e-ISSN: 2580-3239
p-ISSN: 2085-7993
Volume 22 No. 2 November 2023

Simulasi Perhitungan Analisis *Cluster* pada Kasus Penyakit Menular Menggunakan Bahasa Pemrograman Python

Aulia Wanda Puspitasari, Herlina Napitupulu, Nurul Gusriani
Departemen Matematika, Fakultas MIPA, Universitas Padjadjaran
Email: aulia19011@mail.unpad.ac.id; herlina.napitupulu@mail.unpad.ac.id; nurul.gusriani@unpad.ac.id

Abstrak

Peningkatan pencegahan dan pengendalian penyakit menular menjadi salah satu dari tujuan strategis pembangunan kesehatan 2020-2024 yang telah ditetapkan oleh Kementerian Kesehatan. Jika terus dibiarkan, penyakit menular dapat menjadi Kejadian Luar Biasa (KLB) yang banyak memakan kematian. Untuk mengoptimalkan penanganan penularan penyakit menular, perlu ditentukan kelompok yang menjadi prioritas. Salah satu metode pengelompokkan yang dapat digunakan adalah analisis *cluster*. Penelitian ini bertujuan untuk menemukan *cluster* terbaik berdasarkan nilai *Cophenetic Correlation Coefficient* (CPCC) dan *Pseudo-F*. Hasil penelitian menunjukkan bahwa metode *average linkage* merupakan metode terbaik dengan nilai CPCC paling mendekati 1 yaitu sebesar 0,8513. Metode *average linkage* membagi kabupaten/kota di Jawa Barat menjadi lima *cluster* berdasarkan nilai *Pseudo-F* tertinggi.

Kata Kunci: analisis *cluster*, penyakit menular, *Pseudo-F*, *Cophenetic Correlation Coefficient*

Abstract

Increasing the prevention and control of infectious diseases is currently one of the 2020-2024 health development strategic goals set by the Ministry of Health. If it left unchecked, infectious diseases can become Kejadian Luar Biasa (KLB) which result in many deaths. To optimize the handling of infectious disease transmission, it is necessary to determine which groups are the priority. One of the grouping methods that can be used is cluster analysis. This study aims to find the best cluster based on the Cophenetic Correlation Coefficient (CPCC) and Pseudo-F values. The results showed that the average linkage method was the best method with a CPCC value closest to 1, that is 0.8513. The average linkage method divides districts/cities in West Java into five clusters based on the highest Pseudo-F value.

Keywords: *cluster analysis; infectious diseases; Pseudo-F; Cophenetic Correlation Coefficient*

1 PENDAHULUAN

Kementerian Kesehatan telah menetapkan enam tujuan strategis yang dijabarkan menjadi empat belas sasaran strategis dalam mendukung pembangunan kesehatan 2020-2024. Salah satu dari tujuan strategis tersebut adalah peningkatan pencegahan dan pengendalian penyakit dan pengelolaan kedaruratan kesehatan masyarakat. Pengendalian penyakit, khusus penyakit menular, masih merupakan hal yang belum dapat dituntaskan pada sektor kesehatan (Direktorat Jenderal Pencegahan dan Pengendalian Penyakit, 2020). Penyebaran penyakit menular harus menjadi perhatian lebih karena apabila kasus penyakit menular dibiarkan, akan terjadi Kejadian Luar Biasa (KLB). Menurut Peraturan Menteri Kesehatan Republik Indonesia Nomor 1501/MENKES/PER/X/2010 Kejadian Luar Biasa adalah keadaan ketika dalam kurun

e-ISSN: 2580-3239
p-ISSN: 2085-7993
Volume 22 No. 2 November 2023

waktu tertentu, kejadian kesakitan ataupun kematian meningkat yang bermakna secara epidemiologi dan dapat menjurus kepada terjadinya wabah.

