PRISMA 4 (2021): 602-610
PRISMA, Prosiding Seminar Nasional Matematika
https://journal.unnes.ac.id/sju/index.php/prisma/
ISSN 2613-9189

Analisis *Cluster* dengan Metode *K-Means* pada Persebaran Kasus *COVID-19* Berdasarkan Provinsi di Indonesia

Desy Noor Permata Sari[a,*], YL. Sukestiyarno[a]

[a] *Universitas Negeri Semarang, Sekaran, Gunungpati, Semarang 50229, Indonesia*

* *Alamat Surel: desy.permata@students.unnes.ac.id*

Abstrak

Tujuan dalam penelitian menggunakan metode *K-Means Cluster* adalah untuk mengetahui tingkat persebaran kasus COVID-19 kategori tinggi, sedang, dan rendah pada masing-masing provinsi di Indonesia. Ada beberapa aspek yang bisa diukur seperti jumlah penduduk, kepadatan penduduk, kasus positif terinfeksi COVID-19, pasien yang sembuh, dan pasien yang meninggal dunia. Metode pengumpulan data yang digunakan adalah metode dokumentasi berupa data sekunder yang diperoleh dari publikasi buku Badan Pusat Statistik dan data Kemenkes RI di Badan Nasional Penanggulangan Bencana. Data yang digunakan yaitu jumlah penduduk (X1), kepadatan penduduk (X2), positif (X3), sembuh (X4), dan meninggal (X5) dan dianalisis menggunakan software SPSS. Dari hasil penelitian dengan metode *K-Means* terbentuk menjadi 5 *cluster*. *Cluster* 1 termasuk kasus yang tinggi berisi 2 provinsi. *Cluster* 2 termasuk kasus yang sedang berisi 3 provinsi. *Cluster* 3 termasuk kasus yang rendah berisi 29 provinsi dan dibagi lagi menjadi 3 *cluster* dengan mengelompokkan berdasarkan tingakatannya. Karakteristik *cluster* 1 kategori tinggi berisi rata-rata X2, X3, X4, dan X5 berada di atas rata-rata populasi. *Cluster* 2 kategori sedang berisi rata-rata X2, X3, X4, dan X5 berada di bawah rata-rata populasi. *Cluster* 3 kategori rendah berisi rata-rata semua variabel berada di atas rata-rata populasi. Variabel yang memberikan perbedaan paling besar adalah variabel kepadatan penduduk dengan nilai F sebesar 26,641 dan nilai signifikan 0,000. Provinsi yang memiliki nilai paling besar pada variabel kepadatan penduduk adalah Provinsi DKI Jakarta.

Kata kunci:

K-Means, COVID-19, cluster.

1. Pendahuluan

Peran *Data Science* di *Industri 4.0* untuk tahun 2020 sangat dibutuhkan dan memberikan dampak yang sangat kuat dalam kehidupan ekosistem dunia. *Industri 4.0* merupakan sebuah revolusi dari inovasi di bidang elektronik dan teknologi informasi. Konsep *Industri 4.0* adalah realitas baru dari ekonomi *modern*, karena inovasi dan pengembangan teknologi memainkan peran penting dalam setiap organisasi (Ślusarczyk, 2018).

Artificial Intelligence (AI) adalah domain sains dan teknik yang berkaitan dengan teori dan praktik pengembangan sistem yang menunjukkan karakteristik yang kami kaitkan dengan kecerdasan dalam perilaku manusia, seperti persepsi, pemrosesan bahasa alami, pemecahan masalah dan perencanaan, pembelajaran dan adaptasi, dan bertindak atas lingkungan (Tecuci, 2012). *Machine Learning* adalah bagian dalam *Artificial Intelligence* dengan kemampuan mesin untuk mengakses data yang ada dengan perintah mereka sendiri. Tugas-tugas yang dilaksanakan oleh *Machine Learning* berbagai macam. *Machine Learning* mempunyai 3 macam teknik pada umumnya, yaitu (1) *Supervised Learning* yang mempelajari pada data yang ada dan mencari pola dari data set yang akan dijadikan sebagai acuan untuk data-data selanjutnya, (2) *Reinforcement Learning* yang menempatkan agent ke sebuah *environment* yang tidak diketahui tempatnya dan akan mengeksplorasi keseluruhan sampai menemukan istilah reward dan *error*, dan (3) *Unsupervised Learning* bersifat deskriptif dan berguna untuk mengkategorikan atau mengelompokkan data (Rauhan, 2019).

