

# Penerapan *K-Means Cluster* dan Evaluasi *Clustering* pada Pesebaran Kasus *Covid-19*

EUIS SARTIKA<sup>1</sup>, SRI MURNIATI<sup>2</sup>, AGUS BINARTO<sup>3</sup>, ENDANG HABINUDDIN<sup>4</sup>

<sup>1</sup>)Program Studi Manajemen Pemasaran, Administrasi Niaga Politeknik Negeri Bandung, Indonesia

<sup>2</sup>)Program Studi Teknik Refrigerasi dan tata Udara, Teknik Mesin Politeknik Negeri Bandung, Indonesia

<sup>3</sup>)Program Studi Teknik Otomasi Industri, Teknik Elektro Politeknik Negeri Bandung, Indonesia

<sup>4</sup>)Program Studi Teknik Telekomunikasi, Teknik Elektro Politeknik Negeri Bandung, Indonesia

e-mail: euis.sartika@polban.ac.id

## ABSTRAK

Indonesia termasuk negara dengan jumlah pasien *Covid-19* kelompok tinggi. Jumlah penduduk yang besar berpotensi dalam penularan virus. Diperlukan upaya untuk mengurangi penularan virus ini, diantaranya adalah memperoleh keterangan tentang sifat data penderita *Covid-19* untuk tiap wilayah lalu mengelompokkannya berdasarkan kesamaan sifat dari pasien. Analisis yang paling tepat untuk masalah ini adalah *K-Means Cluster* dengan mengklasifikasikan  $n$  objek berdasarkan pesebaran kasus *Covid-19* dan mengevaluasi hasil *Clustering* menggunakan Davies Bouldin Index (DBI). Metode evaluasi optimasi *Cluster* lain yang digunakan adalah Elbow dan Silhouette. Variabel yang digunakan adalah pasien positif ( $X_1$ ), pasien sembuh ( $X_2$ ), pasien meninggal ( $X_3$ ), jumlah penduduk ( $X_4$ ), kepadatan penduduk ( $X_4$ ), usia lansia ( $X_5$ ), dan sarana kesehatan ( $X_6$ ). Informasi penelitian ini diharapkan dapat memberi alternatif dalam mengupayakan pencegahan *Covid-19* lebih luas lagi. Hasil penelitian menunjukkan bahwa *K-Means Clustering* dengan dua *Cluster* menghasilkan nilai DBI sebesar 2,0027 dengan nilai koefisien Silhouette adalah 71,18%. Analisis *K-Means Cluster* dengan tiga *Cluster* menghasilkan nilai DBI sebesar 1,0055 lebih kecil dari nilai DBI dua *Cluster*, berarti *Clustering* dengan tiga *Cluster* lebih baik dibandingkan dua *Cluster*. Variabel-variabel yang paling mendominasi dalam *K-Means Cluster* ini adalah jumlah penduduk, kepadatan penduduk, jumlah penduduk lansia, dan sarana Kesehatan.

*Kata Kunci:* *Covid-19*, analisis *Cluster*, *K-Means Cluster*

## ABSTRACT

Indonesia is a country with a high number of *Covid-19* patients. A large population has the potential for virus transmission. Efforts are needed to reduce the transmission of this virus, including obtaining information about the nature of the data for *Covid-19* sufferers for each region and then grouping them based on similar characteristics of the patients. The most appropriate analysis for this problem is the *K-Means Cluster* by classifying  $n$  objects based on the distribution of *Covid-19* cases and evaluating the results of clustering using the Davies Bouldin Index (DBI). Another *Cluster* optimization evaluation method used is Elbow and Silhouette. The variables used were positive patients ( $X_1$ ), recovered patients ( $X_2$ ), deceased patients ( $X_3$ ), population ( $X_4$ ), population density ( $X_4$ ), elderly age ( $X_5$ ), and health facilities ( $X_6$ ). It is hoped that this research information can provide an alternative in seeking wider prevention of *Covid-19*. The results showed that *K-Means Clustering* with two clusters produced a DBI value of 2.0027 with a Silhouette coefficient value of 71.18%. The *K-Means Cluster* analysis with three clusters produces a DBI value of 1.0055 which is smaller than the DBI value of two clusters, meaning that a cluster with three clusters is better than two clusters. The most dominating variables in this *K-Means Cluster* are population size, population density, number of elderly people, and health facilities.

