ESTIMASI: Journal of Statistics and Its Application
Vol. 5, No. 1, Januari, 2024, Hal. 61-74
DOI: https://doi.org/10.20956/ejsa.v5i1.13605

Implementasi Algoritma Centroid Linkage dan K-Medoids dalam Mengelompokkan Kabupaten/Kota di Sulawesi Selatan Berdasarkan Indikator Pendidikan

Nur Alfianingsih Raja[1*], Georgina Maria Tinungki[2], Nasrah Sirajang[3]
[123]Departemen Statistika, Fakultas MIPA, Universitas Hasanuddin, Makassar, 90245, Indonesia
*** Corresponding author, email: nuralfianingsihraja@gmail.com**

Abstract

Cluster analysis is a multivariate analysis technique that aims to cluster the observational data or variables into clusters in such a way that each cluster is homogeneous according to the factors used for clustering. This study used the Centroid linkage algorithm that was useful for forming groups based on the distance between centroids and the K-Medoids algorithm that was based on the use of the most centered object (medoid) to group districts/cities and obtained comparison results based on the education indicator data in South Sulawesi. The implementation of the Centroid Linkage Algorithm and K-Medoids on the education indicator data in South Sulawesi in 2018, showed that the grouping of districts/cities in South Sulawesi produced 2 clusters with cluster 1 of 21 districts/cities, and cluster 2 of 3. To determine the best method, it was seen from the value of the Standard Deviation ratio in the cluster 〖(S〗_W) and Standard Deviation between Clusters 〖(S〗_B) showed the same standard deviation ratio (S) in the Centroid Linkage algorithm and K-Medoids that was equal to 104,967.

***Keywords**: Centroid Linkage, Cluster Analysis, K-Medoids, Standard Deviation.*

Abstrak

Analisis cluster merupakan suatu teknik analisis multivariat yang bertujuan untuk mengclusterkan data observasi ataupun variabel-variabel ke dalam cluster sedemikian rupa sehingga masing-masing cluster bersifat homogen sesuai dengan faktor yang digunakan untuk melakukan pengclusteran. Penelitian ini menggunakan algoritma Centroid linkage yang berguna untuk membentuk kelompok berdasarkan jarak antar centroidnya dan algoritma K-Medoids yang didasarkan pada penggunaan objek yang paling terpusat (medoid) untuk mengelompokkan kabupaten/kota dan memperoleh hasil perbandingan berdasarkan data indikator pendidikan di Sulawesi Selatan. Implementasi Algoritma Centroid Linkage dan K-Medoids pada data indikator pendidikan di Sulawesi selatan tahun 2018 yakni pengelompokkan kabupaten/kota di Sulawesi Selatan yang menghasilkan 2 cluster dengan cluster 1 sebanyak 21 kabupaten/kota, cluster 2 sebanyak 3. Untuk menentukan metode terbaik dilihat dari nilai rasio Simpangan Baku dalam Cluster (S_W) dan Simpangan Baku antar Cluster (S_B) yaitu memiliki nilai rasio simpangan baku (S) yang sama pada algoritma Centroid Linkage dan K-Medoids sebesar 104.967.

Kata Kunci: Analisis Cluster, Centroid Linkage, K-Medoids, Simpangan baku.

1. Pendahuluan

Pendidikan adalah adalah proses pengubahan sikap dan tata laku seseorang atau kelompok orang dalam usaha mendewasakan manusia melalui upaya pengajaran dan pelatihan. Pendidikan berkualitas merupakan prinsip dasar pembangunan untuk menciptakan sumber daya manusia yang berdaya saing tinggi. Disamping itu, akses untuk memperoleh kesempatan belajar yang sama dan merata dalam setiap jenjang pendidikan

diharapkan dapat tercapai guna menjamin kualitas pendidikan yang inklusif dan merata. Pendidikan juga menjadi salah satu tujuan dalam Sustainable Development Goals (SDG's) [1].

Pentingnya peran pendidikan terhadap kemajuan bangsa, pengukuran dan penghitungan indikator-indikator pendidikan perlu dilakukan untuk melihat sejauh mana pemerataan pendidikan. Agar dapat mengetahui pemerataan pendidikan atau karakteristik indikator pendidikan menurut jenis kelamin dan tingkat pendidikan maka dilakukan pengelompokan menggunakan analisis *cluster*. Analisis *cluster* merupakan suatu metode untuk mengelompokkan objek atau unit penelitian ke dalam beberapa kelompok dan setiap unit penelitian dalam suatu kelompok akan mempunyai ciri yang relatif sama sedangkan antar kelompok unit penelitian memiliki sifat yang berbeda.

Analisis *cluster* dibagi menjadi dua metode yaitu *hierarchical method* dan *non-hierarchical method*. Dalam *hierarchical method* jumlah kelompok yang akan diperoleh belum diketahui, sedangkan dalam *non-hierarchical method* diasumsikan ada *k* kelompok terlebih dahulu. *Hierarchical method* terdiri dari *Single Linkage*, *Complete Linkage*, *Centroid Linkage*, *Average Linkage* dan *Ward's Method*. Sedangkan metode yang termasuk *non-hierarchical method* adalah metode *K-Means* dan *K-Medoids* [2].

