Vol. 4, No. 1, June 2025, pp. 17-26
e-ISSN: 2987-906X
https://ejournal.unib.ac.id/diophantine

Analisis Perbandingan Metode *Hirearchical*, *K-Means* dan *K-Medoids Clustering,* dalam Pengelompokan Kasus Penyakit Menular di Bengkulu Tengah

Desi T. Tasti[1], Fridz M. Gumay[2] , Ulfianida Aysha[1], Winalia Agwil[1] , dan Wingke Y. Pratami[1]*

[1] Program Studi Statistika, Universitas Bengkulu, Indonesia
[2] Badan Pusat Statistik Kabupaten Bangkulu Tengah, Indonesia

* Corresponding Author Email: wingkeyolapratami@gmail.com

Article Information

Article History:
Submitted: 26 December 2024
Accepted: 22 April 2025
Published: 30 June 2025

Key Words:
Cluster
Infectious Diseases
Hirearchical
K-Means
K-Medoids

DOI:
https://doi.org/10.33369/diophantine.v4i1.32048

Abstract

In *Indonesia, infectious diseases are still a long-standing problem. Experts have accumulated knowledge regarding the emergence of this disease. In the last ten years, Indonesia is still experiencing the problem of a triple burden of disease. While Indonesia is still hit by infectious diseases, non-communicable diseases (NCDs) and diseases that should be overcome, apart from that, infectious diseases are also still a big problem that must be faced. Researchers are interested in conducting research on infectious diseases in Bengkulu Tengah Regency. When analyzing infectious diseases, grouping can be done. Where, the sub-districts in Bengkulu Tengah will provide information based on the level of infectious disease. Grouping is carried out so that decision making is better and more focused. Information obtained from clusters can be used to plan strategies and policies that suit the characteristics of each group. Cluster analysis is an approach to looking for similarities in data and placing similar data into groups. There are two grouping methods in cluster analysis, namely hierarchical methods and non-hierarchical methods. One cluster analysis that uses a hierarchical method is the average linkage method, while the non-hierarchical ones are K-Means and K-Medoids. The variables used in this research are TBC and DHF in 2022. The highest rates of TBC and DHF occurred in Pondok Kelapa District, namely 29 and 23 cases. Based on the results of the analysis, it consists of 2 clusters, with cluster 1 consisting of 9 sub-districts, while cluster 2 consists of 2 sub-districts. Based on the results of evaluating the best method using the Calinski-Harabasz Index, it was found that the K-medoids method was the best method with a value of* 0.

1. PENDAHULUAN

Di Indonesia penyakit menular masih menjadi masalah yang berkepenjangan, sedangkan penyakit tidak menular juga menjadi ancamaan dikarenakan gaya hidup tidak sehat serta penyakit-penyakit degeneratif. Bukan hanya karena kecenderungan gaya hidup, ada beberapa hal yang juga mempengaruhi tingkat kesehatan masyarakat diantaranya adalah jumlah penduduk yang semakin meningkat yang mengakibatkaan keterbatasannya akan luas lahan khususnya untuk permukiman, kurangnya ketersedian air bersih, menurunnya kualitas air akibat pencemaraan limbah, meningkatnya pencemaran udara akibat banyaknya trasportasi kota dan semakin banyak pabrik yang beroperasi [10]. . Indonesia dalam sepuluh tahun terakhir masih mengalami masalah *triple burden diseases*. Dimana Indonesia masih dilanda penyakit infeksi, penyakit tidak menular (PTM) dan penyakit yang seharusnya sudah teratasi selain itu penyakit menular juga masih menjadi masalah besar yang harus di hadapi. Penyakit menular dapat timbul diakibatkan beroperasinya berbagai faktor yang mempengaruhi seperti agen, induk semang atau lingkungan. Didalam usaha para ahli untuk mengumpulkan pengetahuan mengenai timbulnya penyakit, mereka telah melakukan eksperimen terkendali untuk menguji sampai dimana penyakit itu bisa di cegah sehinga dapat meningkatkan taraf hidup penderita [3].

Analisis *cluster* adalah sebuah pendekatan untuk mencari kesamaan dalam data dan menempatkan data yang sama ke dalam kelompok-kelompok. Analisis *cluster* membagi sekumpulan data ke dalam beberapa kelompok dimana kesamaan dalam sebuah kelompok lebih besar daripada diantara kelompok-kelompok (Afira dan Arie, 2021). Metode pengklasteran dalam analisis *cluster* ada dua, metode hierarki dan metode non-hierarki. Analisis *cluster* dengan metode hierarki adalah analisis yang pengklasteran datanya dilakukan dengan cara mengukur jarak kedekatan pada setiap objek yang kemudian membentuk sebuah dengdogram. Jenis analisis *cluster* dengan metode hierarki ada beberapa macam, diantaranya yaitu metode single linkage, metode complete linkage, metode *average linkage*, metode centroid, metode ward, dan metode median *clustering* [7].