*Keywords:* *Covid-19*, *Cluster analysis*, *K-Means Cluster*

## 1. PENDAHULUAN

Di Indonesia, kasus pasien penderita *Covid-19* sampai tanggal 8 Februari 2021 sebesar 1.157.837 jiwa dengan kasus meninggal adalah 31.556 jiwa, dimana golongan usia 46 -59 tahun memiliki persentase tertinggi. Provinsi Jabar, Jateng, Jatim, dan DKI Jakarta, dengan laju jumlah penduduk per Km persegi tertinggi mencapai urutan teratas kasus pasien positif. Faktanya, wilayah tersebut mempunyai penduduk dalam jumlah yang besar, dengan rata-rata jumlah penduduk per wilayah bernilai besar juga. Oleh karena itu, banyaknya penduduk per wilayah dan laju jumlah penduduk per Km<sup>2</sup> menjadi peubah yang dipakai dalam riset ini (Ganasegeran, 2021). Luas daratan Indonesia yang mencapai 1,91 juta km<sup>2</sup> dengan pulau-pulau sebanyak 17.504 buah, ditambah banyaknya penduduk, hal ini menjadi tantangan pemerintah agar lebih fokus lagi dalam mencari upaya untuk mengurangi penyebaran virus lebih luas lagi. Upaya tersebut antara lain memperoleh keterangan mengenai sifat atau karakteristik penderita *Covid-19* di Indonesia berdasarkan wilayah yang diklasifikasikan berdasarkan penderita positif ke dalam kelas yang memiliki sifat yang mirip, selanjutnya akan diperoleh karakteristik dari tiap-tiap kelas (Sari and Sukestiyarno, 2021).

*Clustering* adalah salah satu metode penyelesaian multi *variable* yang bertujuan mengklasifikasikan data menurut kemiripan sifat. *Clustering* mengelompokkan data yang memiliki kesamaan sangat mirip dengan data lain yang terletak dalam satu kelompok. Pengelompokkan yang tepat memiliki kehomogenan sifat yang besar antar objek dalam satu kelompok dan keheterogenan yang besar antar kelompok (A.GoreJr, 2000). Novelty penelitian ini adalah menerapkan data terkini yaitu data kasus *Covid-19* pada tahun 2021 dan beberapa riset pendahuluan menerapkan peubah kependudukan, selain faktor kependudukan juga faktor kelompok usia lansia (Dowd et al., 2020) dan sarana kesehatan. Alasan faktor kependudukan dipakai sebagai salah satu peubah penelitian, didasarkan pada hasil riset yang dijelaskan dalam jurnal penelitian di beberapa negara, selain peubah faktor lingkungan (Abed and Lashin, 2021). *K-Means Clustering* adalah salah satu metode *Cluster* non-hierarki yang bertujuan mengelompokkan objek yang mempunyai kemiripan sifat digabung dalam satu *Cluster*. Alat untuk mengukur kesamaan karakteristik tersebut adalah jarak antara objek. Objek dengan jarak paling dekat akan dikelompokkan dalam *Cluster* yang sama, dan mengindikasikan bahwa objek-objek tersebut memiliki kesamaan karakteristik. Kelemahan yang dimiliki *K-Means* adalah pada saat menentukan pusat awal *Cluster*, yang berdampak pada *Cluster* yang terbentuk hanya bersifat local. Kelemahan lainnya adalah *K-Means* cukup bermasalah dalam hal dimensi jika dihadapkan pengolahan data yang mempunyai dimensi besar.

Riset ini bertujuan mengaplikasikan analisis *K-Means Cluster* dalam pengelompokkan wilayah *Covid-19* dan mengetahui kelompok wilayah mana yang termasuk dalam *Cluster* tingkat pesebaran kasus *Covid-19*, serta menentukan *variable* yang menunjukkan ketidaksamaan paling besar pada *Cluster* yang terbentuk. Diharapkan hasil penelitian ini dapat memberi manfaat bagi pihak terkait untuk dapat memutus mata rantai penularan virus lebih luas lagi di Indonesia.

Penelitian yang relevan:

- Andayani. S., (2007) menyatakan bahwa Algoritma *K-Means* dapat digunakan untuk pengklasifikasian *Cluster* dalam sebuah basis data dengan ukuran besar dengan menerapkan aturan kesamaan dan ketidaksamaan data terlebih dahulu data dalam basis data (Andayani, 2016).
- Imam Wahyudi, dkk (2021) menyatakan bahwa penentuan *Cluster* terbaik pada metode *K-Means* terhadap Sentra Industri Produksi Di Pamekasan dengan menggunakan Elbow nilai DBI-nya adalah 0,515 sedangkan tanpa menggunakan Elbow, DBI-nya adalah 0,635 (Wahyudi, Sulthan, & Suhartini, 2021)
- Talakua. M.W, dkk, (2017), menyatakan bahwa kabupaten / kota provinsi Maluku dapat dikelompokkan ke dalam tiga *Cluster* berdasarkan Indeks Pembangunan Manusia (Talakua, Leleury and Taluta, 2017).

