



Pengelompokkan Data Kasus Covid-19 di Sumatera Utara Menggunakan Metode K-Means

Siti Sahara Hsb¹, Pasukat Sembiring²

Universitas Sumatera Utara^{1,2}

e-mail: saharasiti07@gmail.com

Abstract

Cluster analysis is a multivariate technique aimed at grouping objects into distinct categories, where there are significant differences between the groups, while the objects within a group are similar or relatively close. In this study, the distance used for grouping objects is Euclidean distance. This research aims to cluster the regencies/cities in North Sumatra based on Covid-19 case data from November 4, 2021, and to identify the characteristics of each formed cluster based on the categories of Covid-19 case distribution levels, namely high (Cluster 1), moderate (Cluster 2), and low (Cluster 3). The results show that, out of 33 regencies/cities analyzed, 3 clusters were formed. Cluster 1 consists of 1 member, Cluster 2 consists of 13 members, and Cluster 3 consists of 18 members, while 1 regency/city is identified as an outlier. This study provides insights into the grouping of areas based on the level of Covid-19 case distribution, which can serve as a basis for decision-making regarding the management of the pandemic.

Keywords: Cluster, K-Means, Clustering.

Abstrak

Cluster analysis adalah salah satu teknik multivariat yang bertujuan untuk mengelompokkan objek-objek ke dalam kategori yang berbeda, di mana antar kelompok memiliki perbedaan yang signifikan, sementara objek dalam satu kelompok memiliki kesamaan atau kedekatan relatif. Dalam penelitian ini, jarak yang digunakan untuk mengelompokkan objek adalah jarak Euclidean. Penelitian ini bertujuan untuk mengelompokkan Kabupaten/Kota di Sumatera Utara berdasarkan data kasus Covid-19 pada tanggal 4 November 2021 dan mengidentifikasi karakteristik masing-masing Cluster yang terbentuk berdasarkan kategori tingkat persebaran kasus Covid-19, yaitu tinggi (Cluster 1), sedang (Cluster 2), dan rendah (Cluster 3). Hasil penelitian menunjukkan bahwa dari 33 Kabupaten/Kota yang dianalisis, terbentuk 3 Cluster. Cluster 1 terdiri dari 1 anggota, Cluster 2 terdiri dari 13 anggota, dan Cluster 3 terdiri dari 18 anggota, sementara 1 Kabupaten/Kota teridentifikasi sebagai Outlier. Penelitian ini memberikan wawasan mengenai pengelompokan wilayah berdasarkan tingkat persebaran kasus Covid-19 yang dapat digunakan sebagai dasar dalam pengambilan keputusan terkait penanganan pandemi.

Kata Kunci: Cluster, K-Means, Pengelompokan.

PENDAHULUAN

Pengelompokan data Mining dapat dibagi menjadi 6 kelompok yaitu deskripsi, estimasi, prediksi, klasifikasi, Clustering (pengelompokan), dan asosiasi. Clustering akan melakukan pengelompokan data-data ke dalam sejumlah kelompok (Cluster) berdasarkan kesamaan karakteristik masing-masing data pada kelompok-kelompok yang ada (Kusrini dan E.T. Luthfi, 2009). Analisis Cluster adalah suatu teknik analisis multivariat yang tujuan proses utamanya adalah untuk mengelompokkan objek individu yang didasarkan pada karakteristik tertentu. Analisis Cluster mengklasifikasikan objek sehingga suatu objek memiliki kedekatan dengan objek yang lainnya. Analisis Cluster bekerja atas dasar kesamaan objek atau sejenisnya (bukan variabel) (Widayat, 2018).

Analisis Cluster bertujuan untuk mengetahui karakteristik data pasien yang positif Covid-19 di Sumatera Utara. Analisis Cluster mengarah pada pengelompokan suatu objek berdasarkan karakteristik dengan menggunakan kumpulan variabel yang ditentukan oleh peneliti. Salah satu kelebihan analisis Cluster ini adalah mengelompokkan data dalam kuantitas yang cukup besar dan variabel yang berlipat-lipat (Nariawati, 2008).