Pada penelitian ini, digunakan dataset jumlah kasus penyakit menular tahun 2022 di Bengkulu Tengah untuk dapat dilakukan proses data mining dengan salah satu metode Unsupervised Learning, yaitu *clustering* yang mana akan digunakan analisis kelompok dengan metode *K-means*, Hirearchical dan *K-medoids* [6].

2. METODE

2.1 Tahapan Analisis

Jenis data yang digunakan merupakan data yang bersumber dari Badan Pusat Statistik Kabupaten Bengkulu Tengah. Data tersebut merupakan data TBC dan DBD tahun 2022. Berikut langkah-langkah analisis data dengan metode *Avrage Linkage*:

a. Memilih ukuran jarak kedekatan menggunakan jarak *Euclidean*
b. Menentukan banyaknya *cluster* (*cluster* optimal)
c. Menginterpretasikan profil *cluster* (*cluster-cluster* yang terbentuk)

Metode yang digunakan dalam penelitian ini yaitu *K-Means*:

a. Melakukan uji multikolinearitas
b. Menentukan banyak *cluster* yang akan dibentuk
c. Menentukan pusat *cluster* secara acak
d. Menghitung jarak dari setiap objek data terhadap masing-masing pusat *cluster* (*centeroid*) menggunakan jarak *Euclidean*
e. Menentukan jarak terendah atau terdekat
f. Kemudian melakukan *cluster*

Metode ketiga yang digunakan dalam peneitian ini yaitu *K-Medoids*:

a. Menentukan jumlah *cluster* atau *k* optimal yang akan dibentuk
b. Menentukan pusat *cluster* awal secara acak dan menentukan perwakilan medoid
c. Menghitung matriks jarak menggunakan *Euclidean distance* antara *medoids* dengan setiap objek data
d. Tentukan nilai jarak terendah atau terkecil diantara nilai jarak yang diperoleh
e. Kemudian hitunglah total *cost* jarak terdekat keseluruhan jarak objek data ke *medoids*
f. Pilih secara acak objek(data) untuk menjadi *k* baru sebagai perwakilan medoid.
g. Gunakan set medoid baru untuk menghitung ulang *cost* jarak.
h. Jika yang baru lebih besar dari pada lama maka algoritma dapat dihentikan.

2.2 Analisis *Cluster*

Analisis *cluster* adalah salah satu teknik multivariat yang bertujuan mengklasifikasi suatu objek-objek ke dalam suatu kelompok-kelompok yang berbeda antara lain antara kelompok satu dengan lainnya. Objek-objek yang telah memiliki kedekatan jarak relatif sama dengan objek lainnya [7]. Analisis *cluster* merupakan suatu analisis statistik yang bertujuan untuk menggabungkan objek atau variabel ke dalam kelompok yang mempunyai sifat berbeda antara kelompok satu dengan kelompok yang lainnya (Johnson & Wichern, 1992). Analisis *cluster* mengelompokkan elemen yang mempunyai kesamaan karakteristik diantara objek-objek tersebut, sehingga keragam di dalam suatu kelompok tersebut lebih kecil dibandingkan keragaman antar kelompok. Objek dapat berupa barang, jasa, tumbuhan, binatang dan orang (responden,konsumen atau yang

lainnya). Objek tersebut akan diklasifikasi kedalam satu atau lebih *cluster* (kelompok) akan mempunyai satu kemiripan atau kesamaan karakter. Prosedur *cluster* atau pengelompokan data dapat dilakukan dengan dua metode yaitu metode hierarki dan metode *non* hierarki. Pada penelitian ini menggunakan metode hierarki [8].

2.3 Multikolinearitas

Uji multikolinearitas dilakukan untuk menguji ada atau tidaknya variabel independen yang mempunyai kemiripan antar variabel independen lain Jika data menunjukkan adanya multikolinearitas, maka salah satu cara yang dapat dilakukan yaitu melakukan transformasi data ke dalam bentuk logaritma natural. Cara untuk menguji adanya multikolinieritas, yaitu [9]:

a. Melihat nilai *tolerance* terjadi multikolinieritas, jika nilai *tolerance* ≤ 0.1.
b. Melihat nilai VIF (*Variance Inflation Factor*) terjadi multikolinieritas, jika nilai VIF ≥ 10.
c. Melihat matriks korelasi antara semua variabel independen dalam model. Korelasi tinggi menunjukkan adanya multikolinieritas.

