelompok, banyak teknik menggunakan indeks kesamaan atau kedekatan antara setiap pasang pengamatan. Kedekatan atau pendekatan yang biasa digunakan adalah mengukur kemiripan yang dinyatakan dalam jarak antara pasangan objek. Semakin kecil besaran jarak suatu individu terhadap individu lain, maka semakin besar kemiripan individu tersebut, sehingga individu tersebut akan dimasukkan dalam kelompok yang sama [1]. Pada analisis klaster digunakan jarak Euclid sebagai alat ukur kedekatan, yang didefinisikan sebagai berikut:

$$d(i,j) = d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{2}$$

dengan d_{ij} menyatakan jarak antara objek ke-i dan obyek ke-j, dan p menyatakan jumlah variabel klaster. Sedangkan x_{ik} menyatakan data dari objek ke-i pada variabel ke-k.

Variabel yang digunakan untuk menghitung jarak euclid tidak boleh mengandung korelasi antar variabel. Pembentukan klaster hirarki mempunyai sifat sebagai pengembangan suatu hirarki atau struktur mirip pohon bercabang. Metode klaster hirarki merupakan metode pengelompokan yang mana jumlah kelompok yang akan dibuat belum diketahui, teknik ini diproses melalui penggabungan berurutan (*agglomerative*) atau pembagian berurutan (*divissive*). Teknik *agglomerative* terdiri atas 3 metode, yaitu metode *Linkage*, metode *Variance* dan metode *Centroid*. Metode Linkage terdiri dari metode *Single Linkage*, *Complete Linkage* dan *Average Linkage*, sedangkan metode *Variance* terdiri atas metode Ward. Metode *agglomerative* dimulai dengan menganggap bahwa tiap objek adalah sebuah klaster. Pada pendekatan *average linkage*, jarak antara dua klaster didefinisikan sebagai jarak rata-rata antara semua anggota didalam satu klaster dengan semua anggota pada klaster lain [5]. *Average linkage* menggunakan jarak terdekat dan metode ini dapat digunakan untuk mengelompokan objek atau variable

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W} \tag{3}$$

dimana, $d_{(UV)W}$ menyatakan jarak antar klaster (UV) dan klaster W, d_{ik} menyatakan jarak antar objek i pada klaster (UV) dan objek k pada klaster W, $N_{(UV)}$ menyatakan jumlah item pada klaster (UV) serta N_W menyatakan jumlah item pada klaster W.

Menurut Johnson dan Wichern, metode Ward mempertimbangkan pengelompokan secara hirarki berdasarkan meminimalkan informasi yang hilang dalam menggabungkan dua grup [5]. Metode Ward didasarkan pada kriteria *sum of square error* (SSE) dengan ukuran kehomogenan antara dua objek berdasarkan jumlah kuadrat kesalahan yang paling minimal, SSE hanya dapat dihitung jika klaster memiliki elemen lebih dari satu objek. Formula untuk menghitung SSE adalah sebagai berikut:

$$SSE = \sum_{i=1}^{N} (\boldsymbol{y_i} - \bar{\boldsymbol{y}})'(\boldsymbol{y_i} - \bar{\boldsymbol{y}}) \quad (4)$$

Dimana $\boldsymbol{y_i}$ adalah vektor kolom yang entrinya nilai rata-rata objek i, $\bar{\boldsymbol{y}}$ adalah vektor kolom yang entrinya rata-rata nilai objek dalam klaster, dan N menyatakan banyaknya objek.

Indeks Dunn adalah salah satu pengukuran validitas klaster yang diajukan oleh J.C. Dunn. Ukuran validitas klaster ini berlandaskan pada fakta bahwa klaster yang terpisah itu biasanya memiliki jarak antar klaster yang besar dan diameter intra klaster yang kecil [6]. Indeks Dunn dapat dituliskan sebagai berikut:

$$D = \min_{j=i+1,\dots,n_c} \left(\frac{d(c_i, c_j)}{\max_{k=1,\dots,n_c} (diam(c_k))} \right) \quad (5)$$

Dimana nilai $d(c_i, c_j)$ dan $diam(c_k)$ didefinisikan sebagai berikut:

$$d(c_i, c_j) = \min \left(\begin{matrix} x \in c_i \\ y \in c_j \end{matrix} (d(x, y)) \right) \quad (6)$$

$$diam(c_k) = \max \left(\begin{matrix} x \in c_i \\ y \in c_i \end{matrix} (d(x, y)) \right) \quad (7)$$

Nilai pada indeks Dunn ini jika nilainya semakin besar, maka hasil klaster akan semakin baik. Indeks Dunn memiliki rentang nilai dari nol sampai tak hingga.

