Klasterisasi provinsi di Indonesia berbasis perkembangan kasus Covid-19 menggunakan metode K-Medoids

Indra Gunawan*, Galuh Anggraeni, Endang Sulistiyo Rini, Yunanda Mustofa Putri, Yuda Khoirul Zikri
Program Studi Statistika, Fakultas MIPA, Universitas Islam Indonesia

*Penulis Korespondensi: 17611064@students.uii.ac.id

Abstract. COVID-19 is a disease caused by a new type of corona virus Sars-CoV-2 which is affecting almost all countries in the world. The country of Indonesia is one of the states of Southeast Asia which has a high spike in positive confirmed cases. With the growth in the number of positive confirmed patients, patients died, and patients recovered in Indonesia; then clustering is done using K-Medoids to see the development of the COVID-19 case in Indonesia. K-Medoids is a classic partitioning technique from clustering that classifies object n datasets into k clusters. By clustering using the K-Medoids method; then obtained provinces that have confirmed positive corona in Indonesia are divided into 3 groups. Members from group 1 which are in the high category are 2 provinces; members from group 2 in the medium category are 6 provinces; while for members of group 3 with a low category of 26 provinces.
Keywords: k-medoids; covid-19

1. Pendahuluan

Corona Virus Disease 2019 yang disingkat dengan COVID-19 merupakan penyakit yang baru ditemukan dan disebabkan oleh jenis corona virus baru yaitu Sars-CoV-2. Pandemi ini melanda hampir diseluruh negara. COVID-19 pertama dilaporkan muncul di Kota Wuhan, Tiongkok pada tanggal 31 Desember 2019. Penyebaran virus COVID-19 dapat berpindah langsung melalui tetesan kecil dari hidung atau melalui percikan batuk dan bersin yang terhirup oleh orang yang sehat. Selain itu, COVID-19 dapat berpindah secara tidak langsung dengan menyebarkan virusnya melalui benda-benda mati yang bertahan selama dua jam sampai enam hari akibat terpapar virus dari sentuhan tangan seseorang yang terpapar. Indonesia sendiri pertama kali mengumumkan pasien pertamanya pada tanggal 2 Maret 2020.

Indonesia menjadi salah satu negara bagian asia tenggara yang memiliki lonjakan tinggi pada kasus positif menurut data yang diterbitkan oleh Kementerian Kesehatan yang diperbaharui setiap harinya. Upaya pemerintah Indonesia dalam pengurangan kasus positif baru yaitu dengan memberlakukan kebijakan *stay at home* yang mana masyarakat dilarang untuk keluar rumah kecuali melakukan kegiatan penting seperti berbelanja kebutuhan pokok. Semakin berkembangnya jumlah pasien yang terkonfirmasi positif, pasien meninggal, dan pasien sembuh menarik peneliti untuk melakukan klasterisasi dengan menggunakan metode klasterisasi *K-Medoids* untuk melihat bagaimana perkembangan kasus COVID-19 di Indonesia. Pada penelitian kali ini, selain untuk melihat perkembangan kasus COVID-19, peneliti ingin mengetahui bagaimana hasil klasterisasi yang terbentuk dengan metode *K-Medoids*, serta bagaimana karakteristik *cluster* yang terbentuk dengan mengguankan metode *K-Medoids,* dan bagaimana hasil daerah yang terdampak COVID-19 dengan berdasarkan hasil clustering *K-Medoids*.

2. Metode

Dengan adanya pandemi yang sedang terjadi yaitu virus COVID-19 maka penelitian ini dilakukan secara *online*. Data yang digunakan adalah data COVID-19 di seluruh provinsi di Indonesia dengan variable pasien positif, pasien sembuh, dan pasien meninggal. Waktu pengambilan data dilakukan selama 4 bulan dari tanggal 02 Maret 2020 sampai dengan 30 Juni 2020. Populasi yang digunakan adalah seluruh masyarakat Indonesia. Data yang digunakan merupakan data sekunder yang diperoleh pada laman Portal GIS Gugus Tugas Percepatan Penanganan COVID-19 pada halaman website https://bnpb-inacovid19.hub.arcgis.com/ (BNPB, 2020).