Pada penelitian ini, akan digunakan pendekatan *hierarchical method* yaitu *Centroid Linkage*. Metode ini dikenal lebih memiliki beban komputasi yang relatif lebih ringan karena hanya cukup menghitung titik tengah antar *cluster*. Metode ini juga baik untuk data yang mengandung outlier. Selain itu, untuk mendapatkan keterbandingan dengan metode yang lain akan digunakan *non-hierarchical method* yaitu metode *K-Medoids*. Metode *k-medoids* ini menggunakan objek pada kumpulan objek untuk mewakili sebuah *cluster*. Kelebihan dari metode ini mampu mengatasi kelemahan dari metode *k-means* yang *sensitive* terhadap *outlier* dan hasil proses *clustering* tidak bergantung pada urutan masuk pada suatu himpunan data [3].

Penelitian *centroid linkage* sebelumnya dilakukan oleh Rini Silvi (2018) dengan judul Analisis Cluster dengan Data Outlier Menggunakan *Centroid Linkage* dan *K-Means Clustering* untuk Pengelompokan Indikator HIV/AIDS di Indonesia. Perbandingan rasio S_W/S_B, *Centroid Linkage* lebih homogen dibandingkan *K-means*. Adapun penelitian yang menggunakan metode *k-medoids clustering* dilakukan oleh Dini Marlina (2018) dengan judul Implementasi Algoritma *K-Medoids* dan *K-Means* untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak. Validitas yang digunakan pada penelitian ini adalah validitas *Silhouette Coefficient* dan dihasilkan *K-Medoids* lebih baik dalam melakukan pengelompokan pada data sebaran Anak Cacat dibandingkan dengan algoritma *K-Means*. Dalam penelitian Elok Fitriani (2016) yang berjudul "Pengelompokan Kabupaten/Kota di Jawa Timur berdasarkan Indikator Pendidikan Tahun 2013 menggunakan Analisis *Hierarchical Cluster*" dan terbentuk dua *cluster* yakni *cluster 1* dengan 29 Kabupaten memiliki tingkat pendidikan lebih tinggi dibandingan *cluster* 2 dengan 9 kabupaten.

2. Material dan Metode

2.1 Sumber Data

Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari publikasi Badan Pusat Statistik Provinsi Sulawesi Selatan yaitu data persentase indikator pendidikan penduduk (15 Tahun+) menurut jenis kelamin dan ijazah/STTB (Surat Tanda Tamat Belajar) tertinggi yang dimiliki Tahun 2018.

2.2 Analisis *Cluster*

Analisis *cluster* merupakan suatu teknik analisis multivariat yang bertujuan untuk meng*cluster*kan data observasi ataupun variabel-variabel ke dalam *cluster* sedemikian rupa sehingga masing-masing *cluster* bersifat homogen sesuai dengan faktor yang digunakan untuk melakukan peng*cluster*an. Karena yang diinginkan adalah untuk mendapatkan *cluster* yang sehomogen mungkin, maka yang digunakan sebagai dasar untuk meng*cluster*kan adalah kesamaan skor nilai yang dianalisis. Data mengenai ukuran kesamaan tersebut dapat dianalisis dengan analisis *cluster* sehingga dapat ditentukan objek yang masuk dalam *cluster* [4].

2.3 Jarak Euclidean

Jarak *Eculidean* adalah jarak yang paling umum dan paling sering digunakan dalam analisis *cluster.* Jarak antara dua buah *cluster* diukur sebagai jarak *Euclidean* antara kedua rataan (*centroid*) *cluster*. Jarak ini harus memenuhi asumsi bahwa semua variabel yang diamati tidak berkorelasi dan antar variabel memiliki satuan yang sama [5]. Jarak *Euclidean* antara *cluster* ke-*i* dan ke-*j* dari *M* variabel didefinisikan:

$$d_{ij} = \sqrt{\sum_{p=1}^{M} \left(x_{ip} - x_{jp}\right)^2} \tag{1}$$

2.4 *Centroid Linkage*

Centroid linkage merupakan salah satu metode *hierarchical clustering* yang berguna untuk membentuk kelompok berdasarkan jarak antar *centroid*nya. *Centroid* adalah rata-rata dari semua anggota dalam *cluster* tersebut. Metode ini dibangun dengan memperhatikan pengecilan nilai standar deviasi *cluster* sekecil-kecilnya. Metode ini menggabungkan dua *cluster* melalui jarak terdekat di antara titik pusat antar *cluster*. Metode ini sangat ampuh untuk memperkecil *variance within cluster* karena melibatkan titik pusat pada saat penggabungan antar *cluster*. Penggabungan *cluster* pada *centroid linkage* didasarkan pada lokasi titik pusat yang terbentuk pada tahap sebelumnya. Metode ini juga baik untuk data yang mengandung outlier [6].