Menurut Sharma di dalam [7], indeks RS dapat didefinisikan sebagai berikut:

$$RS = \frac{SS_B}{SS_T} = \frac{SS_T - SS_W}{SS_T} = \frac{\left(\sum_{j=1}^{n} (x_j - \bar{x})^2\right) - \left(\sum_{i=1}^{n_c} \sum_{j=1}^{r_i} (x_{ij} - \bar{x})^2\right)}{\left(\sum_{j=1}^{n} (x_j - \bar{x})^2\right)} \quad (8)$$

dimana, x_j adalah data ke – j pada variabel dan x_{ij} adalah data ke-j pada variabel di masing-masing klaster ke-i. Semakin besar nilai RS maka klaster yang dihasilkan akan semakin baik. RS memiliki rentang nilai dari nol sampai satu.

Menurut Johnson dan Wichern [5], analisis komponen utama (AKU), merupakan analisis tertua dalam APG yang diperkenalkan oleh Karl Pearson tahun 1901, yang biasanya digunakan untuk Mereduksi jumlah himpunan peubah yang banyak dan saling berkorelasi menjadi peubah-peubah baru yang tidak berkorelasi dengan mempertahankan sebanyak mungkin keragaman data tersebut.

2. Metodologi Penelitian

Data yang digunakan dalam penelitian ini adalah hasil dari membangkitkan data yang mengandung masalah multikolinearitas dengan menggunakan *software* RStudio versi 1.2.1335 dan menggunakan beberapa *package* yang disediakan oleh *software* RStudio. Menurut Kibria dan Muniz, untuk mendapatkan data yang mengandung multikolinearitas pada setiap himpunan data X_{ij} dibangkitkan menggunakan simulasi Monte Carlo dengan persamaan sebagai berikut [8]:

$$X_{ij} = \sqrt{(1 - \rho^2)}\, x_{ij} + \rho x_{ip} \quad (9)$$

dengan $i = 1, 2, 3, \dots, n$ dan $j = 1, 2, 3, \dots, p$. Adapun x_{ij} dibangkitkan berdistribusi normal dengan μ dan σ ditentukan. Berikut adalah langkah-langkah simulasi yang dilakukan

a) Membangkitkan data $\boldsymbol{X_i} \sim N_p(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$, $\boldsymbol{X_i} = [X_1, X_2, X_3, \dots, X_p]^t$ adalah vektor pengamatan dengan $i = 1, 2, 3, \dots$ dibangkitkan secara acak yang kemudian dikonversi menjadi data multikolinearitas dengan $\rho^2 = 0.96$ dan ketentuan:
   1) Data ke-1 dibangkitkan dengan $n = 10$ yang membentuk dua klaster. Klaster pertama dengan $n = 5$ berdistribusi $X_j \sim N(0,1)$ sehingga $\boldsymbol{X_1} \sim N_4(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$ dengan $\boldsymbol{\mu_1} = [0,0,0,0]^t$ dan $\boldsymbol{\Sigma_1} = 1\mathbf{I_4}$ dimana $j = 1, 2, 3, 4$. Klaster kedua dengan $n = 5$ objek berdistribusi $X_j \sim N(5,2)$ sehingga $\boldsymbol{X_2} \sim N_4(\boldsymbol{\mu_2}, \boldsymbol{\Sigma_2})$ dengan $\boldsymbol{\mu_2} = [5,5,5,5]^t$ dan $\boldsymbol{\Sigma_2} = 2\mathbf{I_4}$ dimana $j = 1, 2, 3, 4$ yang kemudian gabungan data $\boldsymbol{X_1}$ dan $\boldsymbol{X_2}$ dikonversi menjadi data multikolinearitas dengan menggunakan persamaan (9)
   2) Data ke-2 dibangkitkan dengan $n = 20$ yang membentuk tiga klaster. Klaster pertama dengan $n = 7$, berdistribusi sama dengan klaster pertama pada poin a. Klaster kedua dengan $n = 7$, berdistribusi sama dengan klaster kedua pada poin a. Klaster ketiga dengan $n = 6$ berdistribusi $X_j \sim N(8,3)$ sehingga