To cite this article:
Sari, D. N. P., & Sukestiyarno, YL. (2021). Analisis *Cluster* dengan Metode K-Means pada Persebaran Kasus *Covid-19* Berdasarkan Provinsi di Indonesia. *PRISMA, Prosiding Seminar Nasional Matematika 4*, 602-610

Analisis *cluster* merupakan teknik multivariat yang mempunyai tujuan utama untuk mengelompokkan objek-objek berdasarkan karakteristik yang dimilikinya (Awalluddin & Taufik, 2017). Analisis *cluster* mengklasifikasi objek sehingga objek-objek yang paling dekat kesamaanya dengan objek lain berada dalam *cluster* yang sama (Ediyanto *et al*., 2013). *K-Means Cluster Analysis* merupakan salah satu metode *cluster analysis* 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 dikelompokan dalam satu *cluster* yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan ke dalam *cluster* yang lain. Metode *K-Means* berusaha mengelompokkan data yang ada kedalam satu kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang berbeda dengan data yang ada didalam kelompok yang lain (Helilintar & Farida, 2018). Dasar algoritma *K-Means* adalah sebagai berikut:

- Tentukan nilai k sebagi jumlah *cluster* yang ingin dibentuk.
- Inisialisasi k sebagai centroid yang dapat dibangkitkan secara random.
  $d(i,k) = \sqrt{\sum_i^m (C_i j - C_k j)^2}$
- Hitung jarak setiap data ke masing-masing centroid menggunakan persamaan *Eulidiean Distance* yaitu sebagai berikut:
  $min \sum_k^i - a_{ik} - = \sqrt{\sum_i^m (C_i j - C_k j)^2}$
- Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan centroidnya.
  $C_{kj} = \frac{\sum_k^i x_{ij}}{p}$
- Tentukan posisi centroid baru (k)
- Kembali ke langkah 3 jika posisi centroid baru dengan centroid lama tidak sama.

Menurut (Santoso, 2014), *K-Means Cluster*ing memiliki asumsi-asumsi yang harus dipenuhi, yaitu:

- Sampel yang diambil harus benar-benar bisa mewakili populasi yang ada.
- *Multikolinearitas*, yakni kemungkinan adanya korelasi antar objek. Namun sebaiknya tidak ada, bila ada besar *Multikolinearitas* tersebut tidaklah tinggi (misal di atas 0,5).

Penyebaran virus *COVID-19* saat ini semakin luas, bahkan ke negara-negara lain yang terletak jauh, seperti Italia dan Iran. Hingga saat ini pun, vaksin untuk menyembuhkan *COVID-19* masih dalam penelitian, dimana para peneliti terus berpacu dengan waktu dikarenakan makin banyaknya orang yang terinfeksi atau bahkan mengalami kematian diakibatkan keganasan virus ini. Untuk mengetahui tingkat persebaran kasus *COVID-19* kategori tinggi, sedang, dan rendah pada masing-masing Provinsi di Indonesia. Ada beberapa aspek yang bisa diukur seperti jumlah penduduk, kepadatan penduduk, kasus positif terinfeksi *COVID-19*, pasien yang sembuh, dan pasien yang meninggal dunia.

Kuantitas atau jumlah penduduk dapat sebagai potensi maupun menjadi beban bagi suatu negara, akan menjadi potensi apabila jumlah penduduk seimbang dengan sumber daya yang lain serta mempunyai kualitas hidup yang baik. Sebaliknya, menjadi beban apabila jumlah penduduk melampaui kapasitas wilayah Negara tersebut (Christiani *et al.*, 2013).

Kepadatan penduduk adalah perbandingan antara jumlah penduduk dengan luas wilayah yang dihuni (Mantra, 2011). Kepadatan penduduk dapat mempengaruhi kualitas hidup penduduknya. Pada daerah dengan kepadatan yang tinggi, usaha peningkatan kualitas penduduk akan lebih sulit dilakukan (Christiani *et al*., 2013). Daerah yang memiliki kepadatan penduduk yang tinggi akan menyebabkan transmisi penyakit menular lebih cepat dengan rantai penyebaran yang lebih kompak dan kompleks.