Langkah-langkah *Centroid Linkage* [2]:

a. Asumsikan setiap data merupakan *cluster*.
b. Menghitung jarak antar *cluster* dengan *Euclidean Distance* dengan persamaan:

$$d_{ij} = \sqrt{\sum_{p=1}^{M} \left(x_{ip} - x_{jp}\right)^2} \quad (2)$$

$i = 1,2,3, \dots, N$; $j = 1,2,3, \dots, N$ dan *M* adalah banyaknya variabel.

c. Memilih jarak terkecil antar *cluster* lalu menggabungkan kedua objek yang memiliki jarak terkecil tersebut. Misalkan *cluster* U dan *cluster* V memiliki jarak terdekat, maka U dan V bergabung dalam satu *cluster*.
d. Menghitung *centroid* dari U dan V dengan rumus :

$$X_{(UV)} = \frac{(N_U \times \bar{x}_U) + (N_V \times \bar{x}_V)}{N_U + N_V} \quad (3)$$

dengan :
N_U, N_V = Banyak objek pada *cluster* ke U dan V
$\bar{x}_U$, $\bar{x}_V$ = Rata-rata objek pada *cluster* ke U dan V

e. Bentuk matriks data baru dengan data dari *cluster* gabungan U dan V yang diperoleh dari langkah keempat.
f. Ulangi langkah kedua, demikian seterusnya sampai semua data bergabung dengan jumlah *cluster* yang diinginkan.

2.5 *K-Medoids*

K-Medoids juga dikenal sebagai *Partitioning Around Medoids* (PAM) adalah varian dari metode *K-Means*. Hal ini didasarkan pada penggunaan objek yang paling terpusat (*medoid*) bukan dari objek rata-rata (*mean*) yang dimiliki oleh setiap *cluster*, dengan tujuan mengurangi sensitivitas dari partisi sehubungan dengan nilai ekstrim yang ada dalam himpunan data [7]. Algoritma *K-Medoids* menggunakan metode partisi *clustering* untuk mengelompokkan sekumpulan *N* objek menjadi sejumlah *K cluster*. Algoritma ini menggunakan objek pada kumpulan objek untuk mewakili sebuah *cluster*. Objek yang terpilih untuk mewakili sebuah *cluster* disebut dengan *medoid*. *Cluster* dibangun dengan menghitung kedekatan yang dimiliki antara *medoid* dengan objek *non-medoid* [3].

Langkah-langkah algoritma *K-Medoids* [8]:

a. Inisialisasi pusat *cluster* sebanyak *K* (jumlah *cluster*).
b. Mengalokasikan setiap data (objek) ke *cluster* terdekat menggunakan persamaan ukuran jarak *Euclidean Distance* dengan persamaan:

$$d_{ij} = \sqrt{\sum_{p=1}^{M} \left(x_{ip} - x_{jp}\right)^2} \quad (4)$$

$i = 1,2,3, \dots, N$; $j = 1,2,3, \dots, N$ dan *M* adalah banyaknya variabel.

c. Memilih secara acak objek pada masing-masing *cluster* sebagai kandidat *medoid* baru.
d. Menghitung jarak dengan persamaan (1) setiap objek yang berada pada masing-masing *cluster* dengan kandidat *medoid* baru.
e. Menghitung total simpangan (s) dengan menghitung nilai :

s = total *distance* baru – total *distance* lama

Jika s $\geq$ 0 maka proses dihentikan, tetapi sebaliknya jika s < 0, maka tukar objek dengan data *cluster* untuk membentuk sekumpulan *k* objek baru sebagai *medoid*.

Mengulangi langkah 3 sampai 5 hingga tidak terjadi perubahan *medoid*, sehingga didapatkan *cluster* beserta anggota *cluster* masing-masing.

2.6 Algoritma *Least Angle Regression*

Algoritma *Least Angle Regression* (LAR) merupakan suatu metode algoritma untuk mendapatkan hasil estimasi koefisien serta dapat melakukan seleksi variabel yang berkaitan dengan pemilihan model terbaik yang dikembangkan oleh Efron, dkk (2004). Dalam pengestimasiannya, algoritma LAR menggunakan program komputasi dengan bantuan *software.*

Adapun langkah-langkah pada algoritma LAR (Hastie, dkk (2004)) :

a. Membuat semua koefisien parameter sama dengan nol $\beta_1, \beta_2, \dots, \beta_k = 0$.
b. Mencari variabel prediktor $\boldsymbol{x}$ yang paling berkorelasi terhadap y.
c. Menduga koefisien parameter β yang memiliki korelasi tertinggi dengan y.
d. Menghitung korelasi antara variabel prediktor yang tersisa dengan sisaan terbaru.
e. Mengulang langkah sampai seluruh p variabel prediktor masuk ke dalam model atau hingga proses iterasi berakhir.
f. Mengeluarkan variabel dari gugus variabel aktif jika terdapat koefisen yang bernilai nol.

Proses iterasi berakhir ketika seluruh variabel telah masuk ke dalam model sehingga didapatkan hasil optimal dari metode algoritma LAR [8].

2.7 *Principal Component Analysis*

Principal Component Analysis (PCA) atau analisis komponen utama merupakan salah satu analisis multivariat yang digunakan untuk mereduksi dimensi data dari yang berukuran besar dan saling berkorelasi menjadi dimensi yang lebih kecil dan tidak saling berkorelasi. *Principal Component* (PC) merupakan suatu kombinasi linear dari variabel-variabel asal. Pembentukan PC berdasarkan dua cara yaitu matriks kovarians atau matriks korelasi [11].

Principal Component Analysis (PCA) atau analisis komponen utama merupakan salah satu analisis multivariat yang digunakan untuk mereduksi dimensi data dari yang berukuran besar dan saling berkorelasi menjadi dimensi yang lebih kecil dan tidak saling berkorelasi. *Principal Component* (PC) merupakan suatu kombinasi linear dari variabel-

variabel asal. Pembentukan PC berdasarkan dua cara yaitu matriks kovarians atau matriks korelasi [11].