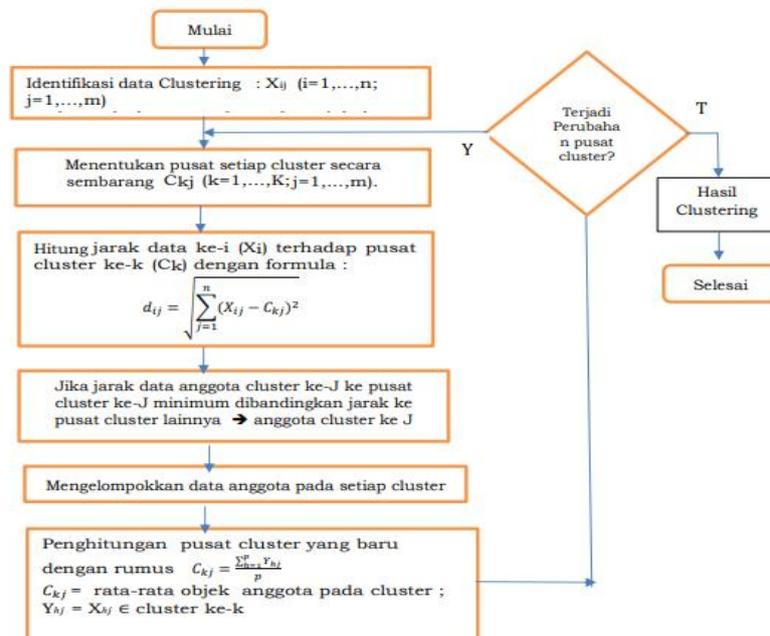
*Clustering* dapat mengelompokkan objek, dimana semua objek yang memiliki sifat yang paling dekat kemiripannya akan mengelompok kedalam satu *Cluster*. Karakteristik *Cluster* mempunyai tingkat homogenitas yang besar antara individu dalam *Cluster* yang sama (within *Cluster*), mempunyai tingkat heterogenitas yang besar antara *Cluster* yang berbeda (between-*Cluster*). Beberapa istilah penting dalam analisis *Cluster* (Ghozali, I, 2013) adalah:

- Distances between *Cluster* centers, yakni menggambarkan jarak antar pusat *Cluster*.
- *Cluster* membership, yakni menunjukkan keanggotaan.
- *Cluster* centers, ialah titik pangkal dilakukan awal pengklasifikasian dalam analisis *Cluster* non-hierarki.
- Rata-rata lama *Cluster* (*Cluster centroid*), ialah nilai average peubah dari semua objek pada suatu *Cluster* tertentu.
- Skedul aglomerasi ialah skedul yang menunjukkan keterangan tentang objek atau kasus yang diklasifikasikan pada setiap langkah dalam analisis hierarki.

Sampel harus mewakili populasi yang diteliti dan nonmultikolinearitas merupakan dua hal (asumsi) yang harus dipenuhi dalam analisis *Cluster*. Uji multikolinearitas digunakan untuk mengetahui apakah terdapat kemiripan antar variabel bebas yang satu dengan *variable* bebas yang lain (Yulianto and Hidayatullah, 2016). Untuk mengatasi gejala multikolinearitas, antara lain dilakukan perubahan data ke dalam bentuk lain, antara lain logaritma natural (ln), fungsi akar, dan lain-lain (Ghozali, 2018). Gejala multikolinieritas ditunjukkan dengan nilai tolerance. Multikolinieritas terjadi, apabila nilai tolerance  $\leq 0,1$  atau dengan melihat nilai VIF (*Variance Inflation Faktor*). Gejala multikolinieritas pada data ditunjukkan dengan nilai VIF  $\geq 10$ .

Algoritma analisis *Cluster* dapat dibagi kedalam dua kelompok yaitu *Clustering* hirarki dan non hirarki (Ghozali, 2018). Analisis cluster hirarki disebut juga klasifikasi *Cluster* yang bersarang atau *Cluster* dengan struktur pohon, tepatnya pohon berjenjang. Langkah dimulai dengan pengklasifikasian dua atau lebih dari dua objek yang memiliki jarak minimum. Kemudian dilanjutkan dengan pengelompokan *Cluster* yang kedua, dan seterusnya sampai diperoleh pohon *Cluster* yang membentuk hierarki, untuk lebih jelasnya ditunjukkan oleh dendrogram. Selanjutnya, analisis non hirarki yang memiliki langkah kebalikan dari prosedur hirarki tidak melakukan pembentukan diagram pohon, tetapi diawali dengan menentukan banyak *Cluster* yang diharapkan. Setelah penentuan banyak *Cluster*, proses *Clustering* dilakukan namun tidak mengikuti tahapan hierarki, metode ini dikenal sebagai *K-Means Clustering*. Asumsi Normalitas harus dipenuhi pada analisis *K-Means Clustering* dan ini merupakan salah satu kelemahan *K-Means Clustering*, kelemahan lainnya adalah data tidak boleh mempunyai pencilan dan hanya bisa diterapkan untuk data yang rata-ratanya dapat dicari dengan mudah (Andayani, 2016). Proses analisis *K-Means* melakukan pengklasifikasian data semula menjadi satu kelompok, sehingga data dalam satu group, mempunyai sifat yang berlainan dengan data yang ada didalam group yang berbeda (Saputro and Helilintar, 2020). Pada prinsipnya, penentuan pusat *Cluster* pada analisis *K-Means* bersifat iteratif. Penentuan pusat *Cluster* dilakukan berdasarkan konsep jarak antar setiap data ke pusat *Cluster*.