2.1. Teknik Analisis Data

Algoritma *K-Medoids* sering disebut juga algoritma PAM (*Partitioning Around Medoid*) yang dikembangkan oleh Leonard Kaufman dan Peter J. Rousseeuw. Metode PAM merupakan algoritma yang menyerupai dengan *k-means.* Kedua algoritma ini *partitional* yang memecah dataset menjadi beberapa kelompok. Dikutip dari Kaur, dkk (2014), perbedaan dari algoritma *k-means* dengan algoritma *K-Medoids* terletak pada penentuan pusat *cluster*, dimana algoritma *k-means* menggunakan nilai rata–rata atau *means* dari setiap *cluster* sebagai pusat *cluster* dan algoritma *K-Medoids* menggunakan objek data sebagai perwakilan (*medoids*) sebagai pusat *cluster.* Algoritma *K-Medoids* digunakan untuk mengatasi kelemahan dari algoritma *k-means* yang sangat sensitive terhadap pencilan (*outlier*) karena objek–objek ini sangat jauh letaknya/karakteristiknya dari mayoritas data lainnya, sehingga jika dimasukkan ke suatu *cluster* data semacam ini bisa mendistorsi nilai rata-rata (*mean*) *cluster* tersebut.

Menurut Abhisek dan Purnima (2013), algoritma *K-Medoids* merupakan teknik partisi klasik dari *clustering* yang melakukan klasterisasi dataset objek *n* ke dalam *k cluster* yang dikenal sebagai *a priori.* Prinsip dari algoritma *K-Medoids* yaitu untuk meminimalkan jumlah kesamaan antara setiap objek dan titik referensi yang sesuai. Algoritma *K-Medoids* dapat dilakukan dengan langkah-langkah sebagai berikut (Bhat, 2014): (1) inisialisasi pusat *cluster* sebanyak *k* atau jumlah *cluster*, (2) hitung setiap objek ke *cluster* terdekat dengan persamaan *Euclidian Distance*, (3) setelah *Euclidian Distance* dihitung, inisialisasikan pusat *cluster* baru secara acak pada masing – masing objek sebagai kandidat *non medoids*, (4) hitung jarak setiap objek yang berada pada masing–masing *cluster* dengan kandidat *non medoids*, (5) hitung total simpangan (*S*) dengan menghitung total *distance* baru dikurang dengan total *distance* lama. Jika nilai $S < 0$ maka tukarkan objek tersebut dengan data *cluster non medoids* untuk membentuk beberapa kumpulan *k* objek baru sebagai *medoids*, serta (6) ulang perhitungan langkah c sampai dengan e sehingga *medoid* tidak mengalami perubahan, hingga didaptakan pengelompokan *cluster* beserta anggota *cluster*.

2.1.1. Metode within cluster sum of square (WCSS)

Metode *Compactness Separation Criterion* (CSC) adalah salah satu metode validitas yang ada pada *clustering* yang menggunakan kriteria minimal sebagai patokannya. CSC memiliki dua nilai yang yaitu *intra-cluster* dan *inter-cluster* yang mana dua nilai tersebut harus dihitung. Untuk mendapatkan nilai *intra-cluster* digunakan *Within Cluster Sum of Square atau WCSS* sebagai perhitungan (Maududie & Wibowo, 2014). Menurut Tallahasse (2020) *cluster cohessien* salah satu metode yang berada dalam *internal measure* serta berfungsi sebagai pengukur seberapa erat obyek dalam *cluster*. Untuk menghitung *Cluster cohessien* dapat menggunakan *Within Cluster Sum of Square* atau *WCSS.*

2.1.2. Metode silhoutte coefficient

Silhoutte Coefficient adalah metode yang biasa digunakan untuk melihat kualitas serta kekuatan kluster dan seberapa baik objek berada pada *cluster*. Metode *Silhoutte Coefficient* adalah metode gabungan dari dua metode yaitu *cohesion* dan *separation*. Menurut Wira, dkk. (2019), metode *cohesion* merupakan ukuran untuk mngetahui relasi antar objek didalam *cluster*. Sedangkan, metode *separation* merupakan ukuran untuk mngetahui seberapa jauh atau terpisah satu *cluster* dengan *cluster* yang lain.