Tujuan dari analisis *cluster* adalah mengelompokan objek berdasarkan kemiripan objek tersebut kedalam *cluster* yang sama. Oleh karena itu dibutuhkan beberapa ukuran guna mengetahui kemiripan dan perbedaan dari objek-objek tersebut. Dalam mengukur kesamaan antar objek ada tiga metode yang bisa digunakan yaitu ukuran asosiasi, ukuran korelasi, dan ukuran jarak. Pada penelitian kali ini peneliti menggunakan ukuran jarak [11].

Ukuran jarak yang digunakan dalam penelitian kali ini yaitu ukuran jarak *Euclidean*. Jarak *Euclidean* adalah besaran jarak pada garis lurus yang menghubungkan antar objek, sebagai contoh ukuran ketidaksamaan atau jarak antar objek ke-i dan objek ke-j dapat disimbolkan dengan d_{ij} dan untuk variable ke-k dengan $k = 1, \dots, p$. nilai d_{ij} didapatkan dari perhitungan pada jarak *Euclidean* dengan rumus [11]:

$$d_{ij} = \sqrt{\left(\sum_{k=1}^{p} x_{ik} - x_{jk}\right)^2}$$

Keterangan :

d_{ij} = jarak *Euclidean* dari objek ke-i dan objek ke-j
p = jumlah variabel *cluster*
x_{ik} = nilai objek ke- i pada variabel ke- k
x_{jk} = nilai objek ke- j pada variabel ke- k

2.4 *Hirearchical Clustering*

Metode *Hirearchical* adalah suatu metode analisis *cluster* yang dilakukan secara bertahap dan bertingkat sehingga membentuk tingkatan seperti pada struktur pohon. Metode ini menghasilkan urutan partisi dengan menggabungkan atau membagi *cluster*. Pada setiap urutan tahapan, partisi baru secara optimal digabungkan atau dipisah dari partisi sebelumnya menurut beberapa kriteria kecukupan. Hasil dari metode ini dapat disajikan dalam bentuk dendogram. Dendogram adalah representatif visual dari seluruh tahapan yang menunjukkan bagaimana *cluster* terbentuk. Selain itu juga terdapat nilai koefisien jarak pada setiap tahapan [1].

Tipe dasar dalam metode hierarki bisa aglomeratif atau devisif. Pada pengklasteran aglomeratif, dimulai dengan menempatkan obyek dalam *cluster-cluster* yang berbeda kemudian mengelompokkan obyek secara bertahap ke dalam *cluster-cluster* yang lebih besar, sedangkan pada pengklasteran devisif dimulai dengan menempatkan semua obyek sebagai satu *cluster*. Kemudian secara bertahap obyek-obyek dipisahkan ke dalam *cluster- cluster* yang berbeda, dua *cluster*, tiga *cluster*, dan seterusnya [8].

Ada lima metode hierarki aglomeratif dalam pembentukan *cluster* yaitu:

a. Pautan Tunggal (*Single Linkage*)
b. Pautan Lengkap (*Complete Linkage*)

c. Pautan Rata-rata (*Average linkage*)
d. Metode *Ward* (*Ward's Method*)
e. Metode *Centroid* (pusat)

2.5 *Average Linkage Method*

Dalam penelitian ini, akan digunakan salah satu pembagian metode algomeratif, yaitu metode *average linkage* adalah metode *clustering* dengan prinsip jarak rata-rata antar setiap pasangan objek yang mungkin pada satu *cluster* dengan seluruh objek pada *cluster* yang lain. *Average linkage* menghitung jarak antara dua *cluster* yang disebut sebagai jarak rata-rata dimana jarak tersebut dihitung pada masing-masing *cluster*. Prosedur *average linkage* dimulai dengan mendefinisikan matrik $D = \{d_{ik}\}$ untuk memperoleh objek-objek paling dekat, sebagai contoh U dan V. Kemudian objek ini digabung ke dalam bentuk *cluster* (UV). Selanjutnya jarak antar (UV) dan *cluster* lainnya, (W) [8].

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

dimana:

D_{ik} = jarak antar objek I dalam *cluster* (UV) dan objek k dalam *cluster* W
$N_{(UV)}$ = jumlah item pada *cluster* UV
N_W = jumlah item pada *cluster* W

Metode *Average linkage* merupakan proses *clustering Average Distance* atau pada jarak rata-rata antar obyeknya. Untuk metode *Average linkage*, jarak diantara dua *cluster* dapat diasumsikan sebagai jarak rata-rata dari semua anggota dalam satu *cluster* dengan semua anggota dalam *cluster* lainnya [10].