Analisis Cluster dibagi menjadi dua metode yaitu metode hirarki dan metode non hirarki. Metode hirarki dibagi menjadi dua, yaitu metode agglomerative (pemusatan) dan metode divisive (penyebaran). Dalam metode hirarki, jumlah kelompok yang akan diperoleh belum diketahui, sedangkan dalam metode non hirarki diasumsikan ada k kelompok terlebih dahulu. Metode-metode yang termasuk dalam metode agglomerative adalah Single Linkage Method, Average Linkage Method, Ward's Method, Centroid Method dan Median Method (Everitt, 1974). Sedangkan metode yang termasuk metode non hirarki adalah metode non hirarki adalah metode K-Means dan Fuzzy Method.

K-Means merupakan salah satu algoritma Clustering yang masuk dalam kelompok Unsupervised learning yang digunakan untuk membagi data menjadi beberapa kelompok dengan sistem partisi. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan K-Nearest Neighbor (KNN) dan algoritma supervised learning lainnya yang menerima masukan berupa vektor. Pada algoritma K-Means, komputer mengelompokkan sendiri data-data yang menjadi masukannya tanpa mengetahui terlebih dahulu target kelasnya. Masukan yang diterima adalah data atau objek dan k buah kelompok (Cluster) yang diinginkan. Algoritma ini akan mengelompokkan data atau objek ke dalam k buah kelompok tersebut (Wanto, Anjar et al. 2020).

K-Means Cluster sebagai salah satu metode data Clustering non-hierarki mempartisi data yang ada ke dalam bentuk atau lebih Cluster atau kelompok, sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam

satu cluster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain. Kelompok atau Cluster yang di dapat merupakan pengetahuan/informasi yang bermanfaat bagi pengguna kebijakan dalam proses pengambilan keputusan (Usman, Hardius dan Nurdin Sobari, 2013).

Pada setiap Cluster terdapat titik pusat (Centroid) yang mempresentasikan Cluster tersebut. Secara sederhana algoritma K-Means dapat dijelaskan sebagai algoritma data mining yang digunakan untuk menyelesaikan masalah pengelompokan. Pada pemrosesan data algoritma K-Means Clustering, akan diawali dengan pengelompokan Centroid pertama yang dipilih secara acak sebagai titik awal untuk setiap Cluster, kemudian menghitung secara berulang agar posisi Centroid optimal (Wanto, Anjar et al. 2020).

Covid-19 adalah penyakit yang disebabkan oleh virus Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Covid-19 dapat menyebabkan gangguan sistem pernapasan, mulai dari gejala yang ringan seperti flu, hingga infeksi paru-paru, seperti pneumonia. Covid-19 (coronavirus disease 2019) adalah jenis penyakit baru yang disebabkan oleh virus dari golongan coronavirus, yaitu SARS-CoV-2 yang juga sering disebut virus Corona (dr. Pittara, 2020).

Dalam menyikapi pandemi Covid-19 masyarakat dan pemerintah melakukan berbagai hal dalam mencegah penyebaran kasus Covid-19. Contohnya masyarakat melakukan isolasi mandiri jika mengalami gejala-gejala terindikasi Covid-19, menunda keberangkatan ke daerah-daerah yang terindikasi Covid-19 dengan kasus positif yang tinggi. Pemerintah juga melakukan berbagai usaha pencegahan seperti memberikan fasilitas cuci tangan di tempat-tempat umum, melakukan Pembatasan Kegiatan Masyarakat (PPKM) di daerah-daerah terdampak Covid-19 terutama di hari libur. Dan dalam satu tahun terakhir sudah di anjurkan juga supaya masyarakat melakukan vaksin Covid-19 di Puskesmas atau tempat-tempat yang di sediakan Pemerintah.