Tahapan menentukan PC berdasarkan matriks korelasi sebagai berikut:

1. Membuat matriks varian kovarians $\mathbf{\Sigma}$.
2. Pereduksian PC dimulai dengan mencari nilai eigen $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_M$ yang diperoleh dari persamaan:

$$|\mathbf{\Sigma} - \lambda \mathbf{I}| = 0 \tag{5}$$

Nilai eigen selalu diurutkan dari yang terbesar sampai terkecil. Nilai eigen menunjukkan besarnya total varian yang dijelaskan oleh PC yang terbentuk. Pasangan nilai eigen dan vektor eigen yang saling ortonormal adalah $(\lambda_1, \gamma_1), (\lambda_2, \gamma_2), \ldots, (\lambda_M, \gamma_M)$ dengan $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_M \geq 0$ maka PC ke-*p* didefinisikan sebagai berikut:

$$PC_p = \gamma_{1p} x_1 + \gamma_{2p} x_2 + \cdots + \gamma_{Mp} x_M, \qquad p = 1, 2, \cdots, M \tag{6}$$

Nilai *eigen* dapat menjelaskan besarnya kontribusi keragaman masing-masing komponen utama dalam menjelaskan keragaman data asal. Apabila komponen utama, tahap selanjutnya adalah menghitung skor komponen utama dari setiap objek yang akan digunakan untuk dianalisis lebih lanjut.

Menurut Johnson dan Wichern (2007) [9], banyak komponen utama menjelaskan keragaman data dengan baik dilihat dari proporsi keragaman komponen utama. Banyak komponen utama dikatakan sangat baik apabila proporsi keragaman sebesar 80%, dihitung menggunakan persamaan:

$$Total\ Varian = \frac{\lambda_p}{\sum_{p=1}^{M} \lambda_p} \times 100\% \tag{7}$$

2.8 Penentuan Jumlah Cluster Optimum

Penentuan jumlah *cluster* optimum dilakukan dengan pendekatan indeks validitas *silhouette*. Indeks validitas *silhouette* merupakan suatu ukuran statistik yang digunakan untuk menyeleksi permasalahan penentuan jumlah *cluster* optimal yang dapat merepresentasikan grafis singkat seberapa baik setiap objek terletak dalam *cluster*.

Asumsikan data sudah dikelompokan ke dalam *cluster*. Untuk setiap objek *i*, misalkan $a(i)$ adalah rata-rata jarak objek *i* ke semua objek dalam *cluster* yang sama dan $b(i)$ adalah rata-rata jarak minimum objek *i* ke semua objek pada suatu *cluster* serta *i* bukan anggota *cluster*. Dari penjelasan yang telah dipaparkan indeks validitas *silhouette* dapat ditulis dengan persamaan sebagai berikut:

$$S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{8}$$

Rata-rata $S(i)$ dari seluruh objek dalam suatu *cluster* menunjukkan seberapa dekat kemiripan objek dalam suatu *cluster* yang juga menunjukkan seberapa tepat objek telah dikelompokan. Semakin dekat $S(i)$ kepada 1, maka semakin baik pengelompokan objek.

Sebaliknya, semakin dekat $S(i)$ kepada -1, maka semakin buruk pengelompokan objek. Jumlah *cluster* K yang optimal merupakan estimasi dari harga K yang paling memaksimalkan nilai rata-rata $S(i)$ dan jika terdapat satu *cluster* yang anggotanya terdiri dari satu objek maka nilai rata-rata $S(i)$ adalah 0 [9].

2.9 Penentuan Metode Terbaik dengan Simpangan Baku

Untuk mengetahui metode mana yang mempunyai kinerja terbaik, dapat digunakan rata-rata simpangan baku dalam *cluster* (S_W) dan simpangan baku antar *cluster* (S_B) [10]. Rumus rata-rata simpangan baku dalam *cluster* (σ_W):

$$S_w = \frac{1}{K}\sum_{k=1}^{K} S_k \tag{9}$$

dengan:

K = Banyaknya *cluster* yang terbentuk

S_k = Simpangan baku *cluster* ke-*k*

Rumus simpangan baku *cluster* ke-*k* (S_k):

$$S_k = \sqrt{\frac{1}{N_k - 1}\sum_{i=1}^{N_k}(x_i - \bar{x}_k)^2} \tag{10}$$

dengan,

N = Jumlah anggota dari setiap *cluster*

$\bar{x}_k$ = Rata-rata *cluster* ke-*k*

x_i = Anggota *cluster*, dari *i* = 1, 2, …, N_k

Rumus simpangan baku antar *cluster* (S_B) :

$$S_B = \left[\frac{1}{K - 1}\sum_{i=1}^{K}(\bar{x}_k - \bar{x})^2\right]^{\frac{1}{2}} \tag{11}$$

dengan,

$\bar{x}_k$ = Rata-rata *cluster* ke-*k*

$\bar{x}$ = Rataan keseluruhan *cluster*

Rumus Rasio simpangan baku (S):

