

Implementasi Clustering Data Kasus Covid 19 Di Indonesia Menggunakan Algoritma K-Means

Nofita Sari¹, Hanny Hikmayanti Handayani², Amril Mutoi Siregar³

Universitas Buana Perjuangan Karawang

if19.nofitasari@mhs.ubpkarawang.ac.id¹, hanny.hikmayanti@ubpkarawang.ac.id²,
amrilmutoi@ubpkarawang.ac.id³

Abstrak - Covid19 merupakan virus yang pertama kali ditemukan di Wuhan, China pada akhir Desember 2019. Kasus Covid19 masuk ke Indonesia pada Maret 2020 tercatat telah mencapai 1.511.712 dengan 40.858 kematian dan 1.348.330 kasus sembuh. Di Indonesia terdapat 34 provinsi tempat penyebaran kasus Covid19. Penelitian ini bertujuan untuk mengelompokkan setiap provinsi di Indonesia ke dalam cluster tertentu untuk mengidentifikasi daerah dengan jumlah kasus tinggi, sedang, dan rendah. Pengelompokan data kasus Covid19 di provinsi Indonesia menggunakan teknik clustering menggunakan algoritma K-means. Dataset yang digunakan adalah 7098 data dari tanggal 1 Maret sampai dengan 11 Oktober 2020. Dataset yang digunakan adalah dari website AtapData (atapdata.ai). Pada penelitian ini dilakukan optimasi cluster yang diperoleh dengan metode elbow dengan pengolahan data menggunakan Google Collaboratory dengan bahasa pemrograman python yang menghasilkan total 3 cluster. Hasil pengujian dilakukan untuk mendapatkan nilai K optimal. Evaluasi menggunakan Sum of Square Error (SSE). Dari hasil evaluasi memiliki jumlah optimal K:3 yaitu 228913736548657.56.

Kata Kunci : Covid19, algoritma K means, *Clustering*, *Metode Elbow*

Abstract - Covid19 is a virus that was first discovered in Wuhan, China at the end of December 2019. Covid19 cases entered Indonesia in March 2020 it was recorded that it had reached 1,511,712 with 40,858 deaths and 1,348,330 cases recovered. In Indonesia there are 34 provinces where the spread of Covid19 cases. This study aims to classify each province in Indonesia into certain clusters to identify areas with high, medium, and low number of cases. The grouping of Covid19 case data in the Indonesian province uses a clustering technique using the K-means algorithm. The dataset used is 7098 data from March 1 to October 11 2020. The dataset used is from the AtapData website (atapdata.ai). In this study, cluster optimization was carried out using the elbow method with data processing using the Google Collaboratory with the Python programming language which resulted in a total of 3 clusters. The test results were carried out to obtain the optimal K value. Evaluation using Sum of Square Error (SSE). From the evaluation results, it has an optimal number of K: 3, namely 228913736548657.56.

Keywords: Covid19, K mean algorithm, *Clustering*, *Elbow Method*

I. PENDAHULUAN

Covid-19 ialah penyakit virus yang menyebabkan Pneumonia dan masalah pernapasan disebabkan oleh virus yang dikenal sebagai Covid-19. Pada akhir Desember 2019, kasus pertama ditemukan di Wuhan, China dan telah menyebar ke seluruh dunia, termasuk Indonesia. Gejala klinis yang dialami oleh Covid-19 tersebut berkisar dari gejala flu biasa hingga komplikasi serius (*pneumonia* atau *sepsis*) (Wiguna et al., 2020).

Penyebab utama penyebaran Covid19 adalah penularan Covid-19 dari manusia ke manusia. Banyak orang meninggal akibat penyebaran virus yang cepat dan signifikan ini. Masyarakat khawatir tentang virus ini karena adalah darurat kesehatan masyarakat. (*Public Health Emergency of International Concern*). Peningkatan jumlah kasus Covid-19 hampir diseluruh wilayah provinsi Indonesia terdapat kasus Covid19. Setiap hari kasus Covid19 terus meningkat. Menurut data Kementerian

Kesehatan RI, pada 31 Maret 2021 jumlah kasus positif Covid19 di Indonesia dengan jumlah mencapai 1.511.712, 40.858 kematian dan 1.348.330 kasus sembuh (Zulfa et al., 2021).