Gejala-gejala orang yang terinfeksi atau positif *COVID-19*, yaitu demam ≥38℃, batuk kering, sesak napas, nyeri tenggorokan, dan pegal-pegal atau merasa kelelahan (Kemenkes RI, 2020). *COVID-19* dapat menyebar melalui percikan air (droplet) dari hidung atau mulut pada saat berbicara, batuk atau bersin. Jika ada orang lain menyentuh benda yang sudah terkontaminasi dengan droplet tersebut, lalu orang itu menyentuh mata, hidung atau mulut (segitiga wajah), maka orang itu dapat terinfeksi virus Corona.

Pasien yang terinfeksi *COVID-19* dapat sembuh jika menerima perawatan yang tepat untuk meredakan dan mengobati gejala, dan mereka yang sakit serius (gejala sedang dan berat) harus dibawa ke rumah sakit. Sebagian besar pasien sembuh karena perawatan untuk gejala yang dialami (Kemenkes RI, 2020).

Usia lanjut dan penyakit komorbid seperti diabetes dan hipertensi dilaporkan menjadi faktor risiko kematian pada orang-orang yang terjangkit *COVID-19*. Karena itu, lansia berisiko paling tinggi meninggal dan menjadi salah satu kelompok yang paling rentan. Penting untuk diketahui bahwa lansia memiliki hak

yang sama dengan orang lain untuk menerima perawatan kesehatan berkualitas tinggi, termasuk perawatan intensif. Pasien-pasien ini bisa saja menunjukkan gejala-gejala ringan tetapi keadaannya berisiko tinggi menjadi lebih buruk dan perlu dirawat di unit yang ditunjuk agar dapat dimonitor dengan baik (World Health Organization, 2020).

Dalam penelitian ini akan dianalisis kasus pengelompokkan *K-Means Cluster* 34 Provinsi di Indonesia berdasarkan tingkat persebaran kasus *COVID-19*. *K-Means Cluster* adalah salah satu metode dan *cluster*ing non-hierarki yang berusaha mengelompokkan data ke dalam suatu *cluster* sehingga data yang memiliki karakteristik sama dikelompokkan ke dalam satu *cluster* yang sama. Berdasarkan tingkat persebaran kasus *COVID-19* akan dilakukan *cluster*ing pada provinsi di Indonesia pada bulan 31 Mei 2020 dengan metode *K-Means Cluster*.

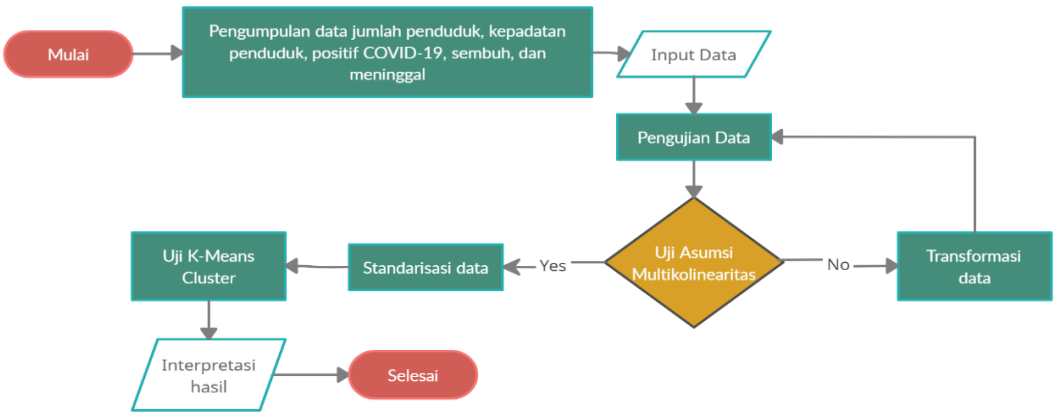
Rumusan masalah yang dikaji dalam penelitian ini adalah (1) bagaimana penerapan metode *K-Means Cluster* untuk mengelompokkan provinsi, (2) karakteristik *cluster* yang terbentuk berdasarkan tingkat persebaran kasus *COVID-19* di Indonesia pada bulan 31 Mei 2020, dan (3) variabel apa yang memberikan perbedaan paling besar pada *cluster* yang terbentuk dalam persebaran kasus *COVID-19* dan Provinsi mana yang memiliki nilai paling besar pada variabel ini.