$$S = \frac{S_W}{S_B} \times 100\% \tag{12}$$

dengan,

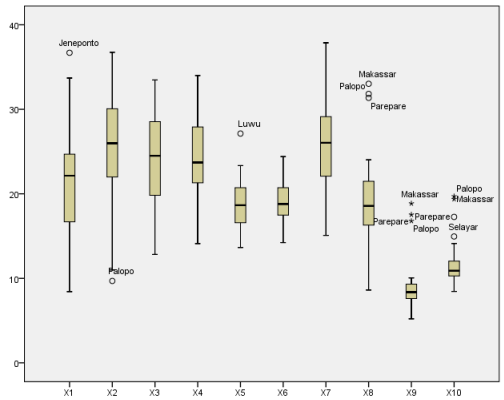
S_W = Simpangan baku dalam *cluster*

S_B = Simpangan baku antar *cluster*

Metode yang mempunyai rasio terkecil merupakan metode terbaik. *Cluster* yang baik adalah *cluster* yang mempunyai homogenitas (kesamaan) yang tinggi antar anggota dalam satu *cluster* (*within cluster*) dan heterogenitas yang tinggi antar *cluster* yang satu dengan *cluster* yang lain (*between cluster*) [11].

3. Hasil dan Diskusi

Sebelum dilakukan analisis *cluster* terlebih dahulu dianalisis secara deskriptif. Berikut adalah statistik deskriptif data kriminalitas di Indonesia yang disajikan pada Grafik 1.



Gambar 1. Boxplot Statistik Deskriptif

Berdasarkan Grafiik 1 dapat diketahui persentase tertinggi, terendah dan outlier pada data indikator pendidikan. Selanjutnya dilakukan uji multikolinearitas dengan menentukan nilai korelasi person, dengan menggunakan program R Studio 3.5.0 diperoleh hasil bahwa terdapat korelasi antara variabel dan salah satunya terjadi korelasi yang kuat adalah antara variabel prediktor x_1 dengan variabel prediktor x_2 ($r_{1,2} = 0.92$). Nilai yang mencapai atau melebihi 0,8 menunjukkan terjadinya multikolinearitas pada variabel prediktor. Metode yang digunakan untuk mengatasi multikolinearitas adalah penggunaan *Principal Component Analysis.*

Setelah dilakukan perhitungan nilai *eigen* dengan persamaan (6), langkah selanjutnya yaitu menghitung proporsi kumulatif dan proporsi variansi kumulatif dengan persamaan (7) yang nilainya dapat dilihat pada tabel 1.

Table 1. Proporsi varians dan proporsi varians kumulatif

PC	Nilai Eigen	Total Varian (%)	Total Kumulatif Varian (%)
PC_1	1.798044e+02	67.58	67.58
PC_2	5.065868e+01	19.04	86.62
PC_3	1.598370e+01	6.008	92.629
PC_4	1.165413e+01	4.38	97.01
PC_5	2.831431e+00	1.064	98.074
PC_6	2.307659e+00	0.008	98.941
PC_7	1.910927e+00	0.007	99.660
PC_8	8.360732e-01	0.003	99.974
PC_9	6.342069e-02	0	99.998
PC_{10}	6.199944e-03	0	100

Sumber: Data diolah, 2020

Pada tabel 1 terlihat bahwa Komponen 1 dan Komponen 2 secara bersama-sama telah dapat menjelaskan keragaman atau varian dari kesepuluh variabel sebesar 86.62%. Dengan kata lain dari dua komponen tersebut sudah mewakili kesepuluh variabel data indikator pendidikan di provinsi Sulawesi Selatan tahun 2018 sebesar 86.62%. Berikut persamaan yang terbentuk dari Komponen 1 dan Komponen 2 berdasarkan tabel 1 :

$$\begin{aligned} PC_1 = {} & -0.41645513x_1 - 0.43480860x_2 - 0.31333627x_3 - 0.29303324x_4 \\ & + 0.09240209x_5 + 0.14866757x_6 + 0.42984238x_7 \\ & + 0.40401532x_8 + 0.20619357x_9 + 0.17974891x_{10} \\ PC_2 = {} & 0.5110681x_1 + 0.4037483x_2 - 0.581168x_3 - 0.4532744x_4 \\ & + 0.00004027808x_5 - 0.04561602x_6 + 0.04865273x_7 \\ & + 0.1591300x_8 + 0.01907573x_9 - 0.04948354x_{10} \end{aligned}$$

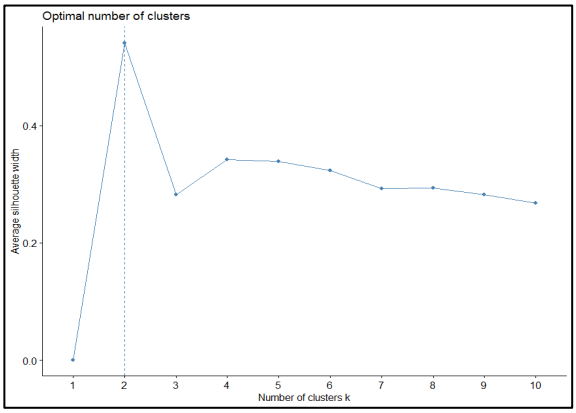
selanjutnya adalah menghitung nilai komponen utama dari kedua persaman PC. Nilai *Principal Component* inilah yang digunakan untuk analisis lebih lanjut.