Langkah-langkah *K-Means Clustering* adalah sebagai berikut:



Gambar 1. *Flowchart K-Means Clustering*

## 2. METODE PENELITIAN

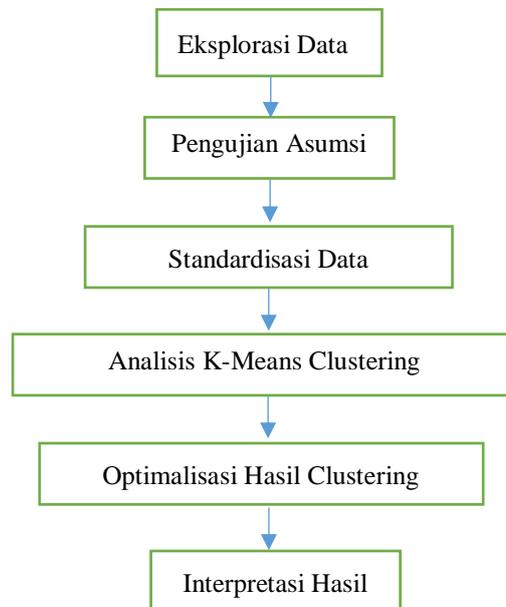
Tahapan yang telah dikerjakan sebelum menginjak langkah penelitian adalah:

- Melakukan pengumpulan artikel-artikel mengenai penerapan *K-Means Clustering*.
- Memilih artikel yang belum dan sudah terbit, dengan mencermati aplikasi *K-Means Clustering* yang sangat diperlukan dan berpeluang untuk dibuat sebuah riset.
- Menganalisis mengenai penerapan *K-Means Clustering* pada persebaran kasus *Covid-19* dengan memakai data tahun 2021 dan peubah penelitian dimana peneliti lain belum ada yang menggunakannya dalam penelitian yang setara.

Tahapan Penelitian adalah:

Riset ini menggunakan data sekunder dengan peubah-peubah: jumlah pasien (X1), jumlah penderita sembuh (X2), dan jumlah penderita meninggal diper (X3) didapat dari GTPPC peride tahun 2021. Peubah banyak penduduk (X4), banyak penduduk per Km2 (X5), kelompok penduduk usia lansia (X6), dan sarana kesehatan (X7) didapat dari BPS 2021. Metode statistik yang digunakan adalah analisis deskriptif dan *K-Means Cluster*.

Langkah Penelitian sebagai berikut:



Gambar 2. Flowchart Langkah Penelitian

## 3. HASIL DAN PEMBAHASAN

### Analisis Deskriptif

Tabel 1. Nilai dari Deskriptif Variabel-variabel

Variabel	Minimum	Maximum	Mean	Std. Deviation
X1	13901	1250715	178085.71	290260.743
X2	13397	1234305	173376.97	284335.839
X3	291	33290	4609.24	7921.124
X4	713600	48782400	7658373.53	11346397.772
X5	9	15978	744.26	2721.058
X6	114176	11867100	1802459.91	2964648.041
X7	11	329	73.29	82.515
Banyak data	34			

Berdasarkan tabel 1, dapat ditunjukkan dimana pasien positif maksimum sebesar 1250715 jiwa dan minimum sebesar 13901 jiwa periode tahun 2021. Untuk pasien sembuh maksimum 1234305 jiwa dan minimum adalah 13397 jiwa. Sedangkan pasien yang meninggal, minimum 291 jiwa dan maksimum adalah 33290 jiwa. Dapat disimpulkan bahwa, pasien positif dan dinyatakan sembuh cukup besar yakni mendekati 99%.

**Standardisasi**

Proses standardisasi perlu dilakukan pada data sebelum data diolah. Hal ini dilakukan agar tidak terjadi ketimpangan data, terutama nilai dari data. Satuan data biasanya berbeda, ada persentase, unit, km, dan sebagainya. Untuk menyamakannya maka dibuat langkah standardisasi sebelum diolah datanya.

**Pengujian Asumsi**

- Multikolinearitas Test ditujukan untuk menentukan ada tidaknya hubungan antar peubah-peubah independent. Untuk mengatasi data non multikolinearitas, dapat diatasi dengan cara pengubahan nilai (logaritma, akar, pangkat, dsb), eliminasi, atau membuang data yang ekstrim. Hasil analisis ditunjukkan pada tabel 2.

Tabel 2. Hasil Pengujian Multikolinearitas

Variabel	Tolerance	VIF
X1	0,122	8,228
X2	0,115	6,448
X3	0,142	7,027
X4	0,271	3,692
X5	0,571	1,934
X6	0,101	9,878
X7	0,112	8,908

Berdasarkan tabel 2, dapat ditunjukkan bahwa setiap nilai tolerance lebih tinggi dari 0,10 dan nilai VIF dari setiap variabel mencapai kurang dari 10. Analisis menunjukkan bahwa tidak terjadi multikolinearitas diantara peubah-peubah penelitian.