Tujuan dalam penelitian ini adalah (1) menerapkan metode *K-Means Cluster* untuk mengelompokkan provinsi dan mengetahui karakteristik *cluster* yang terbentuk berdasarkan tingkat persebaran kasus *COVID-19* di Indonesia tahun 2020, (2) mengetahui Provinsi yang termasuk dalam *CLUSTER* tingkat persebaran kasus *COVID-19* tinggi, sedang, dan rendah, dan (3) mengetahui variabel yang memberikan perbedaan paling besar pada *cluster* yang terbentuk dan Provinsi mana yang memiliki jumlah terbanyak dalam variabel tersebut.

2. Metode

Metode pengumpulan data yang digunakan dalam penyusunan Tugas Akhir ini adalah metode dokumentasi berupa data sekunder. Metode dokumentasi digunakan untuk memperoleh data dan informasi melalui pengumpulan data yang diperoleh dari publikasi buku Badan Pusat Statistik dan data Kemenkes RI di Badan Nasional Penanggulangan Bencana. Dalam hal ini data yang digunakan dari publikasi buku Badan Pusat Statistik adalah data berupa jumlah penduduk dan kepadatan penduduk yang dikumpulkan dari katalog yang berjudul Statistik Indonesia 2020 (BPS-*Statistics* Indonesia, 2020). Sedangkan data yang digunakan dari Kemenkes RI di Badan Nasional Penanggulangan Bencana adalah data Laporan Media Harian *COVID-19* tanggal 31 Mei 2020 yang terdiri dari 3 variabel yaitu yaitu kasus positif terinfeksi *COVID-19*; pasien yang sembuh; dan pasien yang meninggal dunia (BNPB, 2020).

Analisis data pengelompokkan dalam penelitian ini menggunakan metode *K-Means Cluster* dengan jumlah *cluster* yang diinginkan 3 *cluster* dengan bantuan *software* SPSS 20. Alur penelitian dari analisis data dalam penelitian ini adalah sebagai berikut.



Gambar 1. langkah-langkah analisis *K-Means Cluster*

3. Hasil dan Pembahasan

Tabel 1. Hasil uji multikolinieritas

Model	*Collinearity Statistics*	
	Tolerance	**VIF**
Jumlah Penduduk	.434	2.304
Kepadatan Penduduk	.085	11.813
Positif	.008	129.797
Sembuh	.033	30.713
Meninggal	.013	76.156

Uji multikolinearitas dilakukan untuk menguji ada atau tidaknya variabel independen yang mempunyai kemiripan antar variabel independen lain (Yulianto & Hidayatullah, 2014). Jika data menunjukkan adanya multikolinearitas, maka salah satu cara yang dapat dilakukan yaitu melakukan transformasi data ke dalam bentuk logaritma natural (Ghozali, 2018). Cara untuk menguji adanya multikolinieritas, yaitu:

- Melihat nilai *tolerance*
  Terjadi multikolinieritas, jika nilai *tolerance* $\leq 0{,}1$.
- Melihat nilai VIF (*Variance Inflation Factor*)
  Terjadi multikolinieritas, jika nilai VIF ≥ 10.

Dari hasil pengujian multikolinieritas. Didapatkan bahwa hanya variabel Jumlah Penduduk memiliki nilai *tolerance* lebih besar dari 0,1 serta nilai VIF nya kurang dari 10. Sedangkan pada variabel lain memiliki nilai *tolerance* lebih kecil dari 0,1 serta nilai VIF nya lebih besar dari 10. Maka dapat disimpulkan bahwa pada semua variabel selain Jumlah Penduduk terjadi multikolinieritas. Karena data terjadi multikolinearitas, maka dilakukan transformasi data ke dalam bentuk logaritma natural.