Berbagai hal dilakukan oleh pemerintah dalam mencegah penyebaran penyakit menular, mulai dari peningkatan fasilitas kesehatan hingga pemberdayaan masyarakat seperti pengadaan program dalam meningkatkan pengetahuan masyarakat sekitar. Agar program yang dilakukan oleh pemerintah tersebut dapat berjalan secara efektif perlu ditentukan prioritas pengadaan program. Daerah-daerah dengan tingkat kerawanan penyakit menular tinggi menjadi prioritas utama diadakannya program. Salah satu cara penentuan daerah-daerah tersebut dapat dilakukan dengan analisis *cluster*.

Penelitian terdahulu dilakukan oleh Nakayama *et al*. (2012) dengan menggunakan analisis *cluster* untuk mengidentifikasi *clinical phenotype* yang berbeda. Pada penelitian tersebut analisis *cluster agglomerative* hierarki yaitu *Ward's method* dan perhitungan *index Pseudo*-F digunakan untuk menentukan jumlah *cluster*. Penelitian lain telah dilakukan oleh Ulinnuh & Veriani (2020) dengan mengelompokan provinsi di Indonesia menggunakan analisis *cluster Complete Linkage*, *Average Linkage*, dan Metode *Ward* berdasarkan variabel penyakit menular. Penelitian menunjukkan bahwa Metode *Ward* adalah metode yang lebih baik dari pada metode *Complete Linkage* dan *Average Linkage* berdasarkan rasio simpangan baku.

Penelitian mengenai perbandingan metode analisis *cluster* juga telah diteliti oleh Pusdiktasari *et al*. (2021) dalam mengelompokkan provinsi di Indonesia yang perekonomiannya berisiko terdampak Covid-19. Metode analisis *cluster* yang dibandingkan adalah metode *Single Linkage*, *Complete Linkage*, *Average Linkage*, Metode *Centroid*, dan Metode *Ward*. Hasil penelitian menunjukkan bahwa metode *Average Linkage* memiliki nilai *Cophenetic Correlation Coefficient* tertinggi sehingga metode tersebut adalah metode yang terbaik.

Berdasarkan uraian tersebut, penulis menggunakan metode *cluster Agglomerative Hierarchy* yakni *average linkage* dan *centroid method* pada penelitian ini. Penentuan jumlah *cluster* optimal ditentukan dengan menggunakan indeks *Pseudo-F* serta penentuan metode yang terbaik didasarkan nilai *Cophenetic Correlation Coefficient*. Perhitungan yang dilakukan pada penelitian menggunakan bantuan software Ms. Excel dan bahasa pemrograman Python.

2 KAJIAN PUSTAKA

2.1 Pengukuran Jarak

Kemiripan antar objek merupakan dasar dari analisis *cluster*. Terdapat berbagai macam cara untuk mengukur kemiripan, diantaranya adalah pengukuran korelasi dan pengukuran jarak. Dalam analisis *cluster*, tipe pengukuran jarak yang paling umum digunakan untuk mengukur jarak antar objek adalah jarak *Euclidean* (Johnson & Wichern, 2002). Perhitungan jarak *Euclidean* antar objek a dan objek b disimbolkan $d(a,b)$ pada variabel ke-i dengan $i = 1, 2, \dots, p$ adalah sebagai berikut:

$$d(a,b) = \sqrt{\sum_{i=1}^{p} (x_{aix} - x_{bi})^2} \qquad (1)$$

2.2 Analisis *Cluster*

Analisis *cluster* merupakan salah satu cara untuk menemukan kelompok dalam kumpulan data. Metode pengelompokkan dalam analisis *cluster* terbagi menjadi dua yaitu metode hirarki dan non hirarki. Pada pengelompokkan metode hirarki terdapat dua tipe metode yaitu *Agglomerative* dan *Divisive.* Metode hirarki *Agglomerative* menganggap setiap objek sebagai *cluster* tersendiri dan menggabungkan dua *cluster* dengan kesamaan terbanyak hingga semua objek berada pada satu *cluster* yang sama. Algoritma metode *agglomerative hierarchy* diantaranya adalah (Hair *et al*., 2019):