2.6 *K-Means Clustering*

K-means Clustering adalah suatu metode penganalisis data atau metode data mining yang melakukan proses pemodelan tanpa supervisi dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. *K-Means Clustering* merupakan salah satu metode *cluster* analisis *non* hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih *cluster* atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokka dalam satu *cluster* yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokkan kedalam *cluster* yang lain. Metode *k-means clustering* bertujuan untuk meminimalisasikan *objective function* yang diset dalam proses *clustering* dengan cara meminimalkan variasi antar data yang ada di dalam suatu *cluster* dan memaksimalkan variasi dengan data yang ada di *cluster* lainnya juga bertujuan untuk menemukan grup dalam data, dengan jumlah grup yang diwakili oleh variabel K. Variabel K sendiri adalah jumlah *cluster* yang diinginkan dalam membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan supervised learning yang menerima masukan berupa vektor (x_1, y_1) , (x_2, y_2) , ..., (x_i, y_i), di mana x_i merupakan data dari suatu data pelatihan dan y_i merupakan label kelas untuk x_i. Terdapat dua jenis data *clustering* yang sering dipergunakan dalam proses pengelompokan data yaitu *hierarchical* dan *non-hierarchical*, dan *k-means* merupakan salah satu metode data *clustering non-hierarchical* atau *partitional clustering* [2].

2.7 *K-Medoids Clustering*

K-medoids merupakan pengembangan dari metode *k-means clustering*. Jika *k-means* merupakan algoritma yang sensitif terhadap outlier atau jarak data yang terlalu timpang dan noise maka metode *k-medoids* dapat mengatasi hal tersebut. Hal ini dikarenakan *k-medoids* menggunakan nilai median dari data yang tidak terpengaruh walaupun jika ada outlier dan noise dalam data sehingga metode ini lebih robust dibandingkan *k-means* [4].

K-means berusaha meminimumkan nilai total squared error, sedangkan *k-medoids* meminimumkan *sum of dissimilarities* antara data di sebuah *cluster* dan memilih sebuah data di dalam *cluster* sebagai *center* (*medoids*) [5]. Metode *k-medoids* adalah dengan cara mengambil nilai rata-rata objek dalam *cluster* sebagai titik referensi, kita dapat memilih objek aktual untuk mewakili *cluster*, menggunakan satu objek perwakilan per *cluster*. Setiap objek yang tersisa ditugaskan ke *cluster* yang objek perwakilannya paling mirip. Metode partisi kemudian dilakukan berdasarkan prinsip meminimalkan jumlah ketidaksamaan antara setiap objek p dan objek C_i perwakilan yang sesuai. Artinya, kriteria kesalahan mutlak digunakan dan didefenisikan sebagai berikut.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, o_i)^2$$

Dimana E adalah jumlah dari kesalahan absolut untuk semua objek p dalam kumpulan data, dan o_i adalah objek perwakilan dari C_i. Hal ini adalah dasar untuk metode *K-medoids*, yang menegelompokan n objek ke dalam *cluster* k dengan meminimalkan kesalahan absolut.

2.8 Indeks *Calinski-Harabasz*

Indeks validasi *Calinski-Haraabasz* (CH) menghitung perbandingan nilai *sum of square between cluster* (SSB) sebagai *separation* dan nilai *sum of square within cluster* (SSW) sebagai *compactness* yang dikalikan dengan faktor normalisasi, yaitu selisih jumlah data dengan jumlah klaster sibagi dengan jumlah klaster dikurang satu. Jumlah klaster terbaik ditunjukan dengan semakin besar nilai CH (Baarsch, 2012). Misalkan terdapat suatu himpunan data dengan k buah klaster dan N buah titik data, missal C_i adalah klaster ke l dengan x_i adalh titik ke i pada klaster ke l, N_i adalah jumlah titik pada kalster ke l dan $\bar{x}_i$ adalah titik pusat klaster ke l, maka perhitungan indeks validasi CH dapat dilihat pada rumus berikut:

$$CH = \frac{trace(SSB)}{trace(SSW)} \times \frac{N-k}{k-1}$$

$$SSW = \sum_{i=1}^{k} \sum_{x_i \in c_i} (x_i - \bar{x}_i)(x_i - \bar{x}_i)^T$$

$$SSB = \sum_{i=1}^{k} N_i (x_i - \bar{x}_i)(x_i - \bar{x}_i)^T$$

2.9 Evaluasi Hasil Pengelompokan Terbaik

Tingkat keberhasilan usaha ditentukan berdasarkan penilaian kinerja metode tersebut. Penilaian dapat dilakukan dengan membandingkan hasil pengelompokan oleh masing-masing metode dengan menggunakan kriteria dua nilai simpangan baku, yaitu rata-rata baku dalam kelompok (S_w) dan simpangan baku antar kelompok (S_B) (Bunkers, 1996). Rumus rata-rata simpangan baku dalam kelompok [6]:

$$S_w = K^{(-1)} \sum_{k=1}^{K} S_k$$

Keterangan:

K = Banyaknya kelompok yang berbentuk

S_k =simpangan baku kelompok ke k

Rumus rata-rata simpangan antar kelompok

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

Keterangan:

$\bar{X}_k$ = Rataan kelompok k

$\bar{\bar{X}}$ = Rataan keseluruhan kelompok

3. HASIL DAN PEMBAHASAN

3.1 Analisis Deskriptif dan Eksplorasi Dana Analisis

Analisis ini bertujuan untuk mendeskripsikan gambaran secara mendalam dan objektif mengenai jumlah kasus penyakit menular berdasarkan 11 kecamatan di Bengkulu Tengah pada tahun 2022 yang diamati. Kecamatan-kecamatan tersebut terdiri dari Talang Empat, Semidang Lagan, Karang Tinggi, Taba Penanjung, Merigi Kelindang, Pagar Jati, Merigi Sakti, Pondok Kelapa, Pondok Kubang, Pematang Tiga, dan Bang Haji. Berikut rincian dari ringkasan nilai statistik deskriptif yang terdiri dari nilai minimum, nilai maksimum, standar deviasi, varians dan rata-rata terhadap variabel jumlah kasus penyakit menular TBC dan DBD disajikan pada tabel 1.

Tabel 1. Nilai Statistik Deskriptif Dari Data Jumlah Kasus Penyakit Menular Tahun 2022

Variabel	TBC	DBD
Rata-rata	13.180000	6.81800
Standar Deviasi	7.180782	8.44770
Varians	51.563640	71.36364
Nilai Maksimum	29.000000	23.00000
Nilai Minimum	5.000000	0

Berdasarkan tabel tersebut didapatkan nilai minimum, nilai maksimum, standar deviasi, varians dan rata-rata pada variabel TBC dan DBD tahun 2022.

3.2 Uji Multikolinearitas

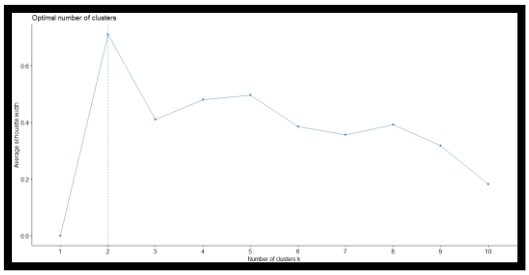
Untuk mengetahui terjadi atau tidaknya multikolinearitas, dapat dilihat dari niai VIF pada masing-masing variabel. Suatu data dikatakan terbebas dari gejala multikolinearitas apabila nilai VIF < 10. Berdasarkan hasil perhitungan uji multikolinearitas terhadap data penyakit menular di Bengkulu Tengah tahun 2022 menggunakan persamaan (2) dan (3) didapatkan hasil yang tertera pada tabel 2.

Tabel 2. Hasil Uji Multikolinearitas Semua Variabel

Variabel	VIF
TBC	2.400825
DBD	2.400825

Berdasarkan tabel 2 diperoleh hasil yaitu variabel TBC memiliki nilai VIF sebesar $2.400825 < 10$, variabel DBD nilai VIF sebesar $2.400825 < 10$. Maka dapat disimpulkan bahwa pada semua variabel tidak terdapat multikolinearitas.

3.3 Menentukan Jumlah *Cluster* (*k* Optimal)



Gambar 1. Grafik Jumlah ***k*** Optimal

Penentuan jumlah *cluster* pada penelitian ini menggunakan metode *silhouette.* Metode ini mengukur seberapa baik setiap individu berada dalam *cluster* yang sama dibandingkan dengan *cluster* lain. Nilai *silhouette* berkisar antara -1 hingga 1. Semakin tinggi nilai *silhouette,* semakin baik *cluster* tersebut.

Berdasarkan gambar diatas, dalam menentukan nilai k optimal dari grafik yang menunjukkan bahwa jumlah *cluster* yang akan terbentuk adalah saat $k = 2$, dikarenakan pada $k = 2$ memiliki nilai *silhouette* yang paling tinggi.