Tabel 2. Hasil uji multikolinieritas dengan transformasi data

Model	*Collinearity Statistics*	
	Tolerance	**VIF**
LN_X1	.852	1.174
LN_X2	.870	1.149
LN_X3	.347	2.884
LN_X4	.368	2.721
LN_X5	.441	2.267

Pada Tabel 2, didapatkan bahwa semua variabel memiliki nilai tolerance lebih besar dari 0,1 serta nilai VIF nya kurang dari 10. Jadi semua variabel dengan data bentuk logaritma natural tidak terjadi multikolinearitas.

Untuk membentuk *cluster* yang diinginkan, maka data dilakukan standarisasi untuk menghasilkan nilai Zscore yang dihasilkan dalam uji ini. Zscore yang didapatkan dari Tabel 3 nantinya untuk mengelompokkan *cluster* dan dapat menganalisis *cluster* yang terbentuk.

Tabel 3. Statistik deskriptif

	N	**Min**	**Max**	**Mean**	**Std. Dev**
Jumlah Penduduk	34	743	49317	7884.58	11202.153
Kepadatan Penduduk	34	9.3	15900.0	741.059	2708.7090
Positif	34	20	7348	778.00	1470.436
Sembuh	34	14	2082	214.94	371.196
Meninggal	34	0	517	47.44	109.530
Valid N (*listwise*)	34				

Tabel 4. *Initial cluster centers*

	Cluster		
	1	**2**	**3**
LN_X1	-2.87	-2.30	-5.90
LN_X2	3.44	-11.36	-2.86
LN_X3	2.99	-3.09	-3.95
LN_X4	3.23	-2.36	-6.87
LN_X5	2.91	-1.85	-4.20

Pada Tabel 4 merupakan proses *cluster*ing data pertama sebelum data tersebut dilakukan iterasi dan data ini adalah proses untuk pembentukan 3 *cluster*.

Tabel 5. *Iteration history*

Iteration	***Change in Cluster Centers***		
	1	**2**	**3**
1	4.946	5.432	6.270
2	.000	.000	.000

Tabel 5 merupakan proses iterasi dalam pengelompokan *cluster* dari tabel initial dan menghasilkan proses iterasi sebanyak 2 kali. Pada iterasi 1 terjadi centeroid yang tidak signifikan dan pada iterasi 2 terjadi centeroid yang signifikan. Jadi, semua *cluster* sudah terbentuk dan iterasi berhenti pada iterasi 2 dengan jarak minimum 10,573.

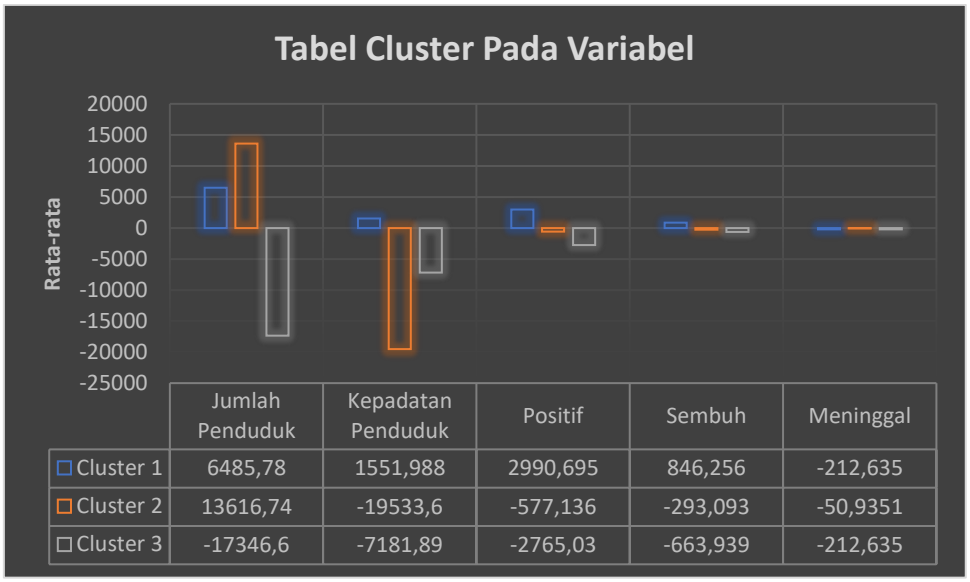
Tabel 6. *Final cluster centers*

	Cluster		
	1	**2**	**3**
LN_X1	-.12	.51	-2.25
LN_X2	.30	-7.48	-2.92
LN_X3	1.50	-.92	-2.41
LN_X4	1.70	-1.37	-2.37
LN_X5	1.33	-.90	-2.37