$\boldsymbol{X_3} \sim N_4(\boldsymbol{\mu_3}, \boldsymbol{\Sigma_3})$ dengan $\boldsymbol{\mu_3} = [8,8,8,8]^t$ dan $\boldsymbol{\Sigma_3} = 3\mathbf{I_4}$ dimana $j = 1,2,3,4$ yang kemudian gabungan data $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ dan $\boldsymbol{X_3}$ dikonversi menjadi data multikolinearitas dengan menggunakan persamaan (9)

3) Data ke-3 dibangkitkan dengan $n = 50$ yang membentuk empat klaster. Klaster pertama dengan $n = 12$, berdistribusi sama dengan klaster pertama pada poin a. Klaster kedua dengan $n = 12$, berdistribusi sama dengan klaster kedua pada poin a. Klaster ketiga dengan $n = 13$, berdistribusi sama dengan klaster ketiga pada poin b. Klaster keempat dengan $n = 13$ berdistribusi $X_j \sim N(10,4)$ sehingga $\boldsymbol{X_4} \sim N_4(\boldsymbol{\mu_4}, \boldsymbol{\Sigma_4})$ dengan $\boldsymbol{\mu_4} = [10,10,10,10]^t$dan $\boldsymbol{\Sigma_4} = 4\mathbf{I_4}$ dimana $j = 1,2,3,4$ yang kemudian gabungan data $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$ dan $\boldsymbol{X_4}$ dikonversi menjadi data multikolinearitas dengan menggunakan persamaan (9)

4) Data ke-4 dibangkitkan dengan $n = 100$ yang membentuk lima klaster. Klaster pertama dengan $n = 20$, berdistribusi sama dengan klaster pertama pada poin a. Klaster kedua dengan $n = 20$, berdistribusi sama dengan klaster kedua pada poin a. Klaster ketiga dengan $n = 20$, berdistribusi sama dengan klaster ketiga pada poin b. Klaster keempat dengan $n = 20$, berdistribusi sama dengan klaster keempat pada poin c. Klaster kelima dengan $n = 20$ berdistribusi $X_j \sim N(13,3)$ sehingga $\boldsymbol{X_5} \sim N_4(\boldsymbol{\mu_5}, \boldsymbol{\Sigma_5})$ dengan $\boldsymbol{\mu_5} = [13,13,13,13]^t$dan $\boldsymbol{\Sigma_5} = 3\mathbf{I_4}$ dimana $j = 1,2,3,4$ yang kemudian gabungan data $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{X_4}$ dan $\boldsymbol{X_5}$ dikonversi menjadi data multikolinearitas dengan menggunakan persamaan (9).

b) Standarisasi data kedalam bentuk nilai Z
c) Melakukan uji asumsi multikolinearitas (VIF)
d) Mengatasi data yang mengandung multikolinearitas menggunakan analisis komponen utama (AKU)
e) Melakukan pengklasteran dengan menggunakan metode *average linkage* dan metode Ward dengan menggunakan data hasil analisis komponen utama
f) Menghitung dan mencatat indeks Dunn pada tiap metode
g) Menghitung dan mencatat indeks RS pada tiap metode
h) Mengulang langkah 1 (satu) sampai langkah 7 (tujuh) sebanyak 1000 (seribu) kali pengulangan
i) Melakukan evaluasi indeks Dunn dan indeks RS dengan menghitung rata-ratanya
j) Analisis hasil

3. Hasil dan Pembahasan

Pada hasil dan pembahasan kali ini akan dilakukan perhitungan menggunakan contoh data yang dibangkitkan dengan $n = 10$, berikut adalah gabungan contoh data yang dibangkitkan dengan $n = 10$ yang membentuk dua klaster:

Tabel 1. Contoh data awal yang mengandung multikolinearitas dengan $\boldsymbol{n = 10}$

Data ke-	$\boldsymbol{X_1\_mul}$	$\boldsymbol{X_2\_mul}$	$\boldsymbol{X_3\_mul}$	$\boldsymbol{X_4\_mul}$
1	0,0165	0,3307	-0,4453	0,1037
2	1,0044	0,8508	0,8556	1,0094
3	0,0767	-0,1488	0,0789	0,0864
4	-0,1254	-0,1640	0,1670	-0,1455
5	-0,9996	-1,0476	-1,0878	-1,2508
6	5,7694	5,5857	6,4945	6,2600
7	5,9804	6,4656	6,0088	5,9094
8	7,7154	7,8891	8,3051	8,2062
9	6,5936	5,2894	5,9093	6,4151
10	5,0673	5,8423	5,5634	5,7868

Standarisasi data dilakukan apabila terdapat perbedaan satuan antar variabel maupun di dalam variabel itu sendiri, berikut adalah hasil standarisasi data pada Tabel 1:

Tabel 2. Contoh data multikolinearitas dengan $\boldsymbol{n = 10}$ yang telah terstandarisasi

Data ke-	$\boldsymbol{X_1}$_stndr	$\boldsymbol{X_2}$_stndr	$\boldsymbol{X_3}$_stndr	$\boldsymbol{X_4}$_stndr
1	-0,924	-0,786	-1,063	-0,881
2	-0,620	-0,667	-0,663	-0,627
3	-0,899	-0,972	-0,868	-0,886

4	-0,956	-0,956	-0,815	-0,951
5	-1,197	-1,198	-1,181	-1,262
6	0,764	0,694	0,954	0,850
7	0,879	1,070	0,806	0,751
8	1,347	1,413	1,451	1,397
9	1,073	0,564	0,725	0,893
10	0,532	0,838	0,653	0,717

Uji asumsi multikolinearitas dilakukan untuk melihat apakah ada hubungan linear di antara variabel atau tidak. Berdasarkan contoh satu gugus data yang sudah dibangkitkan dan telah dilakukan standarisasi kedalam bentuk nilai Z didapat nilai VIF sebagai berikut:

Tabel 3. Output nilai VIF dan kesimpulan

n	X_1_stndr	X_2_stndr	X_3_stndr	X_4_stndr	Kesimpulan
10	94,389	35,899	87,406	225,398	Multikolinearitas

Data multikolinearitas yang telah dicek nilai VIF nya kemudian dilakukan analisis komponen utama untuk mereduksi empat variabel yang saling berkorelasi menjadi dua variabel komponen utama yang tidak saling berkorelasi. Hasil analisis komponen utama, dengan menggunakan contoh data yang telah terstandarisasi dengan $n = 10$ pada Tabel 2 memberikan matriks kovarian

$$\boldsymbol{S} = \begin{bmatrix} 1 & 0{,}976 & 0{,}987 & 0{,}994 \\ 0{,}976 & 1 & 0{,}983 & 0{,}985 \\ 0{,}987 & 0{,}983 & 1 & 0{,}994 \\ 0{,}994 & 0{,}985 & 0{,}994 & 1 \end{bmatrix}$$

Dengan nilai eigen dan proporsi keragaman disajikan pada Tabel 4. Untuk mendapatkan plot sebaran data berdasarkan dua komponen utama, maka diperlukan dua komponen utama yang berpadanan dengan λ_1 dan λ_2 dengan proporsi kumulatif sebesar 99,6%. Vektor eigen yang berpadanan dengan nilai eigen λ_1 dan λ_2 diberikan oleh $\boldsymbol{a_1} = [0{,}450 \quad 0{,}498 \quad 0{,}500 \quad 0{,}502]$ dan $\boldsymbol{a_2} = [0{,}527 \quad -0{,}820 \quad 0{,}086 \quad 0{,}203]$.

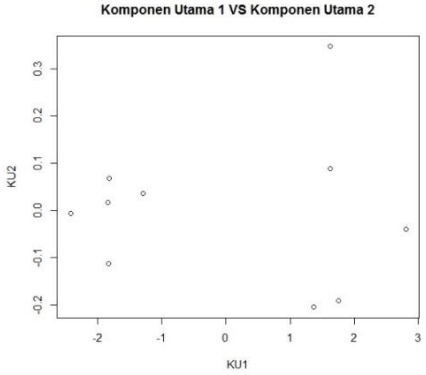
Tabel 4. Nilai eigen dan proporsi keragaman yang dijelaskan setiap nilai eigen

No.	Eigen	Nilai Eigen	Proporsi	Proporsi kumulatif
1.	λ_1	3,960	99%	99%
2.	λ_2	0,025	0,6%	99,6%
3.	λ_3	0,011	0,3%	99,9%
4.	λ_4	0,003	0,1%	100%

Adapun komponen utama yang dihasilkan adalah

$$KU_1 = \begin{bmatrix} -1{,}827 \\ -1{,}288 \\ \vdots \\ 1{,}369 \end{bmatrix} \text{ dan } KU_1 = \begin{bmatrix} -1{,}827 \\ -1{,}288 \\ \vdots \\ 1{,}369 \end{bmatrix}$$

Plot data KU disajikan pada Gambar 1 berikut.