Dalam penelitian ini akan dianalisis kasus pengelompokan K-Means Cluster di seluruh Kabupaten/Kota Provinsi Sumatera Utara adalah salah satu metode dan Clustering non-hierarki yang berusaha mengelompokkan data ke dalam suatu cluster sehingga data yang memiliki karakteristik yang sama di kelompokkan ke dalam satu Cluster yang sama. Ukuran kesamaan yang digunakan adalah ukuran jarak antar objek. Kedua objek yang memiliki jarak paling dekat akan bergabung menjadi satu Cluster. Kedekatan jarak yang dimiliki menunjukkan bahwa kedua objek tersebut memiliki tingkat kesamaan karakteristik.

Pada penelitian Zulia Imami Alfianti dengan judul Pengelompokan Wilayah Penyebaran Covid-19 di Kabupaten Karawang Menggunakan Algoritma K-Means menentukan jumlah Cluster sebanyak 3 yaitu Cluster wilayah penyebaran rendah, Cluster wilayah penyebaran sedang, dan Cluster wilayah penyebaran tinggi. Dari hasil penelitian didapat 50% wilayah yang tingkat penyebarannya rendah, 33,3% wilayah yang tingkat penyebarannya sedang, dan 16,7% wilayah yang tingkat penyebarannya tinggi.

Pada penelitian Darmansah dan Ni Wayan Wardani dengan judul Analisis Persebaran Penularan Virus Corona di Provinsi Jawa Tengah Menggunakan Metode K-Means Clustering yaitu menentukan jumlah cluster terdiri dari tiga macam Cluster yaitu C0, C1 dan C2. Dimana C0 adalah tingkat persebaran virus Covid-19 di kabupaten/kota se-Jawa Tengah dengan kategori rendah, C1 dengan kategori sedang dan C2 dengan kategori tinggi. Kabupaten/kota yang tergolong penyebaran penularan virus Covid-19 terendah (C0) di Jawa Tengah terdapat 18 kabupaten/kota, terdapat 1 kabupaten/kota dengan penularan Covid-19 sedang (C1) dan terdapat 16 kabupaten/kota di Jawa Tengah yang kategori persebaran penularan virus Covid-19 tertinggi (C2).

METODE PENELITIAN

Penelitian ini menggunakan data sekunder yang diperoleh dari catatan dan dokumen di Kantor Dinas Kesehatan Provinsi Sumatera Utara, khususnya di bidang Pencegahan dan Pengendalian Penyakit (P2P). Data tersebut mencakup informasi tentang kasus Covid-19, pasien yang sembuh, dan pasien yang meninggal dunia di Kabupaten/Kota di Sumatera Utara per 4 November 2021. Variabel penelitian mencakup kasus positif terinfeksi Covid-19, pasien sembuh, dan pasien meninggal, dengan analisis menggunakan metode K-Means Cluster. Proses penelitian dimulai dengan studi pendahuluan untuk memahami kondisi awal Covid-19 di wilayah tersebut, dampak ketidaktahuan masyarakat dan pemerintah, serta teori pendukung terkait pemecahan masalah. Data dikumpulkan dengan membaca dokumen yang relevan, kemudian diproses menggunakan K-Means Cluster dengan langkah-langkah seperti standarisasi data, penentuan jumlah cluster, pemilihan centroid awal secara acak, iterasi pertama, dan perhitungan jarak Euclidean untuk mengalokasikan data ke dalam cluster. Iterasi berakhir ketika nilai centroid awal dan centroid baru sudah sama.