Pada Tabel 6 adalah hasil dari proses akhir dalam *cluster*ing yang membentuk *cluster* sebanyak 3 untuk masing-masing variabel. Variabel pada tabel *final cluster centers* merupakan hasil untuk nilai standarisasi. Angka negatif memiliki arti bahwa data berada di bawah rata-rata total dan angka positif memiliki arti bahwa data berada di atas rata-rata total. Untuk mengetahui pengaruh pada variabel di atas dilakukan perhitungan dengan rumus:

$$X = \mu + z.\sigma$$

Keterangan:
X = rata-rata data atau sampel
μ = rata-rata populasi
σ = standar deviasi
z = nilai standarisasi



Gambar 2. rata-rata *cluster* pada masing-masing variabel

Berdasarkan perhitungan yang dihasilkan dari masing-masing variabel, didapatkan karakteristik pada masing-masing *cluster*. berikut adalah penjelasan untuk setiap *cluster*:

- *Cluster* 1 berisi variabel jumlah penduduk berada di bawah rata-rata populasi. Sedangkan variabel kepadatan penduduk, positif, sembuh, dan meninggal berada di atas rata-rata populasi.
- *Cluster* 2 berisi variabel jumlah penduduk yang berada di atas rata-rata populasi. Sedangkan variabel kepadatan penduduk, positif, sembuh, dan meninggal berada di bawah rata-rata populasi.
- *Cluster* 3 berisi variabel jumlah penduduk, kepadatan penduduk, positif, sembuh, dan meninggal berada di bawah rata-rata populasi.

Untuk melihat pengujian tingkat signifikansi antar *cluster* dan mengetahui perbedaan di setiap *cluster*, maka perlu dilakukan uji ANOVA. Ketentuan penggunaan angka F dalam analisis *cluster* ialah bahwa semakin besar angka F hitung (jika dilakukan uji Hipotesis, maka F hitung akan lebih besar dari F tabel) dan tingkat signifikansi (sig) < 0,05; maka semakin besar perbedaan antara ketiga *cluster* yang terbentuk (Bastian et al., 2018). Berikut adalah hasil dari uji ANOVA:

Tabel 7. ANOVA

ANOVA						
	Cluster		***Error***		**F**	**Sig.**
	Mean Square	**df**	***Mean Square***	**df**		
LN_X1	13.671	2	2.815	31	4.857	.015
LN_X2	40.818	2	1.532	31	26.641	.000
LN_X3	16.421	2	2.096	31	7.835	.002
LN_X4	16.222	2	2.157	31	7.522	.002
LN_X5	14.929	2	1.553	31	9.611	.001

Berdasarkan Tabel 7, pada kolom *cluster* adalah besaran *between cluster means* dan kolom *error* adalah besaran *within cluster means*. Sementara pada kolom F didapat dari rumus sebagai berikut.

$$F = \frac{between\ cluster\ mean}{within\ cluster\ mean}$$

Hipotesis:

H_0: Ketiga *cluster* tidak mempunyai perbedaan yang signifikan.

H_1: Ketiga *cluster* mempunyai perbedaan yang signifikan.

Jika angka signifikan > 0.05; H_0 diterima H_1 ditolak.

Jika angka signifikan < 0,05; H_0 ditolak H_1 diterima.

Berdasarkan hasil yang didapatkan bahwa semua variabel memiliki nilai signifikansi (sig.) < 0,05, maka ketiga *cluster* mempunyai perbedaan yang signifikan dan diperoleh hasil bahwa variabel kepadatan penduduk adalah variabel yang paling membedakan anggota dari ketiga *cluster* karena mempunyai nilai F terbesar diantara variabel.