Berdasarkan persamaan (1) dilakukan perhitungan untuk mencari kemiripan antar observasi. Perhitungan kemiripan antara Kabupaten Kepulauan Selayar dan Kabupaten Bulukumba (Objek 1 dan 2):

$$\begin{aligned} d_{(1,2)} &= \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2} \\ &= \sqrt{-9.5745621 - (-4.4008236))^2 + (2.3347165 - (-0.7776142))^2} \\ &= 6.037729081 \end{aligned}$$

Hasil perhitungan jarak dari Kabupaten Kepulauan Selayar dan Kabupaten Bulukumba diperoleh sebesar 6.037729081, dengan menggunakan persamaan yang sama untuk menentukan jarak antara pasangan observasi lainnya yang nantinya membentuk matriks jarak antara pasangan observasi.

Sebelum dilakukan peng*cluster*an maka terlebih dahulu di tentukan jumlah *cluster* terbaik, terdapat beberapa cara untuk menentukan jumlah klaster terbaik, namun pada penelitian ini menggunakan pendekatan Indeks Validitas *Silhouette*.

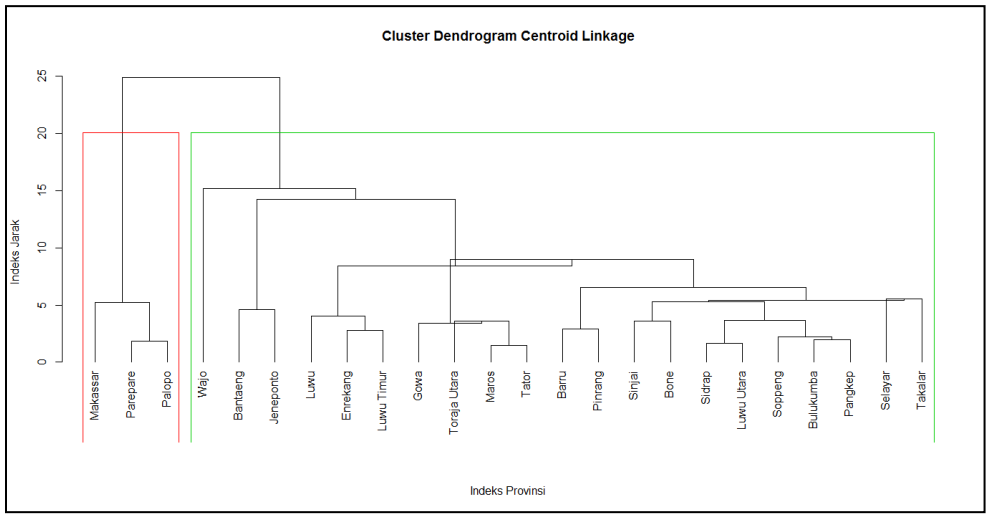


Gambar 2. Plot *Silhouette*

Berdasarkan grafik 2 plot *silhouette* menunjukkan bahwa koefisien *silhouette* tertinggi ketika k = 2, menunjukkan bahwa itu adalah jumlah *cluster* yang optimal. *Number of clusters* menunjukkan pendekatan untuk 1 sampai 10 *cluster*. *Average Silhouette Width* menunjukkan nilai rata-rata *silhouette coefficient* sebesar 0.59. Hal tersebut menunjukkan struktur data yang kuat dan karena nilai rata-rata *silhouette coefficient*nya mendekati angka 1. Selanjutnya proses pengelompokan dengan algoritma *centroid linkage* dan *k-medoid* akan dikelompokan masing-masing menjadi 2 *cluster*.

Proses peng*cluster*an algoritma *centroid linkage* dapat dilakukan menggunakan program *R-Studio* yang terdapat pada lampiran 11**.** Pada tahap pertama, terlihat bahwa observasi nomor 23 (Kota Parepare) dan nomor 24 (Kota Palopo) bergabung menjadi satu *cluster*, karena memiliki jarak paling kecil yaitu 1.32. Hal ini menunjukkan bahwa jarak antara kedua kabupaten tersebut merupakan jarak yang paling dekat dari banyaknya kombinasi jarak 24 kabupaten/kota.

Pada tahap kedua yaitu terlihat bahwa jarak yang terdekat adalah observasi nomor 8 (Kab. Maros) bergabung dengan nomor 18 (Kab. Tana Toraja) dengan jarak yaitu 1.66. Jarak tersebut merupakan jarak rata-rata observasi dari matriks jarak yang baru. Selanjutnya pada tahap ketiga yaitu terlihat bahwa jarak yang terdekat adalah observasi nomor 2 (Kab. Bulukumba) bergabung dengan nomor 12 (Kab. Soppeng) dengan jarak yaitu 1.94. Demikian seterusnya, sehingga semua observasi bergabung menjadi satu *cluster*. Dendogram dengan Algoritma *Centroid Linkage* dapat dilihat pada grafik 4.3 berikut:

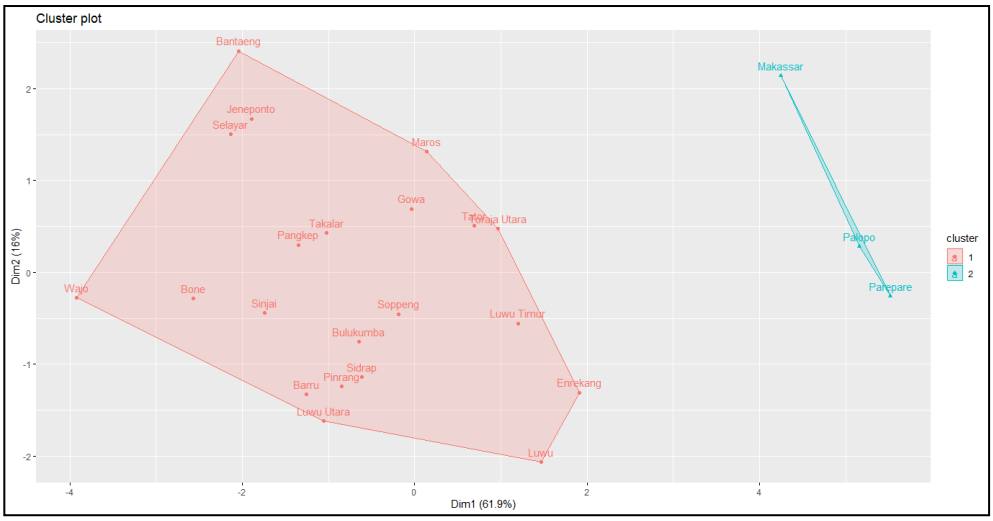


Gambar 3. Dendogram Analisis *Cluster* Hirarki Algoritma *Centroid Linkage*

Berdasarkan grafik 3 Indeks kabupaten/kota menunjuk kepada 24 kabupaten/kota, sedangkan Indeks Jarak menunjukkan ukuran kemiripan antara kabupaten/kota. Dari grafik 3 terlihat bahwa terdapat dua *cluster* yang terbentuk dengan Kabupaten Wajo sampai dengan Kabupaten Takalar masuk *cluster* 1 dan Kota Makassar sampai dengan Kota Palopo masuk *cluster* 2.

Hasil pengelompokan 24 objek dengan algoritma *k-medoids* menggunakan program *R Studio* adalah objek ke-1 sampai dengan objek ke-21 masuk *cluster* 1 dan objek ke-22

sampai dengan objek ke-24 masuk *cluster* 2. Plot algoritma *k-medoids* dapat dilihat grafik 4 berikut ini:



Gambar 4. Plot Analisis *Cluster* Non Hirarki Algoritma *K-Medoids*

Berdasarkan grafik 4, plot pencar dari titik data diwarnai oleh nomor *cluster*. Dalam hal ini, dua komponen utama yang digunakan untuk memplot data. Grafik 3 dan 4 menunjukkan bahwa terdapat 2 *cluster* yaitu *Cluster* 1 terdiri 21 Kabupaten/Kota, *cluster* 2 terdiri 3 Kabupaten/Kota. Dalam hal ini dapat dilihat dalam tabel 2 sebagai berikut:

Table 2. *Cluster* beserta anggotanya dengan Algoritma *Centroid Linkage* dan *K-Medoids*

Cluster	**Kabupaten/ Kota**
Cluster 1	Kepulauan selayar, Bulukumba, Bantaeng, Jeneponto, Takalar, Gowa, Sinjai, Maros, Pangkajene kepulauan, Barru, Bone, Soppeng, Wajo, Sidenreng Rappang, Pinrang, Enrekang, Luwu, Tanah Toraja, Luwu Utara, Luwu Timur, Toraja Utara.
Cluster 2	Makassar, Pare-pare, Palopo.

Sumber: Data diolah, 2020

Selanjutnya menentukan simpangan baku dalam *cluster* dan antar *cluster* pada algoritma *centroid linkage* dan *k-medoids.*

a. Simpangan Baku dalam *cluster* (S_W)

Jika hanya terdapat 1 objek dalam *cluster*, maka S tidak berpengaruh (bernilai 0).

$\bar{x}_k$ = Rata-rata *cluster* ke-*k*

x_i = Anggota *cluster*, dari *i* = 1, 2, …, N_k

Simpangan baku *cluster* 1, untuk rata-rata variabel pada setiap kabupaten/kota dapat dilihat pada lampiran 10, dengan nilai $\bar{x}_I$ = 200.09.

$$S_I = \sqrt{\frac{(x_1 - \bar{x}_I)^2 + (x_2 - \bar{x}_I)^2 + (x_3 - \bar{x}_I)^2 + \cdots + (x_{21} - \bar{x}_I)^2}{N_I - 1}} = 0.702723$$

Simpangan baku *cluster* 2, untuk rata-rata variabel pada setiap kabupaten/kota dapat dilihat pada lampiran 10, dengan nilai $\bar{x}_{II}$ = 199.99.

$$S_{II} = \sqrt{\frac{(x_1 - \bar{x}_{II})^2 + (x_2 - \bar{x}_{II})^2 + (x_3 - \bar{x}_{II})^2}{N_{II} - 1}} = 14.14202$$

Jadi, nilai simpangan baku dalam *cluster* dengan menggunakan algoritma *centroid linkage* dan *k-medoids* :

$$S_W = 2^{-1}(S_I + S_{II})$$

$$S_W = \frac{(S_I + S_{II})}{2}$$

$$S_W = \frac{0.702723 + 14.14202}{2}$$

$$S_W = 7.422371$$

b. Simpangan Baku antar *cluster* (S_B)