PEMBAHASAN

Pembahasan hasil analisis ini berfokus pada penerapan metode K-Means clustering untuk mengelompokkan kabupaten/kota di Provinsi Sumatera Utara berdasarkan tingkat penyebaran kasus Covid-19 pada tanggal 4 November 2021. K-Means merupakan metode clustering non-hierarki yang membagi data menjadi beberapa kelompok (cluster) dengan menggunakan jarak Euclidean sebagai ukuran kesamaan antar objek. Setiap objek yang memiliki kedekatan atau kesamaan karakteristik dikelompokkan ke dalam cluster yang sama.

Berdasarkan hasil analisis, terdapat tiga cluster yang terbentuk: Cluster 1, Cluster 2, dan Cluster 3. Cluster 1, yang terdiri dari satu kabupaten/kota, menunjukkan daerah dengan tingkat penyebaran Covid-19 yang tinggi. Sedangkan Cluster 2, yang beranggotakan 13 kabupaten/kota, mencakup daerah dengan tingkat penyebaran Covid-19 sedang. Cluster 3, yang terdiri dari 18 kabupaten/kota, mewakili daerah dengan tingkat penyebaran Covid-19 yang rendah. Hasil ini menunjukkan bahwa mayoritas kabupaten/kota di Sumatera Utara berada pada kategori penyebaran Covid-19 sedang hingga rendah, dengan hanya satu kabupaten/kota yang terklasifikasi pada kategori penyebaran tinggi.

Proses analisis menggunakan K-Means clustering dimulai dengan pemilihan centroid acak untuk setiap cluster, kemudian dilakukan iterasi berulang untuk mengalokasikan data ke dalam cluster yang sesuai. Proses ini berlanjut hingga tidak ada perubahan signifikan pada nilai centroid, menandakan bahwa pembagian cluster sudah optimal. Pada Tabel 1, terlihat bahwa pada Iterasi keempat dan kelima, nilai centroid pada setiap cluster tidak mengalami perubahan, yang menandakan bahwa proses pengelompokan sudah stabil dan tidak memerlukan perubahan lebih lanjut.

Tabel 1
Proses Iterasi Pertama-Iterasi Kelima

	Iterasi Pertama	Iterasi Kedua	Iterasi Ketiga	Iterasi Keempat	Iterasi Kelima
<i>Cluster 1</i>					
XP	23	45	45	45	45
XS	7655	16119	16119	16119	16119
XM	235,67	425	425	425	425
n_i	3	1	1	1	1
	Iterasi Pertama	Iterasi Kedua	Iterasi Ketiga	Iterasi Keempat	Iterasi Kelima
<i>Cluster 2</i>					
XP	3	4,65	4,93	5,69	5,69
XS	1506,16	1891,18	1998,8	2120,31	2120,31
XM	59,26	74,88	78,33	85,69	85,69
n_i	19	17	15	13	13
	Iterasi Pertama	Iterasi Kedua	Iterasi Ketiga	Iterasi Keempat	Iterasi Kelima
<i>Cluster 3</i>					
XP	4,1	3,07	3	2,67	2,67
XS	336,8	477,21	553,06	625,94	625,94
XM	13,8	19,5	23,19	24	24
n_i	10	14	16	18	18

Sumber; Data diolah, 2024

Pada Tabel 1 dapat dilihat bahwa Iterasi keempat dan Iterasi kelima tidak mengalami perubahan nilai *Centroid*, sehingga proses Iterasi dihentikan. Dari hasil *Clusterisasi*, didapatkan anggota dari *Cluster* 1 sebanyak 1 Kabupaten/Kota, *Cluster* 2 sebanyak 13 Kabupaten/Kota, dan *Cluster* 3 sebanyak 18 Kabupaten/Kota.