Tabel 8. *Cluster membership*

Case Number	**Provinsi**	***Cluster***	***Distance***
1	ACEH	3	1.841
2	BALI	2	5.432
3	BANTEN	3	4.449
4	BANGKA BELITUNG	3	1.939
5	BENGKULU	3	1.796
6	DI YOGYAKARTA	3	1.673
7	DKI JAKARTA	1	4.946
8	JAMBI	3	1.626
9	JAWA BARAT	1	4.946
10	JAWA TENGAH	2	4.211
11	JAWA TIMUR	2	4.925
12	KALIMANTAN BARAT	3	1.196
13	KALIMANTAN TIMUR	3	2.070
14	KALIMANTAN TENGAH	3	1.840
15	KALIMANTAN SELATAN	3	2.281
16	KALIMANTAN UTARA	3	1.716
17	KEPULAUAN RIAU	3	1.197
18	NUSA TENGGARA BARAT	3	2.517
19	SUMATERA SELATAN	3	6.270
20	SUMATERA BARAT	3	2.475
21	SULAWESI UTARA	3	2.572
22	SUMATERA UTARA	3	3.679
23	SULAWESI TENGGARA	3	1.051
24	SULAWESI SELATAN	3	3.789
25	SULAWESI TENGAH	3	1.269
26	LAMPUNG	3	3.869
27	RIAU	3	2.921
28	MALUKU UTARA	3	1.740
29	MALUKU	3	1.504
30	PAPUA BARAT	3	1.740
31	PAPUA	3	3.030
32	SULAWESI BARAT	3	1.804
33	NUSA TENGGARA TIMUR	3	1.779
34	GORONTALO	3	1.902

Tabel 8 merupakan hasil akhir *cluster* yang telah dilakukan, maka dari tabel tersebut dapat disimpulkan sebagai berikut:

Tabel 9. Provinsi pada setiap *cluster*

	***Cluster* 1**	***Cluster* 2**	***Cluster* 3**
Provinsi	DKI Jakarta dan Jawa Barat	Bali, Jawa Tengah, Jawa Timur	Aceh, Banten, Bangka Belitung, Bengkulu, DI Yogyakarta, Jambi, Kalimantan Barat, Kalimantan Timur, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Utara, Kepulauan Riau, Nusa Tenggara Barat, Sumatera Selatan, Sumatera Barat, Sulawesi Utara, Sumatera Utara, Sulawesi Tenggara, Sulawesi Selatan, Sulawesi Tengah, Lampung, Riau, Maluku Utara, Maluku, Papua Barat, Papua, Sulawesi Barat, Nusa Tenggara Timur, dan Gorontalo

Berdasarkan Tabel 8 pada kolom *distance*, didapatkan data provinsi dengan kasus COVID-19 tertinggi pada masing-masing *cluster* yaitu sebagai berikut:

Gambar 3. provinsi yang memiliki kasus tertinggi pada setiap *cluster*

4. Simpulan

Berdasarkan pembahasan yang telah didapatkan, maka dapat diambil kesimpulan bahwa penerapan dengan metode *K-Means Cluster* digunakan untuk mengelompokkan 34 provinsi di Indonesia berdasarkan tingkat persebaran kasus COVID-19 pada bulan 31 Mei 2020. Sebelum analisis *K-Means Cluster*, dilakukan uji asumsi multikolinieritas terlebih dahulu. Setelah memenuhi asumsi, dilanjutkan analisis *K-Means Cluster*. Analisis dengan metode *K-Means Cluster* dalam pengelompokkan 34 provinsi di Indonesia berdasarkan tingkat persebaran kasus COVID-19 dibentuk menjadi 3 *cluster*, yaitu *cluster* 1 berisi 2 provinsi, *cluster* 2 berisi 3 provinsi, dan *cluster* 3 berisi 29 provinsi. Karakteristik *cluster* yang terbentuk adalah pada *cluster* 1 berisi kelompok provinsi yang memiliki jumlah kepadatan penduduk, positif, sembuh, dan meninggal memiliki nilai rata-rata yang lebih tinggi dibandingkan dengan *cluster* 2 dan 3, *cluster* 2 berisi kelompok provinsi yang memiliki jumlah penduduk memiliki nilai rata-rata yang lebih tinggi dan kepadatan penduduk yang lebih rendah dibandingkan dengan *cluster* 1 dan 3, dan *cluster* 3 berisi kelompok provinsi yang memiliki jumlah penduduk memiliki nilai rata-rata yang lebih rendah dibandingkan dengan *cluster* 1 dan