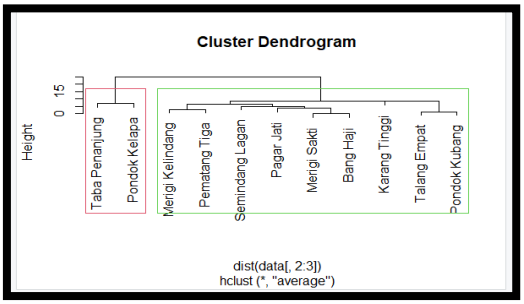
3.4 Matriks Jarak *Euclidean*

Tabel 3. Matriks Jarak *Euclidean*

	1	**2**	⋯	**11**
1	0.00000	3.162278	⋯	5.385165
2	3.162278	0.000000	⋯	2.236068
3	5.000000	5.000000	⋯	5.830952
⋮	⋮	⋮	⋮	⋮
11	5.385165	2.236068	⋯	0.0000

Berdasarkan perhitungan jarak menggunakan *Euclidean*, kecamatan yang memiliki jarak paling kecil sebesar 0.00 yaitu jarak antara *cluster* objek 7 (Kecamataan Merigi Sakti) dengan objek 11 (Kecamatan Bang Haji) yang kemudian kedua kecamatan tersebut digabung menjadi satu *cluster*.

3.5 Pengklasteran Dengan Metode *Average Linkage*



Gambar 2. Dendogram *Average Linkage*

Sebelum malakukan peng-*cluster*-an, dikakukan penentuan banyaknya *cluster* yang digunakan agar terbentuk *cluster* optimal dengan menggunakan metode *silhouette.* Dari metode *silhouette* didapatkan bahwah jumlah *cluster* yang optimal adalah 2 *cluster.*

Proses pengelompokan *cluster* dapat digambarkan dalam bentuk dendogram. Dendogram pengelompokan kecamatan di Bengkulu Tengah menggunakan metode *Average linkage* akan membentuk dua *cluster* yang disajikan pada Gambar 2. Berikut disajikan tabel pengelompokan yang terbentuk berdasarkan dendogram tersebut.

Tabel 4. Hasil *Cluster Average linkage*

Cluster	**Jumlah Kecamatan**	**Kecamatan**
1	9	Talang Empat, Semindang Lagan, Karang Tinggi, Merigi Kelindang, Pagar Jati, Merigi Sakti, Pondok Kubang, Pematang Tiga, Bang Haji
2	2	Taba Penanjung, Pondok Kelapa

Berdasarkan pengklasteran yang membentuk dua *cluster*, maka akan menghasilkan dua kelompok anggota kecamatan. Hasil anggota kecamatan jika akan membentuk dua *cluster* menggunakan metode *Average linkage* dapat dilihat pada tabel 4.

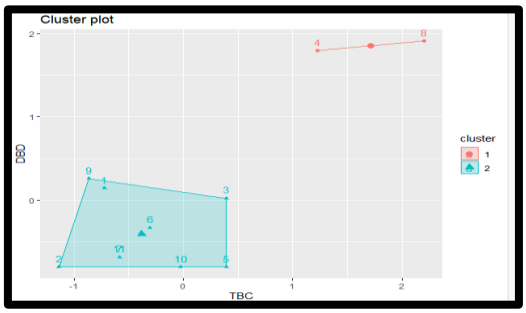
3.6 Pengklasteran Dengan Metode *K-Means*

Berdasarkan hasil analisis, metode *non* hirarki dimana pada gambar dendrogram diperoleh 2 *cluster* yaitu untuk *cluster* 1 terdapat hanya 2 kecamatan yaitu Taba Penanjung dan Pondok Kelapa. Sedangkan untuk *cluster* 2 terdapat 9 kecamatan yaitu Talang Empat, Semidang Lagan, Karang Tinggi, Merigi Kelindang, Pagar Jati, Merigi Skati, Pondok Kubang, Pematang Tiga, dan Bang Haji. Maka, dapat disimpulkan bahwa dari kedua *cluster* ini yaitu jumlah kasus penyakit menular paling banyak adalah berada pada *cluster* 2. Pengelompokan *cluster* ini dilihat berdasarkan kesamaan karakteristik yang dilihat dari jarak dimana langkahnya yaitu

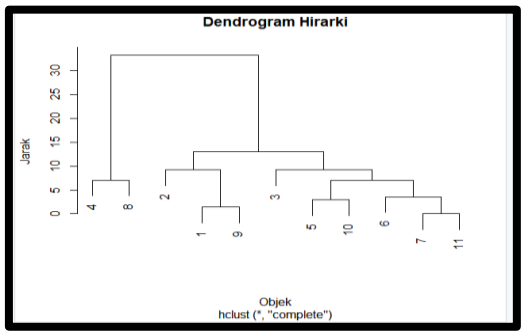
mencari jarak matriks terdekat pada pasangan *cluster*. Kemudian menggabungkan *cluster* U dan V menjadi *cluster* UV. Ulangi sebanyak $N-1$ kali hingga tersisa satu *cluster* yang memuat seluruh objek atau pengamatan, yang dimana N merupakan jumlah total objek atau pengamatan yang akan dikelompokan. Sehingga akhirnya terbentuk dua kelompok *cluster* seperti pada Gambar 3.