- Hasil Uji KMO (Kaiser-Meyer-Olkin)
  - a) Uji KMO (Kaiser-Meyer-Olkin) digunakan untuk menyelidiki apakah analisis faktor layak dilakukan pada riset ini. Range besaran nilai KMO adalah  $0 < \text{nilai KMO} < 1$ .
  - b) Jika nilai KMO  $< 0.5$  maka tidak dapat ditentukan analisis selanjutnya artinya analisis faktor tidak dapat layak dilakukan. Sebaliknya, jika nilai KMO  $> 0.5$  maka layak dilakukan analisis selanjutnya artinya analisis faktor dapat ditentukan. Hasil analisis menunjukkan nilai KMO sebesar 0.639 ( $0 < \text{nilai KMO} < 1$ ), maka analisis faktor layak ditentukan. Dapat ditunjukkan juga bahwa nilai *Bartlett's test of sphericity* atau nilai probabilitasnya menunjukkan nilai 0.000 (nilai KMO  $< 0,05$ ), dapat disimpulkan bahwa peubah-peubah penelitian saling berhubungan dan layak diteruskan ke langkah berikutnya.

Tabel 3. Uji KMO dan Barletts

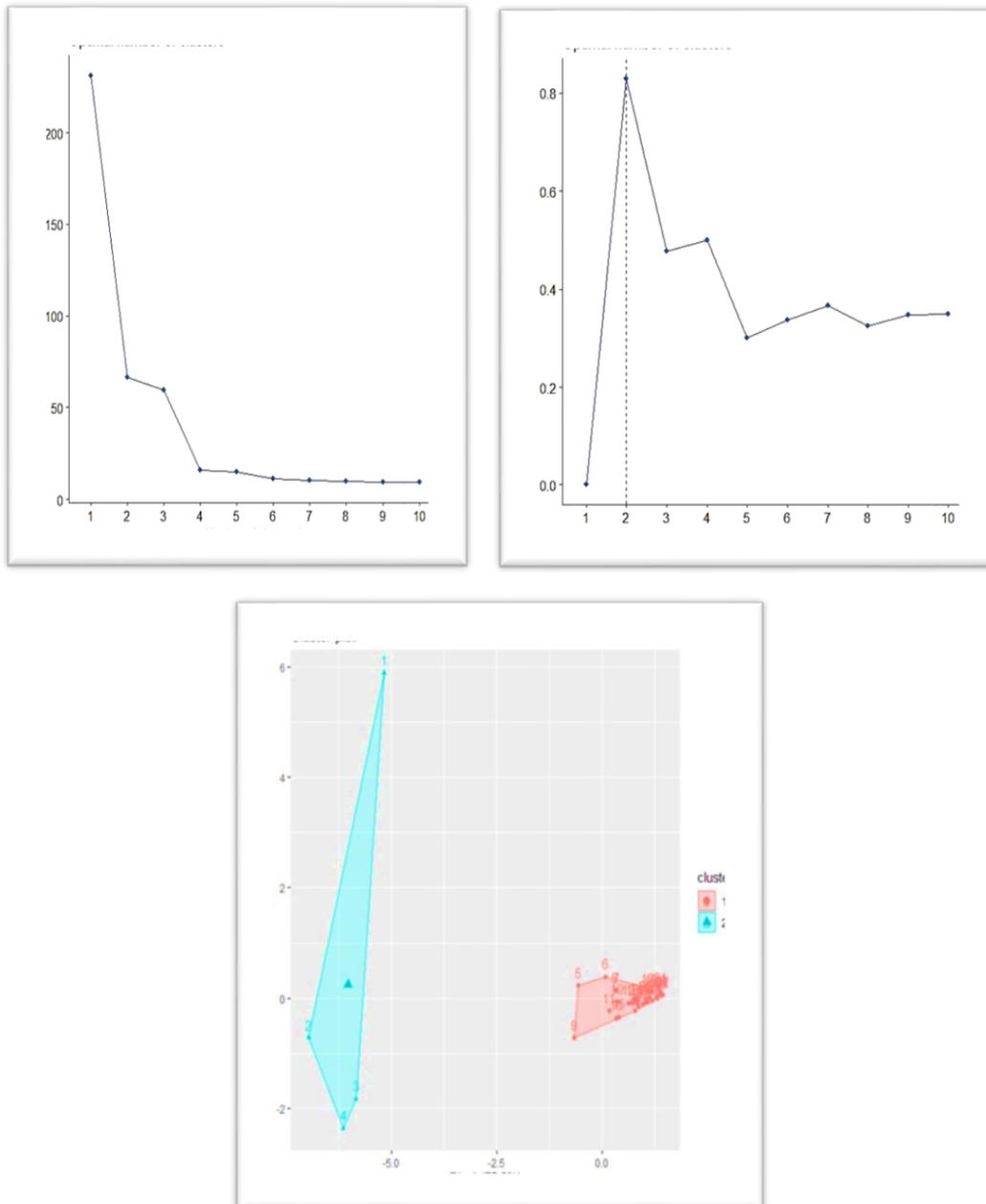
Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy	0,639
Bartlett's Test of Sphericity (Sig)	0,000

Tabel 3 menunjukkan dimana nilai Bartlett's Test of Sphericity (probabilitas) adalah 0,000 lebih kecil dari 0,05 maka layak ditentukan analisis berikutnya.