Organisasi Kesehatan Dunia (WHO) telah menetapkan fenomena Covid19 pada 11 Maret 2020 (Negari & Eryando, 2021). Secara nasional melalui keputusan Badan Nasional Penanggulangan Bencana (BNPB). Menyikapi pandemi Covid-19, Pemerintah Indonesia salah satu kebijakannya adalah pada awal Maret 2020 masyarakat melakukan social distancing untuk memutus rantai penyebaran Covid19 yang tersebar di berbagai provinsi di Indonesia (Zulfa et al., 2021). Namun yang terjadi adalah masyarakat tidak mematuhi dengan baik, masih rendahnya kesadaran masyarakat dalam menangani kasus ini. Jumlah kasus Covid-19 terus meningkat dan membutuhkan upaya penanganan. Penyebaran kasus Covid-19 menimbulkan banyak korban jiwa, mengganggu

kehidupan masyarakat Indonesia, serta aspek perekonomian Indonesia.

Data provinsi dapat ditemukan dan diakses di situs web Kaggle, namun data yang tersedia sangat banyak, sehingga Big Data sangat sulit untuk digunakan dengan baik. Agar dapat memperoleh informasi yang berguna untuk digunakan, dapat menggunakan teknik Data Mining untuk mengolah data dengan jumlah besar. Teknik dapat digunakan adalah teknik clustering, dimana data berdasarkan mean terdekat, sehingga provinsi yang dapat dikategorikan menjadi cluster berdasarkan jumlah banyaknya kasus Covid-19 pada provinsi tersebut (Ekasetya & Jananto, 2020).

Clustering adalah suatu cara menganalisa dengan mengelompokkan ke dalam kelompok kesamaan tertentu. Clustering mengacu pada pengelompokan record, mengamati dan membuat kelas objek yang memiliki kesamaan. Cluster adalah kumpulan record yang mirip satu sama lain dan berbeda dari kumpulan data cluster lainnya. Clustering mencoba membagi beberapa bagian informasi kepada kelompok yang relatif yang memiliki kemiripan, dimana kemiripan dalam kelompok bernilai minimal (Zulfa et al., 2021). Ada dua jenis metode pengelompokan yang digunakan dalam data mining untuk mengelompokkan data, yaitu pengelompokan hierarkis dan pengelompokan non-hierarkis. Prosedur untuk non-hierarchical (parsial) clustering dimulai dengan menentukan jumlah cluster yang dibutuhkan. Jika jumlah cluster diketahui, maka dilakukan proses clustering.

Metode elbow digunakan dalam penelitian ini untuk menentukan nilai k terbaik. Dengan menggunakan hasil perbandingan jumlah cluster yang terbentuk pada suatu titik tertentu, metode elbow dapat digunakan untuk menghasilkan informasi penentuan jumlah cluster yang optimal. Hitung nilai SSE (Sum of Square Error) untuk setiap cluster untuk membuat perbandingan dengan menggunakan grafik sebagai sumber data. Hasil persentase yang berbeda dari setiap nilai cluster dapat dilihat menggunakan grafik. Cluster yang terbaik jika nilai mengalami penurunan, maka cluster tersebut yang terbaik (Putu et al., n.d.).

Pada penelitian sebelumnya berjudul menentukan jumlah cluster SMK ideal di Jawa Tengah dengan menggunakan metode X-means clustering dan K-means clustering, membahas tentang revitalisasi dan pengembangan SMK dengan mengelompokkan SMK berdasarkan data informasi Kemendikbud dan kebudayaan Jawa Tengah. Menggunakan

metode X-Means dan K-Means untuk menghitung nilai Davies-Buldin Index (DBI) menggunakan data MOC dasar untuk menemukan distribusi cluster yang optimal. Hasil perhitungan X-Means dan K-Means membagi data menjadi 4 kelompok yaitu. H. kurang, cukup, baik dan unggul (Sindi et al., 2020)(Ekasetya & Jananto, 2020)(Adhitama et al., 2020).