Dapat dilihat pada lampiran 10, bahwa nilai $\bar{x}_I$ = 200.09 dan $\bar{x}_{II}$ = 199.99.

$$\bar{x} = \frac{\bar{x}_I + \bar{x}_{II}}{2}$$

$$\bar{x} = \frac{200.09 + 199.99}{2}$$

$$\bar{x} = 200.04$$

$$S_B = \left[(2-1)^{-1} \sum_{i=1}^{K} (\bar{x}_k - \bar{x})^2\right]^{\frac{1}{2}}$$

$$S_B = \left(\frac{(200.09 - 200.04)^2 + (199.99 - 200.04)^2}{2-1}\right)^{\frac{1}{2}}$$

$$S_B = (0.005)^{\frac{1}{2}} = 0.070711$$

Jadi, nilai simpangan baku antar *cluster* dengan algoritma *centroid linkage* dan *k-medoids* adalah 0.070711.

c. Rasio Simpangan Baku dalam *cluster* (S_W) dan antar *cluster* (S_B)

$$S = \frac{S_W}{S_B} \times 100\%$$

$$S = \frac{7.422371}{0.070711} \times 100\%$$

$$S = 104.967 \times 100\%$$

$$S = 104.967$$

Jadi nilai rasio simpangan baku dalam dan antar *cluster* dengan menggunakan algoritma *centroid linkage* dan *k-medoids* adalah 104.967.

Metode terbaik yaitu metode yang mempunyai nilai rasio simpangan baku terkecil. Pada kasus ini, kedua algoritma memiliki hasil yang sama dalam pengelompokan kabupaten/kota di Sulawesi Selatan berdasarkan data indikator pendidikan. Hal ini

dikarenakan pada proses pengelompokan sebelumnya, diperoleh pengelompokan dengan anggota yang sama pada setiap *cluster*.

4. Kesimpulan

Ada 2 kelompok kabupaten/kota yang terbentuk menggunakan algoritma *centroid linkage* dan *k-medoids* berdasarkan data indikator Pendidikan yaitu, *cluster* 1 terdiri dari 21 kabupaten/kota diantara lain, Kepulauan selayar, Bulukumba, Bantaeng, Jeneponto, Takalar, Gowa, Sinjai, Maros, Pangkajene Kepulauan, Barru, Bone, Soppeng, Wajo, Sidenreng Rappang, Pinrang, Enrekang, Luwu, Tanah Toraja, Luwu Utara, Luwu Timur, Toraja Utara. *Cluster* 2 terdiri dari tiga kabupaten/kota yaitu Makassar, Pare-pare, Palopo.

Jika ditinjau dari nilai rasio simpangan S_W terhadap S_B, menunjukkan nilai rasio simpangan baku (S) yang sama pada algoritma *centroid linkage* dan *k-medoids* yaitu 104.967. Hal ini dikarenakan pada pada proses pengelompokan kabupaten/kota di Sulawesi Selatan berdasarkan data indikator pendidikan tahun 2018, diperoleh pengelompokan dengan anggota yang sama pada setiap *cluster*.

Daftar Pustaka

[1] Badan Pusat Statistik. *Indikator Kesejahteraan Rakyat Provinsi Sulawesi Selatan 2019.* Sulawesi Selatan. 2019.

[2] Rachmatin, D. Aplikasi Metode-metode Agglomerative dalam Analisis Cluster pada Data Tingkat Polusi Udara. *Infinity Jurnal Ilmiah Program Studi Matematika STKIP Siliwangi,* 3(2), 2014.

[3] Setiyawati, A. W. *Implementasi Algoritma Partitioning Aroud Medoids (PAM) untuk Pengelompokan Sekolah Menengah Atas di DIY berdasarkan Nilai Daya Serap Ujian Nasional.* Yogyakarta: Universitas Sanata Dharma. 2017.

[4] Gudono, P. *Analisis Data Multivariat Edisi Pertama.* Yogyakarta: BPFE-YOGYAKARTA. 2011.

[5] Mongi, C. E. Penggunaan Analisis Two Step Clustering untuk Data Campuran. *JdC,* 4(1)*,* 2015.

[6] Karlita, T. *Algoritma Perbaikan Penentuan Titik Pusat Awal Berbasis Hirarki untuk Klasterisasi Data Katerogikal.* Surabaya: Institut Teknologi Sepuluh November. 2006.

[7] Vercillis, Carlo. *Business Intelligence: Data Mining and Optimization for Decision Making.* Milan: WILEY. 2009.

[8] Dini Marlina, N. F. Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak. *Jurnal CoreIT,* 4(2), 2018.

[9] Johnson, R.A., Wichern, D. W. *Applied Multivariate Statistical Analysis 6th Edition.* New Jersey: Prentice Hall. 2007.

[10] Nicolaus, E. S. Penentua Jumlah Cluster Optimal pada Median Linkage dengan Indeks Validitas Silhouette. *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster),* 5(2), 2016.

[11] Laeli, S. *Analisis Cluster dengan Average Linkage Method dan Ward's Method untuk Data Responden Nasabah Asuransi Jiwa Unit Link* . Yogyakarta: Universitas Negeri Yogyakarta. 2014.