Asumsi *K-Means Clustering* sudah dipenuhi, berikutnya ditentukan langkah *K-Means Clustering* menggunakan Rstudio 4.2.2, dengan hasil analisis sebagai berikut:

- a) Hasil analisis Elbow menunjukkan dataset dikelompokkan ke dalam dua *Cluster*, karena terdapat dua siku, kurva pertama turun sangat tajam dan membentik siku.

Kurva kedua turun walaupun tidak tajam namun membentuk siku. Selanjutnya kurva landai, sehingga dapat disimpulkan bahwa dataset berdasarkan Elbow menunjukkan dua *Cluster*. Sillhoutte menunjukkan nilai rata-rata *Cluster* sebesar 2, artinya dataset dikelompokkan kedalam dua *Cluster*. *Cluster* plot memperlihatkan bahwa dataset berkelompok membentuk dua area. Hasil output Rstudio ditunjukkan dalam gambar berikut:



Gambar 4. Optimasi *Cluster* Menggunakan Elbow dan Silhouette serta *Cluster* Plot

Nilai *Cluster* means ditunjukkan oleh tabel berikut:

Tabel 4. Nilai *Cluster Means* untuk Masing-masing Variabel

<b>Cluster</b>	<b>X1</b>	<b>X2</b>	<b>X3</b>	<b>X4</b>	<b>X5</b>	<b>X6</b>	<b>X7</b>
1	-0.3271657	-0.3247002	-0.3266037	-0.3125218	-0.2003503	-0.3031904	-0.3117918
2	2.4537431	2.4352515	2.4495280	2.3439136	1.5026269	2.2739277	2.3384386

Berdasarkan tabel 4, dapat diperlihatkan bahwa nilai *Cluster* means untuk variabel-variabel dominan ada pada *Cluster* 2. Nilai koefisien Sillhoutte dari penelitian adalah 71,3%, hal menunjukkan bahwa cluster 2 cukup baik karena nilai koefisien Sillhoue mendekati 100%.

Perbandingan Analisis *Cluster* yang menggunakan dua *Cluster* dan tiga *Cluster* disajikan dalam table berikut:

Tabel 5. Banyaknya Provinsi Berdasarkan *Cluster*

<b>Banyaknya Cluster</b>	<b>Cluster ke-</b>	<b>Banyak Provinsi tiap Cluster</b>	<b>Provinsi untuk Masing-masing Cluster</b>
2	1	4	DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur
	2	30	Banten, DI Yogyakarta, Kalimantan Timur, Bali, Sumatra Utara, Riau, Sulawesi Selatan, Sumatra Barat, Nusa Tenggara Timur, Kalimantan Selatan, Sumatra Selatan, Lampung, Kalimantan Barat, Sulawesi Utara, Aceh, Nusa Tenggara Barat, Kepulauan Riau, Kepulauan Bangka Belitung, Sulawesi Tengah, Kalimantan Tengah, Papua, Kalimantan Utara, Jambi, Papua Barat, Sulawesi Tenggara, Bengkulu, Maluku, Sulawesi Barat, Maluku Utrara, Gorontalo
Jumlah		34	
3	1	4	DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur
	2	16	Banten, DI Yogyakarta, Kalimantan Timur, Bali, Sumatra Utara, Riau, Sulawesi Selatan, Sumatra Barat, Nusa Tenggara Timur, Kalimantan Selatan, Sumatra Selatan, Lampung, Kalimantan Barat, Sulawesi Utara, Aceh, Nusa Tenggara Barat,
	3	14	Kepulauan Riau, Kepulauan Bangka Belitung, Sulawesi Tengah, Kalimantan Tengah, Papua, Kalimantan Utara, Jambi, Papua Barat, Sulawesi Tenggara, Bengkulu, Maluku, Sulawesi Barat, Maluku Utrara, Gorontalo
Jumlah		34	

Berdasarkan tabel 5, jika banyaknya *Cluster* 2, *Cluster* dengan jumlah provinsi terbanyak adalah *Cluster* dua yakni sebanyak 30 provinsi dan *Cluster* dengan jumlah provinsi terendah adalah *Cluster* satu yaitu 5. Jika banyaknya *Cluster* 3, maka jumlah provinsi terbanyak adalah *Cluster* 2 sebanyak 16 provinsi, *Cluster* 3 dengan 14 provinsi, dan *Cluster* 1 dengan 4 provinsi.