Gambar 1. Plot KU_1 dan KU_2

Hasil pengklasteran menggunakan metode *average linkage* dan Ward terhadap gabungan data KU_1 dan KU_2 diperoleh Klaster satu terdiri atas objek 1, 2, 3, 4 dan 5 sedangkan klaster dua terdiri atas objek 6, 7, 8, 9, dan 10. Indeks Dunn dan RS untuk metode *average linkage* dan Ward disajikan pada Tabel 5:

Tabel 5. Indeks Dunn dan RS

Indeks	**Metode *Average Linkage***	**Metode Ward**
Indeks Dunn	1,8482	1,8482
Indeks RS	0,9410	0,9410

Pengulangan sebanyak 1000 kali diartikan sebagai cerminan 1000 kasus penelitian yang sama dengan data kasus yang berbeda-beda yang bertujuan untuk menarik kesimpulan yang bersifat umum dari kasus penelitian yang sama tersebut. Proses pengulangan pada penelitian ini akan menghasilkan indeks Dunn dan indeks RS pada masing-masing metode sebanyak 1000 indeks, yaitu 1000 indeks pada metode *average linkage* dan 1000 indeks pada metode Ward dan mengikuti prosedur seperti langkah sebelumnya, didapat hasil indeks Dunn dan indeks RS yaitu sebagai berikut:

Tabel 6. Nilai indeks Dunn dan indeks RS dilakukan pengulangan sebanyak 1000 kali

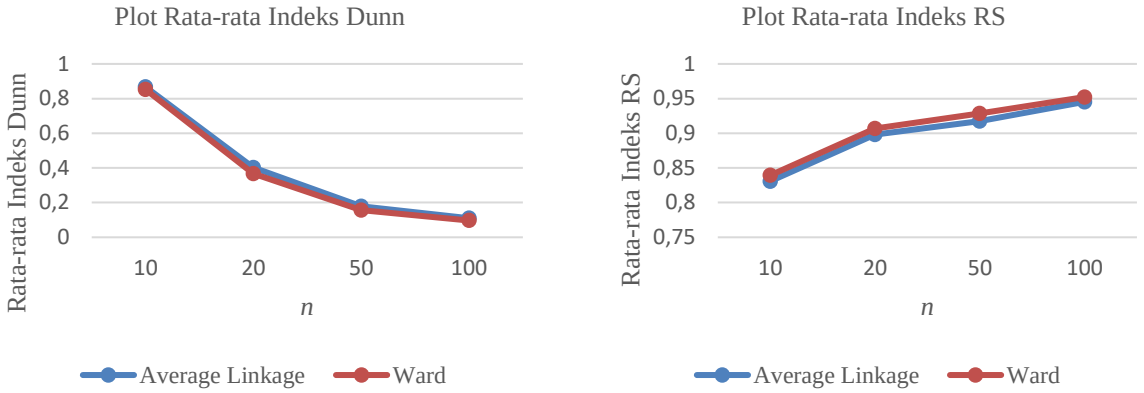
n	**Pengulangan ke-**	**Indeks Dunn**		**Indeks RS**	
		Metode *Average Linkage*	**Metode Ward**	**Metode *Average Linkage***	**Metode Ward**
10	1	0,6868	0,6868	0,9410	0,9410
	2	0,5723	0,1954	0,8700	0,8700
	3	0,6855	0,6855	0,7481	0,7772
	⋮	⋮	⋮	⋮	⋮
	1000	2,0518	2,0518	0,7970	0,7970
20	1	0,3502	0,3502	0,8494	0,8533
	2	0,2718	0,0948	0,8795	0,9036
	3	0,3000	0,3000	0,9118	0,9412
	⋮	⋮	⋮	⋮	⋮
	1000	0,4790	0,4182	0,8437	0,9154
50	1	0,2644	0,2644	0,9507	0,9507
	2	0,1644	0,1337	0,9040	0,9344
	3	0,2443	0,2040	0,8948	0,9216
	⋮	⋮	⋮	⋮	⋮
	1000	0,2165	0,1369	0,9141	0,9176
100	1	0,1385	0,1189	0,9507	0,9507
	2	0,0909	0,0651	0,9040	0,9344
	3	0,1003	0,1003	0,9019	0,9325
	⋮	⋮	⋮	⋮	⋮
	1000	0,1165	0,0974	0,9141	0,9176

Indeks Dunn dan RS pada Tabel 6 akan berubah-ubah apabila dilakukan proses simulasi pengulangan kembali dikarenakan data yang dibangkitkan bersifat acak, oleh karena itu dilakukan penarikan kesimpulan dengan cara menghitung nilai rata-rata indeks Dunn dan RS pada masing-masing metode analisis klaster.