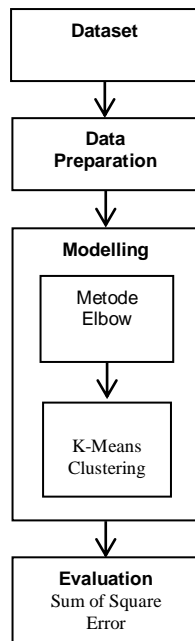
Penelitian sebelumnya (Abdullah et al., 2022) bertujuan untuk mengelompokkan provinsi-provinsi di Indonesia yang berisiko terkena pandemi COVID-19 berdasarkan data penyakit Coronavirus 2019 (COVID-19). Pengelompokan tersebut berdasarkan informasi dari Satuan Tugas Percepatan Penanganan Covid-19 Indonesia (SATGAS COVID-19) pada 19 April 2020. Algoritma yang digunakan adalah K-Means, dengan perhitungan cluster-optimized menggunakan metode Elbow. Pengelompokan tersebut menghasilkan 3 kelompok provinsi. Menggunakan algoritma K-Means untuk segmentasi kasus Covid-19 di Provinsi Lampung. Data atribut yang digunakan meliputi suspek, kabupaten, isolasi, terkonfirmasi positif dan kematian. Dibagi menjadi empat kelompok berdasarkan warna merah, jingga, kuning dan hijau. Validasi cluster dengan indeks Davies-Bouldin (Abdullah et al., 2022).

Algoritma K-Means Clustering digunakan untuk mengolah data Covid-19 dari Provinsi Indonesia. Sebuah pendekatan analisis kluster non-hierarkis yang dikenal dengan algoritma K-means clustering mencoba mengklasifikasikan data atau objek ke dalam satu atau lebih kluster berdasarkan karakteristik masing-masing cluster (Zulfa et al., 2021). Dimana untuk menentukan kelompok optimal menggunakan grafik siku. Penelitian ini bertujuan untuk mengelompokkan setiap provinsi di Indonesia ke dalam cluster tertentu untuk mengidentifikasi daerah dengan jumlah kasus tinggi, sedang, dan rendah. Terdapat 3 cluster data kasus Covid19 provinsi di Indonesia antara lain cluster 0, cluster 1 dan cluster 2. Dengan penelitian ini diharapkan hasil cluster menggunakan Algoritma K-Means dan optimasi jumlah cluster dengan metode elbow (siku) akan menghasilkan cluster terbaik (Winarta & Kurniawan, 2021).

II. METODE PENELITIAN

Metode penelitian adalah langkah-langkah yang dilakukan peneliti untuk mengumpulkan data penelitian. Informasi tersebut diolah dan dianalisis secara ilmiah untuk mencapai tujuan penelitian. Dalam tahapan penelitian ini terdapat 4 langkah yaitu Dataset, Data

Preparation, Modelling dan Evaluation. Seperti yang terlihat pada Gambar 1.



Sumber: (Isy Karima Fauzia et al., 2020)
 Gambar 1. Alur penelitian

1. Dataset

Data yang digunakan dalam artikel ini adalah dataset mengenai Covid-19 di Indonesia yang diperoleh dari website AtapData. jumlah total data yang akan digunakan adalah 7098 dari 1 Mei hingga 11 Oktober 2020 di provinsi Indonesia.

2. Algoritma K-Means

Penelitian yang dilakukan menggunakan algoritma K-means clustering untuk menganalisis Data Covid-19 di Indonesia. K-means clustering merupakan algoritma non-hierarchical clustering. Algoritma K-Means adalah teknik pengelompokan yang menggunakan jarak untuk mengelompokan data menjadi beberapa cluster (Cahyo Wiguno & Nataliani, 2022). Algoritma k-means clustering termasuk ke dalam tipe unsupervised learning. Untuk menghitung jarak semua data ke setiap titik cluster dapat menggunakan teori jarak Euclidean digunakan rumus sebagai berikut:

$$D(i, j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \dots + (X_{ki} - X_{kj})^2}$$

Dimana:

D(i,j) = jarak data i ke pusat cluster j

X_{ki} = Data ke i pada atribut data ke k

X_{kj} = Jarak antara titik pusat ke j pada atribut k

3. Evaluation

Validasi data Untuk mendapatkan perbandingan dengan menghitung Sum of Square Error (SSE) dari nilai cluster. Karena jumlah K cluster yang banyak, maka nilai SSE akan menurun secara perlahan hingga nilai K semakin kecil

(Purwayoga, 2021). Dengan menghitung persamaan rumus:

$$SSE = \sum_{K=1} k \sum_{x_i \in S_k} (X_i - C_k)^2$$

Dimana:

K = jumlah kelompok yang digunakan pada algoritma K Means

X_i = banyak nya data ke i

C_k = jumlah cluster pada cluster ke nilai K.