2. Variabel yang memberikan perbedaan paling besar pada *cluster* yang terbentuk dalam persebaran kasus COVID-19 adalah variabel kepadatan penduduk dan provinsi yang memiliki nilai paling besar pada variabel kepadatan penduduk adalah Provinsi DKI Jakarta.

Adapun beberapa saran yang dapat digunakan untuk penelitian selanjutnya, seperti mendapatkan lebih banyak data dari kasus COVID-19 sehingga dapat menganalisis pengaruh jumlah kasus COVID-19 terhadap akurasi klasifikasi penyebab terjadinya meningkatnya kasus COVID-19 di setiap provinsi dan aplikasi yang digunakan dalam analisis *K-Means Clustering* bisa menggunakan aplikasi yang lebih modern dan lebih canggih untuk bisa menciptakan visualisasi dari *clustering* berupa gambar titik map di setiap provinsi Indonesia.

Daftar Pustaka

Awalluddin, A. S., & Taufik, I. (2017). Analisis *Cluster* Data Longitudinal pada Pengelompokan Daerah Berdasarkan Indikator IPM di Jawa Barat. *978, 187–194*.

BNPB. (2020). Laporan Media Harian Covid19 Tanggal 9 April 2020 Pukul 12.00 Wib. 2020, 1, 6575. *https://loker.bnpb.go.id/s/GugusTugasCovid19?path=%2FData Kemenkes#pdfviewer*

BPS-Statistics Indonesia. (2020). Statistik Indonesia 2020 Statistical Yearbook of Indonesia 2020. *Statistical Yearbook of Indonesia, April. https://doi.org/10.3389/fpsyg.2015.00002*.

Christiani, C., Tedjo, P., & Martono, B. (2013). Analisis Dampak Kepadatan Penduduk Terhadap Kualitas Hidup Masyarakat Provinsi Jawa Tengah. *102–114*.

Ediyanto, Novitasari Mara, M., & Satyahadewi, N. (2013). Pengklasifikasian Karakteristik Dengan Metode *K-Means Cluster* Analysis. *Buletin Ilmiah Mat. Stat. Dan Terapannya (Bimaster), 02(2), 133–136*.

Ghozali, I. (2018). Aplikasi Analisis Multivariate Dengan Program IBM SPSS 25 Edisi 9. Semarang: *Badan Penerbit Universitas Diponegoro*.

Helilintar, R., & Farida, I. N. (2018). Penerapan Algoritma *K-Means Cluster*ing Untuk Prediksi Prestasi Nilai Akademik Mahasiswa. *Jurnal Sains Dan Informatika, 4(2), 80.*

Kemenkes RI. (2020). Tanya Jawab Seputar Virus Corona. *119–135*.

Mantra, I. B. (2011). Demografi Umum. Yogyakarta: *Pustaka Belajar*.

Rauhan, A. (2019). Pengolahan Data Menggunakan Machine Learning. *021, 2–4.*

Ślusarczyk, B. (2018). Industry 4.0–Are we ready? *Polish Journal of Management Studies, 17(1), 232–248.*

Santoso, S. (2014). Mahir Statistik Multivariat Dengan SPSS. Jakarta: PT Elek Media Komputindo.

Tecuci, G. (2012). Artificial intelligence. Wiley Interdisciplinary Reviews: *Computational Statistics, 4(2), 168–180.*

World Health Organization. (2020). Tatalaksana klinis infeksi saluran pernapasan akut berat (SARI) suspek penyakit *COVID-19. World Health Organization, 4(March), 1–25*.

Yulianto, S., & Hidayatullah, K. H. (2014). Analisis Klaster Untuk Pengelompokan Kabupaten/Kota Di Provinsi Jawa Tengah Berdasarkan Indikator Kesejahteraan Rakyat. Statistika, 2(1), 56–63.