2.2. Alat dan Bahan

Penelitian ini menggunakan data sekunder yang didapatkan pada laman Portal GIS Gugus Tugas Percepatan Penanganan COVID-19. Peneliti mengambil data perkembangan COVID-19 pada tiap provinsi dari tanggal 2 Maret 2020 sampai dengan tanggal 30 Juni 2020. Adapun populasi yang digunakan yaitu seluruh masyarakat di Indonesia. Untuk mengolah data tersebut, maka peneliti menggunakan bantuan alat analisis yang berupa *R Studio*, *Geoda*, *Statscan, QGis.*

3. Hasil dan Pembahasan

Sebelum melakukan *clustering* dilakukan analsis deskriptif untuk mengetahui gambaran data yang dibantu dengan menggunakan program R. Berikut adalah analsis deskriptifnya :

Tabel 1. Analisis Deskriptif Data Covid 19

Hasil	Variabel		
	Terkonfirmasi	Meninggal	Sembuh
Min.	86	0	31
1st Qu.	211.5	5.5	159.8
Median	640.5	16	270
Mean	1746.6	87.85	784.1
3rd Qu.	1677.5	81.25	862.5
Max	12695	948	6871

Berdasarkan Tabel 1 didapatkan beberapa hasil dari analisis deskriptif diantaranya yaitu nilai minimal, kuartil 1, median, mean, kuartil 3, dan nilai maksimum. Nilai minimum merupakan nilai terendah diantara semua nilai yang ada pada sebuah kelompok data. Kuartil 1 merupakan kuartil bawah yang memuat 25% dari data dengan nilai terendah. Median atau nilai tengah yang merupakan suatu nilai ukuran pemusatan yang menempati posisi tengah setelah data diurutkan. Mean atau rata-rata adalah ukuran pemusatan data yang digunakan untuk gambaran data yang sedang diamati. Kuartil 3 merupakan kuartil atas yang memuat 25% dari data dengan nilai tertinggi. Nilai maksimum merupakan nilai tertinggi diantara semua nilai yang ada pada sebuah kelompok data.

Dari tabel tersebut dapat diketahui bahwa pada variabel pasien terkonfirmasi atau positif memiliki nilai minimal sebesar 86, kuartil 1 sebesar 211.5, median sebesar 640.5, mean atau nilai tengah sebesar 1746.6, kuartil 3 sebesar 1677.5, dan nilai maksimum sebesar 12695. Untuk variabel pasien meninggal memiliki nilai minimal sebesar 0, kuartil 1 sebesar 5.5, median sebesar 16, mean atau nilai tengah sebesar 87.85, kuartil 3 sebesar 81.25, dan nilai maksimum sebesar 948. Sedangkan, variabel pasien sembuh memiliki nilai minimal sebesar 31, kuartil 1 sebesar 159.8, median sebesar 270, mean atau nilai tengah sebesar 784.1, kuartil 3 sebesar 862.5, dan nilai maksimum sebesar 6871.

Sebelum dilakukan analisis *cluster*, terdapat asumsi yang harus terpenuhi, yakni tidak terjadi multikoliniearitas, maka dilakukan uji multikolinearitas menggunakan progam R dengang hasil seperti berikut:

Hipotesis

H_0: Tidak terjadi multikolinearitas

H_1: Terjadi multikolinearitas

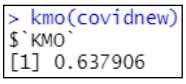
Tingkat Signifikansi

$\alpha = 0.05$

Daerah Kritis

Tolak H_0 jika p-value < α

Statistik Uji



Gambar 1. Uji KMO

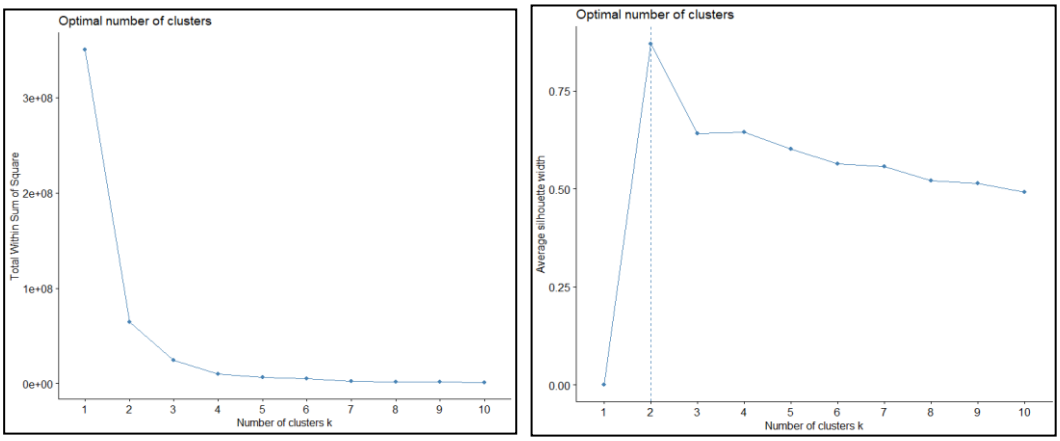
Keputusan

Gagal tolak H_0 karena nilai p-value = 0.637906 > α

Kesimpulan

Dengan menggunakan tingkat kepercayaan 95%, dapat disimpulkan bahwa data yang ada tidak terjadi *multikolinearitas*

Setelah multikolinearitas terpenuhi, maka dilakukan analisis dengan metode *K-Medoids.* Metode *cluster non* hirarki yang merupakan varian dari metode *K-Means*. *K-Medoids* menggunakan metode pengelompokan partisi untuk mengelompokan sekumpulan n objek menjadi sejumlah k-*cluster*. Banyaknya *cluster* yang akan dibentuk (*k*) pada proses peng-*cluster*-an dengan metode *K-Medoids* adalah dengan menggunakan metode untuk menentukan nilai clusteringnya yaitu *WCSS*.



Gambar 2. Plot WCSS dan Silhoutte untuk Menentukan *Cluster*

Pada Gambar 2, plot WCSS, sumbu x menunjukkan ukuran cluster yang akan ditentukan, dengan rentang grup 1 sampai 10. Sumbu y menyatakan *total within sum of square* yaitu jumlah kuadrat dan hasil kali silang dalam grup untuk ukuran variabilitas pengamatan dalam setiap cluster. Berdasarkan grafik pada Gambar 2, ketika titik *cluster* di angka 3 pada plot WCSS menunjukkan pergerakan yang mulai landai, tidak seperti perubahaan titik *cluster* angka – angka sebelumnya yang menunjukkan perubahan yang cukup curam. Semakin landai titik *cluster* maka semakin optimum jumlah *cluster* dan semakin banyaknya jumlah *cluster* yang terbentuk maka semakin sedikit jumlah *cluster* yang tergabung. Kemudian, plot Silhoutte menunjukkan titik optimumnya pada nilai 2 dengan ditunjukkan pada garis yang berada pada angka 2. Namun terdapat opsi lain yaitu nilai yang berada dibawah *cluster* optimumnya yaitu pada *cluster* 3. Pada grafik sama halnya pada plot WCSS sumbu x adalah banyaknya jumlah cluster yang akan di bentuk lalu untuk sumbu y adalah rata-rata dari lebar cluster menggunakan silhouette. Dari 2 plot tersebut memiliki nilai *cluster* yang berbeda-beda maka diambil adalah nilai yang terbesar yaitu pada metode WCSS yaitu 3. Setelah melakukan pendekatan WCSS, maka diperoleh anggota setiap grup sebagaimana tertera pada Gambar 3. Sedangkan rata-rata banyaknya kasus pada kategori pasien terkonfirmasi positif, pasien meninggal, dan pasien sembuh, sebagaimana pada Tabel 2.