Tabel 5. Anggota Hirarki

Kecamatan	TBC	DBD	*Cluster*
1	8	8	2
2	5	0	2
3	16	7	2
4	22	22	1
5	16	0	2
6	11	4	2
7	9	1	2
8	29	23	1
9	7	9	2
10	13	0	2
11	9	1	2



Gambar 3. Plot *Cluster K-means*



Gambar 4. Dendogram *K-means*

3.7 Pengklasteran Dengan Metode *K-Medoids*

Berdasarkan k optimal yang didapatkan, maka dapat dilakukan pengklasteran dengan $k = 2$.

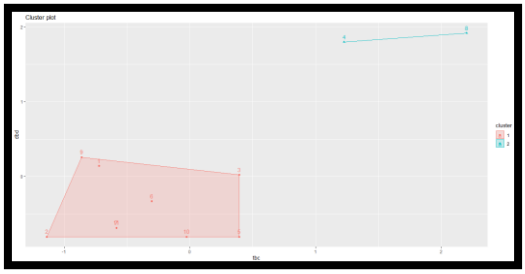
Tabel 6. *Medoids*

	ID	TBC	DBD
[1,]	6	11	4
[2,]	8	29	23

Berdasarkan tabel diatas menunjukkan bahwa data yang dipilih sebagai perwakilan medoid adalah data ke-6 dan data ke-8 dengan kedua data tersebut merupakan kecamatan Pagar Jati dan Pondok Kelapa.

Tabel 7. Hasil *Cluster K-medoids*

No	Kecamatan	*Cluster*
1	Talang Empat	1
2	Semindang Lagan	1
3	Karang Tinggi	1
4	Taba Penanjung	2
5	Merigi Kelindang	1
6	Pagar Jati	1
7	Merigi Sakti	1
8	Pondok Kelapa	2
9	Pondok Kubang	1
10	Pematang Tiga	1
11	Bang Haji	1



Gambar 5. Plot *Cluster K-medoids*

Tabel 8. Nilai Rata-Rata *Cluster*

Cluster	TBC	DBD
1	10.4	3.33
2	25.5	22.5

Berdasarkan hasil pengklasteran diperoleh 2 *cluster* dengan *cluster* 1 terdapat 9 kecamatan yaitu Talang Empat, Semidang Lagan, Karang Tinggi, Merigi Kelindang, Pagar Jati, Merigi Skati, Pondok Kubang, Pematang Tiga, dan Bang Haji. Sedangkan, pada *cluster* 2 hanya terdapat 2 kecamatan yaitu Taba Penanjung dan Pondok Kelapa. Berdasarkan nilai rata-rata pada tabel , dapat disimpulkan bahwa kecamatan yang masuk dalam *cluster* 1 adalah kecamatan dengan jumlah penyakit menular yang rendah. Sedangkan kecamatan yang masuk pada *cluster* 2 adalah kecamatan yang memiliki jumlah penyakit menular tertinggi.

3.8 Penentuan Dan Evaluasi Metode Terbaik

Ada dua pendekatan metode yang dilakukan dalam penelitian ini, yaitu dengan metode *nonhierarchical clustering* yaitu *K-means*, *Hirearchical*, dan *K-medoids*. Untuk penerapan kedua metode tersebut, terlebih dahulu dilakukan validitas untuk jumlah *cluster* yang optimum dari masing-masing metode yang digunakan menggunakan *Calinski-Harabasz Index* yang disajikan pada table berikut.

Tabel 8. Perhitungan nilai *Calinski-Harabasz Index* (CH)

Jumlah Kelompok	*K-Means*	*Hirearchical*	*K-Medoids*
	CH	CH	CH
K=2	34.01127	34.01127	34.01127
K=3	26.89272	26.89272	26.89272
K=4	29.50722	23.86121	28.61958

Pada index validasi *Calinski-Harabasz Index* (CH), untuk menentukan k optimum adalah dengan melihat nilai CH yang paling besar. Berdasarkan tabel diatas, diketahui bahwa k optimum ada pada jumlah kluster k=2 untuk setiap metode yang digunakan.

3.9 Evaluasi Model Dan Metode Terbaik

Untuk evaluasi model dilihat dari nilai *average within* dan *average between cluster. Cluster* terbaik ialah yang memiliki nilai *average within* yang sangat kecil dan nilai *average between* yang sangat besar. Dalam

pemilihan jarak yang menghasilkan kualitas pengelompokkan tarbaik, dipilih jarak yang meminimalkan nilai rasio rata-rata simpangan baku dalam kelompok dan simpangan baku antar kelompok. Untuk membandingkan kedua model, dapat dilihat dari nilai rasio rata-rata simpangan baku paling kecil. Nilai ratio diperoleh dari hasil pembagian rata-rata simpangan baku dalam kelompok/*average within* (SW) dan simpangan baku antar kelompok/ *average between* (SB) yang disajikan pada tabel di bawah.