Hasil evaluasi *Clustering* berdasarkan nilai koefisien DBI untuk *K-Means Clustering* sebanyak 2 *Cluster* dan 3 *Cluster* adalah sebagai berikut:

Tabel 6. Nilai DBI Berdasarkan Banyaknya *Cluster*

<b>Banyaknya Cluster</b>	<b>Nilai DBI</b>
2	2,00266872
3	1,00545175

Tabel 6 memperlihatkan bahwa nilai DBI untuk banyak *Cluster* 3 lebih kecil dibandingkan banyak *Cluster* 2. Hal ini menunjukkan bahwa pengelompokan dataset dalam tiga *Cluster* lebih baik dibandingkan pengelompokan dataset dalam dua *Cluster*.

Tabel 7. Jarak antara Pusat *Cluster* Akhir (*Centroid*)

<i>Cluster</i>	1	2	3
1		4,477	7,193
2	4,477		2,731
3	7,193	2,731	

Tabel 7 menunjukkan bahwa :

- Jarak *Centroid Cluster* 1 dan *Centroid Cluster* 3 adalah 7,193
- Jarak *Centroid Cluster* 1 dan *Centroid Cluster* 2 adalah 4,477
- Jarak *Centroid Cluster* 2 dan *Centroid Cluster* 3 adalah 2,731

Makin tinggi nilai jarak antar pusat *Cluster*, makin jauh jarak antar pusat *Cluster* dan hal ini menunjukkan hasil yang baik.

Tabel 8. Anova

<b>Variabel</b>	<b>Fhit</b>	<b>Sig</b>
Pasien Terkonfirmasi	51,247	0,000
Pasien Sembuh	50,018	0,000
Pasien Meninggal	55,886	0,000
Jumlah Penduduk	51,553	0,000
Kepadatan Penduduk	20,624	0,000
Jumlah Penduduk Lansia	56,897	0,000
Sarana Kesehatan	53,370	0,000

Tabel 8 menunjukkan bahwa semua *variable* dalam penelitian ini memberi pengaruh berarti. Hal ini dapat dilihat dari nilai probabilitas (Sig), semua *variable* bernilai 0,000 lebih rendah dari 0,05. Faktor yang paling mendominasi adalah jumlah penduduk, jumlah penduduk lansia, dan sarana Kesehatan. Hal ini ditunjukkan dengan nilai F hitung mempunyai nilai yang hampir berimbang, yakni sekitar 50-an. Faktanya, provinsi-provinsi pada cluster 1 menunjukkan jumlah penduduk yang cukup tinggi, hal ini mendorong munculnya aktivitas yang tinggi pula. Tidak menutup kemungkinan jika masing-masing warga tidak menerapkan konsep 5M (Mencuci tangan, Memakai masker, Menjaga jarak, Menghindari kerumunan, Mengurangi mobilitas), akibatnya kasus positif pada cluster 1 meningkat.

#### 4. SIMPULAN DAN SARAN

Berdasarkan analisis pembahasan, dapat disimpulkan bahwa pengelompokan dalam *K-Means Clustering* yang terbaik adalah 3 *Cluster*, hal ini didasarkan pada evaluasi DBI yang memberikan nilai DBI untuk 3 *Cluster* lebih kecil dibandingkan 2 *Cluster*, dan menunjukkan bahwa optimasi *Clustering* sudah baik. Wilayah Indonesia yang terdiri dari 34 provinsi menurut kasus persebaran *Covid-19* lebih tepat diklasifikasikan ke dalam tiga *Cluster*. Provinsi dengan jumlah penduduk tertinggi, jumlah penduduk per Km<sup>2</sup> tertinggi, jumlah penduduk lansia terbanyak, dan fasilitas kesehatan terbanyak terdapat pada *Cluster* 1. Dan pada *Cluster* 1 ini juga, jumlah pasien positif, jumlah pasien sembuh, dan jumlah pasien meninggal juga tertinggi. Sehingga dapat disimpulkan *variable*-variabel yang paling mendominasi dalam analisis cluster ini adalah jumlah penduduk, jumlah penduduk lansia, dan sarana kesehatan. Sebagai saran untuk penelitian berikutnya diharapkan melibatkan *variable*-variabel lain yang lebih spesifik dengan periode data terbaru. Sehingga diharapkan dapat memberi masukan lebih berarti lagi kepada pemerintah dalam upaya mengurangi jumlah kasus *Covid-19* di berbagai provinsi.