	COVID19.Provinsi	pam.hasil.clustering
1	Aceh	1
2	Sumatera Utara	2
3	Sumatera Barat	1
4	Riau	1
5	Jambi	1
6	Sumatera Selatan	2
7	Bengkulu	1
8	Kepulauan Bangka Belitung	1
9	Lampung	1
10	Kepulauan Riau	1
11	DKI Jakarta	3
12	Jawa Barat	2
13	Jawa Tengah	2
14	Daerah Istimewa Yogyakarta	1
15	Jawa Timur	3
16	Banten	2
17	Bali	2
18	Nusa Tenggara Barat	2
19	Nusa Tenggara Timur	1
20	Kalimantan Barat	1
21	Kalimantan Tengah	1
22	Kalimantan Selatan	2
23	Kalimantan Timur	1
24	Kalimantan Utara	1
25	Sulawesi Utara	1
26	Sulawesi Tengah	1
27	Sulawesi Selatan	2
28	Sulawesi Tenggara	1
29	Gorontalo	1
30	Sulawesi Barat	1
31	Maluku	1
32	Maluku Utara	1
33	Papua	2
34	Papua Barat	1

Gambar 3. *Output Software R Clustering* Provinsi

Tabel 2. Analisis Cluster *K-Medoids*

Cluster	Variabel		
	Terkonfirmasi	Meninggal	Sembuh
1	12259	793	5631
2	2632	108	1077
3	388	1	210

Berdasarkan Tabel 2, dapat diketahui rata–rata pada tiga *cluster K-Medoids,* dengan interpretasi Sebagai berikut. *Cluster* 1: 2 Provinsi yang termasuk di dalam *cluster* ini yaitu pada DKI Jakarta dan Jawa Timur. Pada tingkatan ini dikategorikan tinggi karena berdasarkan rata-rata clusternya variabel terkonfirmasi, meninggal, dan sembuh memiliki rata-rata tertinggi dibandingkan dengan cluster 2 dan 3. *Cluster* 2: terdapat 10 Provinsi yang termasuk di dalam *cluster* ini yaitu pada Sumatra Utara, Kalimantan Selatan, Sumatera Selatan, Jawa Tengah, Jawa Barat, Bali, Banten, Papua, NTB, dan Sulawesi Selatan. Pada tingkatan ini dikategorikan sedang karena berdasarkan rata-rata clusternya variabel terkonfirmasi, meninggal, dan sembuh memiliki rata-rata menengah diantara cluster 1 dan 3. *Cluster* 3: terdapat 22 Provinsi yang termasuk di dalam *cluster* ini yaitu pada Aceh, Bangka-Belitung, Bengkulu, Gorontalo, Irian Jaya Barat, Jambi, Kalimantan Barat, Kalimantan Tengah, Kalimantan Timur, Kalimantan Utara, Kepulauan Riau, Lampung, Maluku Utara, Maluku, NTT, Riau, Sulawesi Barat, Sulawesi Tengah, Sulawesi Tenggara, Sulawesi Utara, Sumatra Barat, Yogyakarta. Pada tingkatan ini dikategorikan rendah karena berdasarkan rata-rata clusternya variabel terkonfirmasi, meninggal, dan sembuh memiliki rata-rata terendah diantara cluster 1 dan 2. Berikut adalah peta *thematic* untuk berdasarkan penjabaran *cluster* di Indonesia.

Gambar 3. *Thematic Map* Berdasarkan *Cluster*

Berdasarkan hasil analisis, bahwasanya provinsi yang terkonfirmasi positif corona terbagi menjadi 3 kelompok, 2 provinsi masuk ke kelompok 1, 10 provinsi masuk ke kelompok 2, dan 22 provinsi masuk ke kelompok 3. Pada Kelompok 1 (kategori tinggi) dengan nilai rata rata kasus terkonfirmasi sebesar 12259 kasus, kasus meninggal sebesar 793, dan rata-rata kasus sembuh sebesar 5691. Pada Kelompok 2 (kategori sedang) dengan nilai rata rata kasus terkonfirmasi sebesar 2632 kasus, kasus meninggal sebesar 108, dan rata-rata kasus sembuh sebesar 1077. Pada Kelompok 3 (kategori rendah) dengan nilai rata rata kasus terkonfirmasi sebesar 388 kasus, kasus meninggal sebesar 14.6, dan rata-rata kasus sembuh sebesar 210.