1. *Average Linkage*
   Metode *average linkage* menghitung jarak antar dua *cluster* dengan rata-rata jarak

e-ISSN: 2580-3239
p-ISSN: 2085-7993
Volume 22 No. 2 November 2023

setiap pasangan objek dimana salah satu anggota pasangan adalah milik setiap *cluster*. Untuk menghitung jarak antar *cluster* a dengan *cluster* b yang digunakan persamaan

$$d_{ab} = \frac{\sum_{u=1}^{n_a}\sum_{v=1}^{n_b} d(a_u, b_v)}{n_a n_b} \quad (2)$$

di mana $d(a_u, b_v)$ adalah jarak antar objek u pada *cluster* a dengan objek v pada *cluster* b, n_a merupakan banyaknya anggota *cluster* a dan n_b merupakan banyaknya anggota *cluster* b.

2. Metode *Centroid*
   Pada metode *centroid*, jarak antara dua *cluster* a dan b didefinisikan sebagai jarak *Euclidean* antara vektor rata-rata dari dua *cluster*. Selanjutnya *centroid* dari *cluster* ab yang merupakan penggabungan *cluster* a dan *cluster* b dapat dihitung dengan *weighted sum* berikut

$$\overline{x}_{ab} = \frac{n_a\overline{x}_a + n_b\overline{x}_b}{n_a + n_b}. \quad (3)$$

2.3 Standarisasi Data

Standarisasi data dilakukan jika antar variabel memiliki perbedaan ukuran satuan data yang besar. Standarisasi data dengan mentransformasi data kedalam bentuk *z-score* pada persamaan sebagai berikut (Tri *et al*., 2019)

$$z_{ai} = \frac{x_{ai} - \overline{x_i}}{S_{x_i}} \quad (4)$$

di mana
x_{ai} : data objek a variabel i
$\overline{x_i}$: rata-rata data pada variabel i
S_{x_i} : standar deviasi data variabel ke-i.

2.4 Asumsi Analisis *Cluster*

Analisis *cluster* adalah metode untuk mengukur karakteristik struktural dari suatu data himpunan observasi. Peneliti harus fokus pada tiga asumsi pada analisis *cluster* (Hair *et al*., 2019), yaitu:

1. Terdapat struktur
   Asumsi mendasar dari semua teknik interdependensi pada analisis multivariat adalah bahwa terdapat struktur alami yang harus diidentifikasi dengan teknik ini. Analisis *cluster* mengasumsikan ada beberapa kelompok alami dalam sampel yang akan dianalisis.
2. Sampel representatif
   Sampel yang digunakan pada penelitian harus dipastikan dapat mewakili populasi. Tidak ada ketentuan jumlah sampel yang diperlukan agar data representatif, namun semakin mewakili sampel peneliti, semakin baik analisis *cluster* yang dihasilkan.
3. Multikolinearitas antar variabel
   Variabel-variabel pada analisis *cluster* diasumsikan independen. Salah satu cara untuk mendeteksi adanya multikolinearitas adalah dengan menggunakan *Variance Inflation Factor* (VIF). Jika nilai VIF lebih dari 10, maka terdapat multikolinearitas pada variabel prediktor. Nilai VIF dari variabel j dihitung menggunakan rumus:

$$VIF_j = \frac{1}{1 - R_j^2} \quad (5)$$

   di mana R_j^2 adalah koefisien determinasi dari variabel x_j yang diregresikan terhadap variabel lainnya.

2.5 Penentuan Jumlah Cluster

Pemilihan jumlah *cluster* yang tepat merupakan salah satu hal penting dalam analisis *cluster*. Jumlah *cluster* yang optimal dapat ditentukan dengan menggunakan kriteria nilai *Pseudo-F*. Nilai *Pseudo-F* terbesar menunjukkan bahwa *cluster* mencapai hasil yang optimal (Nakayama *et al*., 2012). Nilai *Pseudo-F* dihitung dengan menggunakan persamaan:

$$Pseudo - F = \frac{\frac{SSB}{k-1}}{\frac{SSW}{n-k}} \quad (6)$$

dengan

$$SSB = SST - SSW \quad (7)$$

$$SST = \sum_{j=1}^{p}\left(x_j - \overline{x}_j\right)^2 \quad (8)$$

$$SSW = \sum_{a=1}^{n_g}\sum_{g=1}^{k}\sum_{i=1}^{p}\left(x_{aig} - \overline{x}_{ig}\right)^2 \quad (9)$$