## DAFTAR PUSTAKA

- A.Gore,Jr, P. (2000) Handbook of Applied Multivariate Statistics and Mathematical Modeling. ACADEMIC PRESS. doi: <https://doi.org/10.1016/B978-012691360-6/50012-4>.
- Aabed, K. and Lashin, M. M. A. (2021) 'An analytical study of the factors that influence COVID- 19 spread', Saudi Journal of Biological Sciences, pp. 1177–1195. doi: 10.1016/j.sjbs.2020.11.067.
- Andayani, S. (2016) 'K-Pembentukan *Cluster* dalam Knowledge Discovery in Database dengan K-Pembentukan *Cluster* dalam Knowledge Discovery in Database dengan Algoritma *K-Means* Kata kunci : *Cluster* , Knowledge Discovery in Database , Algoritma *K-Means* , Pendahuluan Dewasa ini p', (June).
- Bates, A., & Kalita, J. (2016). Counting *Clusters* in Twitter Posts. Proceedings of the 2nd International Conference on Information Technology for Competitive Strategies, (p. 85).
- Bholowalia, P., & Kumar, A. (2014). EBK-Means : A *Clustering* Techniques based on Elbow Method and *K-Means* in WSN. International Journal of Computer Application (0975-8887), IX(105), 17 - 24.
- Dowd, J. B. et al. (2020) 'Demographic science aids in understanding the spread and fatality rates of COVID-19', Leverhulme Centre for Demographic Science, Nuffield College, University of Oxford, Oxford OX1 3UQ, United Kingdom. doi: <https://doi.org/10.1073/pnas.200491117>.
- E. Rendón, I. A. (2011). Internal versus External *Cluster* validation indexes. 5(1).
- Febrianti, A. F., Cabral, A. H. and Anuraga, G. (2018) '*K-Means Clustering* Dengan Metode Elbow Untuk Pengelompokan Kabupaten Dan Kota Di Jawa Timur', Seminar Nasional Hasil Riset dan Pengabdian -SNHRP, pp. 863–870. Available at: <http://karyailmiah.unipasby.ac.id/wp-content/uploads/2019/04/K-Means-Artikel.pdf>.
- Ganasegeran, K. et al. (2021) 'Influence of population density for Covid-19 spread in malaysia: An ecological study', International Journal of Environmental Research and Public Health, 18(18). doi: 10.3390/ijerph18189866.
- Ghozali, I (2013) Aplikasi Analisis Multivariate dengan Program IBM SPSS 21 Update PLS Regresi | Perpustakaan Mikroskil. Semarang: UNDIP. Available at: [https://mikroskil.ac.id/pustaka/index.php?p=show\\_detail&id=7026&keywords](https://mikroskil.ac.id/pustaka/index.php?p=show_detail&id=7026&keywords).
- Ghozali, I. (2018) Aplikasi Analisis Multivariate Dengan Program IBM SPSS 25. 9th edn. Semarang: Badan Penerbit Universitas Diponegoro.
- Irwanto, e. a. (2012). Optimasi Kinerja Algoritma Klasterisasi *K-Means* untuk Kuantisasi Warna Citra. Jurnal Teknik ITS, 1(1), 197-202.
- Izzuddin, A. (2015) 'Optimasi *Cluster* pada Algoritma *K-Means* dengan Reduksi Dimensi Dataset Menggunakan Principal Component Analysis untuk Pemetaan Kinerja Dosen', Edisi Nopember, 5(2), pp. 41–46.
- Maulik, U., & Bandyopadhyay, S. (2002). Performance Evaluation of Some *Clustering* Algorithm Validity Indices. IEEE Transaction On Pattern Analysis and Machine Intelligence, 24(12), 1650-1654
- Madhulata, T. (2012, April). An Overview on *Clustering* Methods. IOSR Juornal of Engineering, 2(4), 719-725.
- Saputro, a a and helilintar, r (2020) 'Perancangan Prediksi Prestasi Nilai Akademik Mahasiswa Menggunakan Metode *K-Means Clustering*', Prosiding Semnas Inotek .... Available at: <https://proceeding.unpkediri.ac.id/index.php/inotek/article/view/172>.
- Sari, D. N. P. and Sukestiyarno, Y. L. (2021) 'Analisis *Cluster* dengan Metode *K-Means* pada Persebaran Kasus Covid-19 Berdasarkan Provinsi di Indonesia', PRISMA, Prosiding Seminar Nasional Matematika, 4, pp. 602–610. Available at: <https://journal.unnes.ac.id/sju/index.php/prisma/>.
- Talakua, M. W., Leleury, Z. A. and Taluta, A. W. (2017) 'Analisis *Cluster* Dengan Menggunakan Metode *K-Means* Untuk Pengelompokan Kabupaten/Kota Di Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014', BAREKENG: Jurnal Ilmu Matematika dan Terapan, 11(2), pp. 119–128. doi: 10.30598/barekengvol11iss2pp119-128.
- Yulianto, S. and Hidayatullah, K. H. (2016) 'Analisis Klaster Untuk Pengelompokan Kabupaten/Kota Di Provinsi Jawa Tengah Berdasarkan Indikator Kesejahteraan Rakyat', Statistika, 2(1), pp. 56–63. Available at: <https://jurnal.unimus.ac.id/index.php/statistik/article/view/1115>.
- Wani, M., & Riyaz, R. (2017). A Novel Point Density based Validity Index for *Clustering* Gene Expression Datasets. International Journal of Data Mining and Bioinformatics, 17, pp. 66-84

Wahyudi, I., Sulthan, M., & Suhartini, L. (2021). Analisa Penentuan *Cluster* Terbaik pada Metode *K-Means* Menggunakan Ebow terhadap Sentra Industri Produski Di Pamekasan. *Aplikasi Teknologi Informasi dan Manajemen (JATIM)*, 72 -81.