4. Penutup

Berdasarkan hasil analisis *clustering* didapatkan kesimpulan bahwa dengan pendekatan K-Medoids diperoleh optimum jumlah kelompok sebanyak tiga. Profil dari ketiga grup berdasarkan rata-rata jumlah kasus, maka diperoleh bahwa kelompok 1 masuk kategori tinggi, kelompok 2 masuk kategori sedang, dan kelompok 3 masuk kategori rendah. Adapun anggota dari setiap grup tersebut adalah anggota kelompok 1 sebanyak 2 provinsi yaitu DKI Jakarta dan Jawa Timur. Lalu, anggota kelompok 2 sebanyak 10 provinsi yaitu Sumatra Utara, Kalimantan Selatan, Sumatera Selatan, Jawa Tengah, Jawa Barat, Bali, Banten, Papua, NTB, dan Sulawesi Selatan. Selanjutnya anggota kelompok 3 sebanyak 22 provinsi yaitu Aceh, Bangka-Belitung, Bengkulu, Gorontalo, Irian Jaya Barat, Jambi, Kalimantan Barat, Kalimantan Tengah, Kalimantan Timur, Kalimantan Utara, Kepulauan Riau, Lampung, Maluku Utara, Maluku, NTT, Riau, Sulawesi Barat, Sulawesi Tengah, Sulawesi Tenggara, Sulawesi Utara, Sumatra Barat, Yogyakarta.

Daftar Pustaka

Abhishek, P. & Purnima, S. (2013). *New Approach for K-means and K-Medoids Algoritm. International journal of Computer Application Technology and Research*. Vol. 2. India.

Bhat, A. (2014). *K-Medoids Clustering Using Partitioning Around Medoids for Performing Face Recognition. International Joural of Soft Computing, Mathematics and Control* (IJSCMC), Vol. 3, No.3.

BNPB. (2020, Juni 30). *Indonesia COVID-19 Hub Site*. Retrieved from Gugus Tugas Percepatan Penanganan COVID-19 Republik Indonesia: https://bnpb-inacovid19.hub.arcgis.com/

Kaur, Noor K., Kaur, Usvir., & Singh, Dr.Dheerendra. (2014). *K-Medoids Clustering Algorithm – A Review*. [pdf] International Journal of Computer Application and Technology (IJCAT). ISSN. 2349-1841 Vol. 1, Issue 1. April 2014.

Maududie, & Wibowo. (2014). Metode Adaptive-Sectting Divisive Clustering dengan Pendekatan Graf Hutan Yang Minimum. *Konferensi Nasional Ilmu Komputer (KONIK) 2014 ISSN : 23 38-3899*. Jember.

Tallahasse, F. (2020, Juli 28). *Introduction to Data Mining Clustering Peixiang Zhao*. Retrieved from slideplayer.com/slide/10431363/

Wira, B., Endy Budianto, A., & Sartika Wiguna, A. (2019). IMPLEMENTASI METODE K-MEDOIDS CLUSTERING UNTUK MENGETAHUI POLA PEMILIHAN PROGRAM STUDI MAHASISWA BARU TAHUN 2018 DI UNIVERSITAS KANJURUHAN MALANG. *Jurnal Terapan Sains & Teknologi*.

Ucapan Terimakasih

Penulis mengucapkan rasa terima kasih yang sangat dalam kepada pihak Universitas PGRI Semarang yang memberikan waktu kepada kami dalam seminar ini. Tidak lupa kepada Dosen Pembimbing mata kuliah *Statistical Consulting* yang telah membimbing penulis sampai akhir. Serta, kepada teman-teman yang sudah mendukung dari awal sampai akhir pengerjaan.