e-ISSN: 2580-3239
p-ISSN: 2085-7993
Volume 22 No. 2 November 2023

di mana
SST : *Sum of Square Total*
SSB : *Sum of Square Between Group*
SSW : *Sum of Square Within Group*
n_g : banyaknya objek pada *cluster* ke-g
k : banyaknya *cluster* yang dibentuk dalam pengamatan
x_{aig} : sampel objek a variabel i pada *cluster* ke-g
x_i : sampel variabel i
$\overline{x}_i$: rata-rata sampel variabel i
$\overline{x}_{ig}$: rata-rata sampel variabel i pada *cluster* ke-g.

2.6 Pemilihan Metode Terbaik

Metode yang populer digunakan dalam mengukur kualitas *cluster* yang dihasilkan pada algoritma *hierarchical clustering* adalah metode *Cophenetic Correlation Coefficient* (*CPCC*). Metode yang paling mendekati nilai 1 merupakan metode yang paling baik dan tepat untuk digunakan (Li *et al*., 2022). Perhitungan nilai *Cophenetic Correlation Coefficient* menggunakan persamaan

$$CPCC = \frac{\sum_{a<b}\left(d(a,b)-\overline{d}\right)(w(a,b)-\overline{w})}{\sqrt{\sum_{a<b}(d(a,b)-d)^2}\sqrt{\sum_{a<b}(w(a,b)-\overline{w})^2}} \quad (10)$$

dengan
$d(a,b)$: *Euclidean distance* antara objek a dan b
$w(a,b)$: *dendrogram distance* antara objek a dan b
$\overline{d}$: rata-rata *Euclidean distance*
$\overline{w}$: rata-rata *dendrogram distance*.

3 METODE PENELITIAN

Metode yang digunakan pada penelitian ini adalah metode *cluster agglomerative* hierarki yaitu metode *average linkage* dan *centroid method* dengan bantuan bahasa pemrograman Python dan Ms. Excel.

4 HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Data yang dikumpulkan adalah data sekunder berupa data jumlah kasus penyakit menular di Provinsi Jawa Barat tahun 2021. Data kasus penyakit menular yang digunakan adalah data kasus penyakit pneumonia, tuberkulosis, HIV, diare dan DBD. Data diambil berdasarkan kabupaten/kota di Provinsi Jawa Barat.

4.2 Uji Multikolinearitas

Uji multikolinearitas dilakukan untuk mendeteksi adanya korelasi antar variabel. Pengujian dilaksanakan dengan mencari nilai VIF variabel prediktor terhadap variabel lainnya. Terdapat lima variabel yang digunakan pada penelitian ini. Perhitungan nilai VIF variabel-variabel dilakukan dengan menggunakan Ms.Excel dan didapat hasil seperti pada Tabel 1.

Tabel 1. Nilai VIF.

Variabel	R^2	Nilai VIF
Pneumonia	0,3151	1,4601
Tuberkulosis	0,4406	1,7878
HIV	0,2576	1,3470
Diare	0,4514	1,8227
DBD	0,2404	1,3165

Berdasarkan Tabel 1, tidak ada variabel yang memiliki nilai VIF lebih dari 10, sehingga dapat disimpulkan bahwa tidak terjadi multikolinearitas antar variabel yang digunakan.

4.3 Standarisasi Data

Data yang digunakan pada penelitian ini memiliki perbedaan rentang nilai antar variabel yang besar. Oleh karena itu dilakukan proses standarisasi data dengan transformasi data ke dalam bentuk *z-score*. Standarisasi data ke dalam bentuk *z-score* dilakukan dengan meggunakan bahasa pemrograman Python dengan *script* sebagai berikut.

```
import scipy.stats as stats
data = df.apply(stats.zscore)
```