Berikut adalah hasil *Clusterisasi* untuk setiap Kabupaten/Kota di Sumatera Utara beserta kategori tingkat persebarannya:

Tabel 2
Hasil *Clusterisasi*

No	Kabupaten/Kota	Cluster	Kategori
1	Deli Serdang	1	Tinggi
2	Kota Pematang Siantar	2	Sedang
3	Simalungun	2	Sedang
4	Asahan	2	Sedang
5	Serdang Berdagai	2	Sedang
6	Langkat	2	Sedang
7	Kota Binjai	2	Sedang
8	Dairi	2	Sedang
9	Karo	2	Sedang
10	Tapanuli Utara	2	Sedang
11	Labuhan Batu	2	Sedang
12	Batu Bara	2	Sedang
13	Toba	2	Sedang
14	Samosir	2	Sedang
15	Kota Tebing Tinggi	3	Rendah
16	Kota Gunung Sitoli	3	Rendah
17	Labuhan Batu Utara	3	Rendah
18	Kota Padangsidempuan	3	Rendah
19	Tapanuli Tengah	3	Rendah
20	Kota Sibolga	3	Rendah
21	Labuhan Batu Selatan	3	Rendah
22	Tapanuli Selatan	3	Rendah
23	Mandailing Natal	3	Rendah
24	Kota Tanjung Balai	3	Rendah
25	Humbang Hasundutan	3	Rendah
26	Papak Bharat	3	Rendah
27	Padang Lawas	3	Rendah

28	Padang Lawas Utara	3	Rendah
29	Nias	3	Rendah
30	Nias Selatan	3	Rendah
31	Nias Utara	3	Rendah
32	Nias Barat	3	Rendah

Sumber; Data diolah, 2024

Pada tabel 2 dapat dilihat bahwa anggota *Cluster 1* untuk kategori kasus persebaran *Covid-19* tinggi adalah Kota Deli Serdang. Sedangkan *Cluster 2* untuk kategori kasus *Covid-19* sedang anggotanya adalah Kota Pematang Siantar, Simalungun, Asahan, Serdang Berdagai, Langkat, Binjai, Dairi, Karo, Tapanuli Utara, Labuhan Batu, Batu Bara, Toba, Samosir. Dan untuk *Cluster 3* kategori kasus persebaran *Covid-19* rendah anggotanya adalah Kota Tebing tinggi, Gunung Sitoli, Labuhan Batu Utara, Padangsidempuan, Tapanuli Tengah, Sibolga, Labuhan Batu Selatan, Tapanuli Selatan, Mandailing Natal, Tanjung Balai, Humbang Hasundutan, Papak Bharat, Padang Lawas, Padang Lawas Utara, Nias, Nias Selatan, Nias Utara, Nias Barat.

Tabel 2 menunjukkan hasil akhir pengelompokan untuk setiap kabupaten/kota, dengan kategorisasi tingkat penyebaran Covid-19 yang diperoleh dari clusterisasi ini. Kota Deli Serdang termasuk dalam Cluster 1 dengan kategori penyebaran tinggi, sementara kabupaten/kota lainnya terdistribusi dalam Cluster 2 dan Cluster 3, yang masing-masing memiliki kategori penyebaran sedang dan rendah. Hasil ini memberikan gambaran yang jelas tentang distribusi kasus Covid-19 di Sumatera Utara, yang dapat digunakan oleh pemerintah dan pihak terkait untuk merencanakan intervensi yang lebih tepat sasaran berdasarkan tingkat penyebaran yang ada di masing-masing daerah. Penelitian ini berhasil memetakan distribusi penyebaran Covid-19 di Sumatera Utara dengan menggunakan metode K-Means clustering. Hasil yang diperoleh dapat dijadikan dasar dalam pengambilan keputusan untuk kebijakan pencegahan dan penanganan Covid-19 di setiap kabupaten/kota, terutama dalam merancang program vaksinasi atau pembatasan kegiatan masyarakat yang lebih spesifik berdasarkan tingkat risiko penyebaran yang ada.