Analisis hasil dilakukan dengan cara menghitung nilai rata-rata indeks Dunn dan indeks RS pada masing-masing metode analisis klaster, hasil nilai rata-rata indeks Dunn dan indeks RS adalah sebagai berikut:

Tabel 7. Rata-rata nilai indeks Dunn dan indeks RS dilakukan pengulangan sebanyak 1000 kali

n	**Pengulangan**	**Rata-rata Indeks Dunn**		**Rata-rata Indeks RS**	
		Metode *Average Linkage*	**Metode Ward**	**Metode *Average Linkage***	**Metode Ward**
10	1000	0,8668	0,8523	0,8308	0,8393
20	1000	0,4016	0,3676	0,8982	0,9068
50	1000	0,1781	0,1569	0,9177	0,9285
100	1000	0,1099	0,0968	0,9453	0,9522



Gambar 2. Plot Rata-rata indeks Dunn dan RS

Rata-rata nilai indeks Dunn pada Tabel 7 disetiap jumlah objek yang berbeda dan jumlah simulasi pengulangan sebanyak 1000 kali menunjukan bahwa metode *average linkage* adalah metode analisis klaster hirarki terbaik dibandingkan metode Ward pada data yang mengandung masalah multikolinearitas, hal ini dikarenakan nilai rata-rata indeks Dunn pada metode *average linkage* lebih besar dibandingkan metode Ward.

Rata-rata indeks RS pada Tabel 7 disetiap jumlah objek yang berbeda dan jumlah simulasi pengulangan sebanyak 1000 kali menunjukan bahwa metode Ward adalah metode analisis klaster hirarki terbaik dibandingkan metode *average linkage* pada data yang mengandung masalah multikolinearitas, hal ini dikarenakan nilai rata-rata indeks RS pada metode Ward lebih besar dibandingkan metode RS.

4. Kesimpulan

Berdasarkan hasil dan pembahasan, maka diperoleh kesimpulan bahwa analisis klaster metode *average linkage* dan Ward pada data yang mengandung multikolinearitas dapat diatasi dengan analisis komponen utama. Berdasarkan nilai indeks Dunn, metode *average linkage* memberikan hasil yang lebih baik dibandingkan metode Ward dalam pengklasteran data. Ukuran indeks Dunn berlandaskan pada fakta bahwa klaster yang terpisah itu biasanya memiliki jarak antar klaster yang besar dan diameter intra klaster yang kecil, yang berarti klaster-klaster yang dibentuk oleh metode *average linkage* memiliki jarak antar klaster yang lebih besar dan diameter intra klaster yang lebih kecil dibandingkan metode Ward. Berdasarkan nilai indeks RS, metode Ward memberikan hasil yang lebih baik dibandingkan metode *average linkage* dalam pengklasteran data. Indeks RS mengukur apakah karakteristik antar klaster saling berbeda, yang berarti klaster yang terbentuk dengan menggunakan metode Ward memiliki karakteristik yang lebih berbeda dibanding dengan metode *average linkage*

Daftar Pustaka:

[1] Usman, H. dan Nurdin, S. 2013. Aplikasi Teknik Multivariate untuk Riset Pemasaran. PT. Raja Grafindo Persada, Jakarta.
[2] Hair, J.F., Black, W.C., Babin, B.J. dan Anderson, R.E. 2014. *Multivariate Data Analysis*. 7th Edition. Pearson Education Limited, England.
[3] Widarjono, A. 2010. *Analisis Statistika Multivariat Terapan*. UPP STIM YKPN, Yogyakarta.
[4] Supranto. 2004. *Analisis Multivariat: Arti dan Interpretasi*. Rineka Cipta, Jakarta.
[5] Johnson, R. dan Wichern, D. 2007. *Applied Multivariate Analysis*. 6th Edition. Prentice Hall Inc., New Jersey
[6] Satoto, B.D., Khotimah, B.K. dan Muhammad, A. 2015. Pengelompokan Tingkat Kesehatan Masyarakat Menggunakan *Shelf Organizing Maps* dengan *Cluster Validation Idb* dan *I-Dunn*. *Seminar Nasional Aplikasi Teknologi Informasi (SNATi) 2015.*
[7] Sharma, S. 1996. *Applied Multivariate Techniques*. A John Wiley & Sons, Inc., Canada.
[8] Kibria, B. dan Muniz, G. 2009. On Some Ridge Regression Estimator: An Emprical Comparison. *Communication in Statistics – Simulation and Computation*. **38**: 621-630.