4.4 Perhitungan Analisis Cluster

Perhitungan analisis *cluster* dengan metode *average linkage* dan *centroid method*. Perhitungan dilakukan menggunakan bahasa

e-ISSN: 2580-3239
p-ISSN: 2085-7993
Volume 22 No. 2 November 2023

pemrograman Python dengan *script* sebagai berikut.

```
from scipy.cluster.hierarchy import
dendrogram, linkage
import matplotlib.pyplot as plt

average = linkage(data,
method="average",
metric="euclidean")
plt.figure(figsize=(8, 6))
dendrogram = dendrogram(average)
plt.show()

centroid = linkage(data,
method="centroid",
metric="euclidean")
plt.figure(figsize=(8, 6))
dendrogram = dendrogram(centroid)
plt.show()
```

4.5 Penentuan Metode Terbaik

Pada penelitian ini, metode terbaik ditentukan dengan melihat nilai *Cophenetic Correlation Coefficient* (CPCC). Nilai CPCC menghitung korelasi dari jarak *euclidean* antar objek dengan jarak dendogram yang dihasilkan pada proses analisis *cluster*. Perhitungan nilai CPCC dari masing-masing metode cluster menggunakan bahasa pemrograman Python dengan script sebagai berikut.

```
from scipy.cluster.hierarchy import
dendrogram, linkage
from scipy.cluster.hierarchy import
cophenet
from scipy.spatial.distance import
pdist
import pylab

ca = linkage(data, 'average')
c, coph_dists_ca = cophenet(ca,
pdist(data))
c

cc = linkage(data, 'centroid')
c, coph_dists_cc= cophenet(cc,
pdist(data))
c
```

Nilai CPCC dari metode analisis *cluster* yang digunakan pada penelitian ini tertera pada Tabel 2.

Tabel 2. Nilai CPCC masing-masing metode.

Metode Analisis *Cluster*	Nilai CPCC
Average Linkage	0,8513
Centroid Method	0,8380

Berdasarkan Tabel 2, metode *average linkage* adalah metode dengan nilai CPCC yang paling mendekati nilai 1. Maka dari itu dapat disimpulkan bahwa metode *average linkage* merupakan metode terbaik dalam mengelompokkan kabupaten/kota di Provinsi Jawa Barat berdasarkan kasus penyakit menular.

4.6 Penentuan Jumlah Cluster Optimal

Penentuan jumlah *cluster* optimal dilakukan dengan menggunakan nilai *Pseudo-F*. Perhitungan nilai *Pseudo-F* dilakukan pada metode terbaik yaitu *average linkage* menggunakan bahasa pemrograman Python dengan *script* sebagai berikut.

```
from scipy.cluster.hierarchy import
average, fcluster
from sklearn.metrics import
calinski_harabasz_score

aver = average(data)
ca2 = fcluster(aver, 2,
criterion='maxclust')
calinski_harabasz_score(data,ca2)

ca3 = fcluster(aver, 3,
criterion='maxclust')
calinski_harabasz_score(data,ca3)

ca4 = fcluster(aver, 4,
criterion='maxclust')
calinski_harabasz_score(data,ca4)

ca5 = fcluster(aver, 5,
criterion='maxclust')
calinski_harabasz_score(data,ca5)
```

Nilai *Pseudo-F* hasil analisis *cluster* dengan metode *average linkage* tersaji pada Tabel 3.

e-ISSN: 2580-3239
p-ISSN: 2085-7993
Volume 22 No. 2 November 2023

Tabel 3. Nilai *Pseudo-F average linkage.*

Jumlah *Cluster*	Nilai *Pseudo-F*
2	7,5053
3	7,9841
4	7,1540
5	9,8854

Tabel 3 Menunjukkan bahwa nilai *Pseudo-F* tertinggi adalah 9,8854 dengan jumlah *cluster* yang terbentuk sebanyak 5 *cluster*. Maka dari itu, jumlah *cluster* optimal dalam pengelompokkan data dengan metode *average linkage* adalah sebanyak 5 *cluster*.

4.7 Interpretasi Hasil

Berdasarkan penjelasan sebelumnya, didapatkan metode terbaik yaitu *average linkage* dengan jumlah *cluster* optimal sebanyak 5 *cluster*. Dengan menggunakan bahasa pemrograman Python, anggota-anggota dari kelima *cluster* tersebut diberikan dalam Tabel 4.