KESIMPULAN

Berdasarkan hasil dan pembahasan diperoleh hasil Clusterisasi berdasarkan data kasus Covid-19 pada tanggal 4 November 2021 untuk 32 Kabupaten/Kota, dimana sebanyak 1 Kabupaten/Kota yaitu Kota Medan tidak di sertakan dalam proses Clusterisasi karena termasuk dalam data Outlier. Untuk Cluster 1 terdapat sebanyak 1 Kabupaten/Kota, Cluster 2 terdapat sebanyak 13 Kabupaten/Kota, dan Cluster 3 terdapat sebanyak 18 Kabupaten/Kota. Maka disimpulkan bahwa terdapat 1 Kabupaten/Kota dari 32 data yang di Clusterisasi yang termasuk dalam tingkat kasus persebaran Covid-19 tinggi, 13

Kabupaten/Kota yang termasuk dalam tingkat kasus persebaran Covid-19 sedang, dan 18 Kabupaten/Kota yang termasuk dalam tingkat kasus persebaran Covid-19 rendah.

DAFTAR PUSTAKA

- Kusrini dan E.T. Luthfi. 2009. *Algoritma Data Mining*. Yogyakarta: C.V Andi Offset.
- Widayat. 2018. *Statistik Multivariat pada Bidang Manajemen dan Bisnis*. Malang: Universitas Muhammadiyah Malang.
- Nariawati, Umi. 2008. *Teknik-teknik Analisis Multivariat untuk Riset Ekonomi*. Surabaya: Graha Ilmu.
- Everitt, B. 1974. *Cluster Analysis*. London: Heinemann.
- Usman, Hardius dan Nurdin Sobari. 2013. *Aplikasi Teknik Multivariate*. Jakarta: PT Raja Grafindo Persada.
- Wanto, Anjar et al. 2020. *Data Mining Algoritma dan Implementasi*. Jakarta: Yayasan Kita Menulis.
- Pittara, dr. 2020. "Covid-19". <http://www.alodokter.com/covid-19>, diakses pada 10 November 2021 pukul 20.23.
- Darmansah, Ni Wayan Wardani. 2021. Analisis Pesebaran Penularan Virus Corona di Provinsi Jawa Tengah Menggunakan Metode K-Means Clstering. *Jurnal Teknik Informatika dan Sistem Informasi* Vol. 8 No. 1. Institut Teknologi Telkom Purwokerto.
- Zulia Imami Alfianti. 2021. Pengelompokan Wilayah Penyebaran Covid-19 di Kabupaten Karawang Menggunakan Algoritma K-Means. *Jurnal Universitas Bina Sarana Informatika* Vol. 26 No. 2. Kampus Kabupaten Karawang.
- Fathia, A. N. et al. 2016. Analisis Klaster Kecamatan di Kabupaten Semarang Berdasarkan Potensi Desa Menggunakan Metode K-Means. *None*, 5(4), 801-810.
- Irwinsyah, Edy dan Muhammad Faisal. 2015. *Advanced Clustering Teori dan Aplikasi*. Bandung: Graha Ilmu.
- Ediyanto et al. 2013. Pengklasifikasian Karakteristik Dengan Metode K-Means Cluster Analysis. *Buletin Ilmiah Mat. Stat. Dan Terapannya (Bimaster)*, 02(2), 133-136
- Wang, Xiaoyan and Yanping Bai. The Global Minimax K-Means Algorithm. *Springer Plus* (2016) 5:1665 DOI 10.1186/s40064-3329-4
- Kim Sung Hwan. Weighted K-Menas Support Vektor Machine For Cancer Prediction. *Kim SpringPlus* (2016) 5:1162 DOI 10.1186/s40064-016-2677-4.
- Kurniawan, Robert et al. 2019. *Cara Mudah Belajar Statistik Analisis Data dan Eksplorasi*. Jakarta: Kencana.
- Johnson dan Winchen. 1998. *Applied Multivariate Statistical Analysis* 4th Edition. Pearson Education International, New Jersey.
- Kemenkes RI. 2020. Tanya Jawab Seputar Virus Corona. 119-135.