Tabel 8. Hasil Rasio Simpangan Baku

Metode *Clustering*	**Jumlah Kelompok**	**SW/SB**
Hirearchical	2	0.030930
K-means	2	0.104988
K-medoids	2	0

Berdasarkan tabel di atas, diketahui bahwa nilai ratio yang paling kecil ada pada model *K-Medoids* yaitu 0, sehingga model *clustering* dengan algoritma *K-Medoids* lebih baik dibandingkan *K-Means* dan *Hirearchical*.

4 SIMPULAN

Berdasarkan hasil analisis yang telah dilakukan, diperoleh bahwa metode terbaik dalam pengelompokan jumlah kasus penyakit menular di Kabupaten Bengkulu Tengah tahun 2022 yaitu *K-medoids* dengan 2 *cluster*. Metode *K-medoids* terpilih sebagai metode terbaik setelah dilakukan pengujian atau evaluasi validasi. Berdasarkan hasil evaluasi validasi menggunakan *Calinski-Harabasz Index* diketahui bahwa metode *K-medoids* memiliki hasil evaluasi validasi terkecil yaitu sebesar 0. Terdapat 2 *cluster* yang terbentuk berdasarkan hasil pengelompokan tersebut. *Cluster* 1 merupakan *cluster* dengan kategori jumlah kasus penyakit menular yang rendah terdiri dari 9 kecamatan. *Cluster* 1 terdiri dari kecamatan Talang Empat, Semidang Lagan, Karang Tinggi, Merigi Kelindang, Pagar Jati, Merigi Skati, Pondok Kubang, Pematang Tiga, dan Bang Haji. *Cluster* 2 merupakan *cluster* dengan kategori jumlah kasus penyakit menular yang tinggi terdiri dari 2 kecamatan. *Cluster* 2 terdiri dari kecamatan Taba Penanjung dan Pondok Kelapa.

REFERENSI

[1] N. Afira dan A. W. Wijayanto, "Analisis *Cluster* Kemiskinan Provinsi di Indonesia Tahun 2019 dengan Metode Partitioning dan Hierarki", *Jurnal Sistem Komputer*, Vol. 10, No. 2, 101-109, 2021

[2] Anonim. *K-means Clustering*. Universitas Raharja. 2020

[3] Irwan, *Epidemiologi Penyakit Menular*, Yogyakarta, CV. Absolute Media, 2017

[4] J. Han & M. Kamber, *Data Mining: Concepts and Techniques*, Second Edition, San Francisco, Elsevier Inc, 2006

[5] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, San Francisco, Morgan Kaufmann Publishers, 2012

[6] E. Luthfi dan A. W. Wijayanto, 2021, "Analisis Perbandingan Metode Hirearchical, *K-means*, Dan *K-medoids Clustering* Dalam Pengelompokkan Indeks Pembangunan Manusia Indonesia", *INOVASI: Jurnal Ekonomi, Keuangan dan Manajemen*, Vol 17, No 4, 2021

[7] Q. Nafisah dan N. E. Chandra, "Analisis *Cluster Average linkage* Berdasarkan Faktor-Faktor Kemiskinan di Provinsi Jawa Timur", *Zeta – Math Journal*, Vol 3, No 2, 2017

[8] M. Paramadina, Sudarmin, dan M. K Aidid, 2019, "Perbandingan Analisis *Cluster* Metode *Average linkage* dan Metode Ward (Kasus: IPM Provinsi Sulawesi Selatan)", *Journal of Statistics and Its Application on Teaching and Research*, Vol. 1, No.2, 22-23, 2019

[9] D. N. P. Sari, *Analisis Cluster Dengan Metode K-means Pada Persebaran Kasus Covid-19 Berdasarkan Provinsi Di Indonesia,* Semarang, Universitas Semarang, 2020

[10] N. Ulinnuh dan R. Veriani, "Analisis *Cluster* dalam Pengelompokan Provinsi di Indonesia Berdasarkan Variabel Penyakit Menular Menggunakan Metode Complete Linkage, *Average linkage* dan Ward", *Jurnal Nasional Informatika dan Teknologi Jaringan*, Vol. 5, No.1, 2020

[11] R. Veriani, *Analisis Cluster Dalam Pengelompokan Provinsi Di Indonesia Berdasarkan Variable Penyakit Menular Menggunakan Metode Complete Linkage, Average linkage Dan Ward*, Surabaya, Universitas Islam Negri Sunan Ampel, 2020