Tabel 4. Anggota *cluster*

Cluster	Jumlah Anggota *Cluster*	Anggota *Cluster*
1	3	Kab. Bandung, Kota Bandung, dan Kota Depok
2	4	Kab. Sukabumi, Kab. Cianjur, Kab. Ciamis, dan Kab. Cirebon
3	17	Kab. Garut, Kab. Tasikmalaya, Kab. Kuningan, Kab. Majalengka, Kab. Sumedang, Kab. Indramayu, Kab. Subang, Kab. Karawang, Kab. Bandung Barat, Kab. Pangandaran, Kota Bogor, Kota Sukabumi, Kota Cirebon, Kota Bekasi, Kota Cimahi, Kota Tasikmalaya, dan Kota Banjar
4	1	Kab. Bogor
5	2	Kab. Purwakarta dan Kab. Bekasi

Interpretasi hasil dilakukan dengan mencari nilai rata-rata masing-masing variabel dari setiap *cluster* yang terbentuk berdasarkan karakteristik atau ciri-ciri dari masing masing anggota *cluster* sebagai berikut:

1. *Cluster* 1 memiliki jumlah anggota sebanyak 3 kabupaten/kota dengan nilai rata-rata kasus penyakit menular yang rendah yaitu sebanyak 5.252 kasus. Namun perlu adanya perhatian khusus pada kasus penyakit Demam Berdarah Dengue (DBD) dimana *cluster* pertama mempunyai rata-rata kasus penyakit DBD yang paling tinggi dibandingkan *cluster* lainnya.
2. *Cluster* 2 terdiri dari 4 kabupaten, dimana nilai rata-rata kasus penyakit menular dari *cluster* ini adalah sebesar 8.348 kasus. Nilai rata-rata kasus penyakit menular ini tergolong tinggi dengan nilai rata-rata kasus penyakit Pneumonia yang paling tinggi dari pada *cluster* lainnya. Selain itu *cluster* ini juga memiliki nilai rata-rata kasus diare yang tinggi yaitu sebesar 33.006 kasus.
3. *Cluster* 3 merupakan *cluster* dengan anggota terbanyak yaitu sebanyak 17 kabupaten/kota. *Cluster* ini adalah *cluster* dengan nilai rata-rata kasus penyakit menular yang paling rendah. Tidak ada yang perlu dikhawatirkan dari *cluster* ini, mengingat tidak ada nilai rata-rata penyakit yang tergolong tinggi. Bahkan nilai rata-rata penyakit Pneumonia, Tuberkulosis dan diare *cluster* ini merupakan yang paling rendah dibandingkan *cluster* lainnya.
4. *Cluster* 4 mempunyai anggota yang paling sedikit dengan beranggotakan satu kabupaten saja, yaitu Kabupaten Bogor. Akan tetapi *cluster* ini merupakan *cluster* dengan nilai rata-rata kasus penyakit menular yang paling tinggi yaitu sebanyak 21.768 kasus. Penyakit Tuberkulosis, HIV dan diare merupakan prioritas utama dari *cluster* ini mengingat nilai rata-rata penyakit tersebut merupakan yang paling tinggi dibandingkan dengan *cluster* lain. Selain itu nilai rata-rata dua penyakit lainnya, yaitu pneumonia dan DBD juga tergolong tinggi.

e-ISSN: 2580-3239
p-ISSN: 2085-7993
Volume 22 No. 2 November 2023

5. *Cluster* 5 terdiri dari dua kabupaten dengan nilai rata-rata penyakit menular yang tergolong sedang yaitu sebesar 6.065 kasus. Pada *cluster* ini penyakit yang perlu menjadi perhatian adalah penyakit Tuberkulosis dan HIV dimana nilai rata-rata kedua penyakit ini yang tergolong tinggi.

5 SIMPULAN

Berdasarkan hasil dan pembahasan yang telah dilakukan mengenai data jumlah kasus penyakit di Provinsi Jawa Barat menggunakan metode *average linkage* dan *centroid method* dapat ditarik kesimpulan bahwa berdasarkan nilai *Cophenetic Correlation Coefficient* (CPCC), metode terbaik dalam mengelompokkan Kabupaten/Kota Provinsi Jawa Barat tahun 2021 berdasarkan kasus penyakit menular adalah metode *Average Linkage* dengan nilai CPCC sebesar 0,8513. Metode *Average Linkage* membagi Kabupaten/Kota di Provinsi Jawa Barat menjadi 5 *cluster* berdasarkan nilai *Pseudo-F* tertinggi.

DAFTAR PUSTAKA

Afira, N. & Wijayanto, A.W. (2021). Analisis Cluster dengan Metode Partitioning dan Hierarki pada Data Informasi Kemiskinan Provinsi di Indonesia Tahun 2019. *Komputika: Jurnal Sistem Komputer*, *10*(2), 101-109.

Dani, A.T.R., Wahyuningsih, S. & Rizki, N.A. (2019). Penerapan Hierarchical Clustering Metode Agglomerative pada Data Runtun Waktu. *Jambura Journal of Mathematics*, *1*(2), 64-78.

Dinas Kesehatan Provinsi Jawa Barat (2022). *Profil Kesehatan Provinsi Jawa Barat 2021*.

Direktorat Jenderal Pencegahan dan Pengendalian Penyakit (2020). *Rencana Aksi Program P2P 2020-2024*.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis, 5th Edition. Wiley Series in Probability and Statistics*. West Sussex: John Wiley and Sons Ltd.

Gudono (2011). *Analisis Data Multivarat (Edisi Pertama)*. Yogyakarta: BPFE.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis, Eighth Edition. United Kingdom: Cengage Learning EMEA*.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*. Amsterdam: Elsevier.

Johnson, R. A. & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis.* New York: Prentice-Hall, Inc.

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons.

Li, T., Rezaeipanah, A. & El Din, E.M.T. (2022). An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 3828-3842.

Nakayama, T., Asaka, D., Yoshikawa, M., Okushi, T., Matsuwaki, Y., Moriyama, H. & Otori, N. (2012). Identification of chronic rhinosinusitis phenotypes using cluster analysis. *American journal of rhinology & allergy*, *26*(3), 172-176.

Pusdiktasari, Z.F., Sasmita, W.G., Fitrilia, W.R., Fitriani, R. & Astutik, S. (2021). The Clustering of Provinces in Indonesia by The Economic Impact of Covid-19 using Cluster Analysis: Pengelompokkan Provinsi di Indonesia dengan Ekonomi Terdampak Covid-19 Menggunakan Analisis Cluster. *Indonesian Journal of Statistics and Its Applications*, *5*(1), 117-129.

Sembiring, R.K. (2003). *Analisis Regresi Edisi Kedua*. Bandung: ITB.

Srinath, K.R. (2017). Python–the fastest growing programming language. *International Research Journal of*

e-ISSN: 2580-3239
p-ISSN: 2085-7993
Volume 22 No. 2 November 2023

Engineering and Technology, 4(12), 354-357.

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to data mining*. London: Pearson Education.

Ulinnuh, N., & Veriani, R. (2020). Analisis Cluster dalam Pengelompokan Provinsi di Indonesia Berdasarkan Variabel Penyakit Menular Menggunakan Metode Complete Linkage, Average Linkage dan Ward. *InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan*, *5*(1), pp. 102-108.

Whendasmoro, R.G. & Joseph, J. (2022). Analisis Penerapan Normalisasi Data Dengan Menggunakan Z-Score Pada Kinerja Algoritma K-NN. *JURIKOM (Jurnal Riset Komputer)*, *9*(4), pp.872-876.

Widayat (2018). *Statistika Multivariat (Pada Bidang Manajemen Dan Bisnis)*. Malang: Universitas Muhammadiyah Malang (UMM Press).

Wijaya, T., & Budiman, S. (2016). *Analisis Multivariat untuk Penelitian Manajemen*. Yogyakarta: Penerbit Pohon Cahaya.