III. HASIL DAN PEMBAHASAN

1. Data Preparation

a) Select Data

Dataset Covid19 di Indonesia disimpan dalam bentuk file Microsoft Excel atau csv. Atribut data yang digunakan dalam tabel data Covid19 yaitu Province, Total Deaths, Total Recovered, Total Active Cases.

b) Data Cleaning

Setiap dataset yang diimpor ke file Excel akan memiliki hasil penjumlahan yang berbeda. Tabel memiliki banyak baris kosong atau tidak ada nilai. Oleh karena itu, gunakan jumlah yang dihitung dari file Excel di kumpulan dataset final. Pada langkah ini, proses imputasi dilakukan dengan menggunakan teknik imputasi mean (Ruwandara et al., 2021).

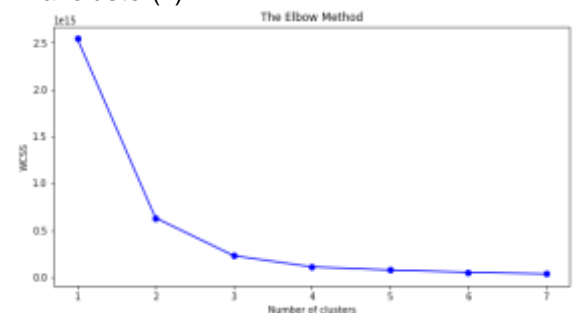
Tabel 1 Data Cleaning

Province	Total Deaths	Total Recovered	Total Active Cases
Jawa Barat	85.171033	1144358	13436
DKI Jakarta	127.589654	1386134	10864
Banten	134.458453	328482	2443
Jawa Tengah	428.736279	601517	1403
Jawa Timur	731.366324	569003	778

Sumber: Hasil Penelitian (2022)

2. Modelling

Hasil jumlah k cluster yang optimal akan digunakan sebagai prosedur pengelompokan K-Means pada hasil penentuan jumlah k cluster yang ideal. Menggunakan Google Collaboratory dengan bahasa pemrograman python, pencarian jumlah cluster optimal dilakukan selama proses ini. memanfaatkan WCSS dan Metode Elbow untuk memastikan nilai cluster(k).



Sumber: Hasil Penelitian (2022)

Gambar 2 Metode Elbow

Pada visualisasi ini dapat terlihat elbow pada data Covid-19 diujikan melalui algoritma K-Means dengan jumlah cluster K=2 dan K=3 dan dalam pengujian ini menggunakan k=3. Agar lebih terlihat dan dapat lebih dapat insight yang diberikan.

3. Pengelompokan K-Means

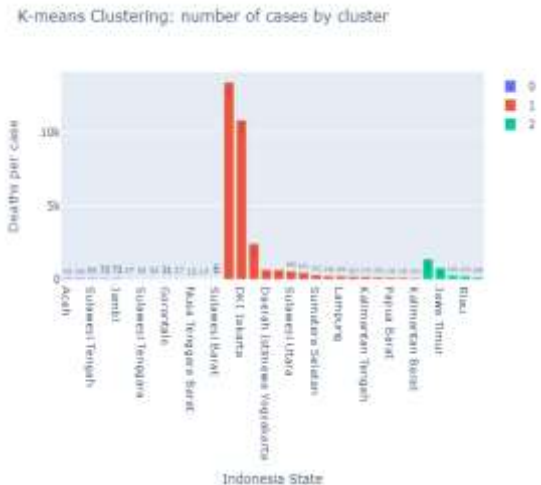
Dataset Covid-19 terdapat 34 Provinsi di Indonesia. Dari semua data yang diperoleh terdapat 4 atribut, dilakukan menggunakan proses perhitungan data dengan algoritma Kmeans. Contoh perhitungan dipilih cluster dapat dilihat pada tabel 2.

Tabel 2 Cluster data kasus Covid-19

Province	Total Deaths	Total Recovered	Total Active Cases	K-means
Jatim	1.00000	1.00000	1.00000	2
Kaltim	1.00000	1.00000	1.00000	2
Sulsel	1.00000	1.00000	0.50000	2
Jateng	0.50000	1.00000	1.00000	2
Riau	0.50000	1.00000	0.50000	2
Sumbar	0.50000	0.50000	0.50000	1
Kalsel	0.50000	0.50000	0.50000	1
Kep. Riau	0.50000	0.50000	0.50000	1
Kalbar	0.50000	0.50000	0.50000	1
Jabar	0.00000	1.00000	1.00000	1
Jakarta	0.00000	1.00000	1.00000	1
Banten	0.00000	1.00000	1.00000	1
Yogyakarta	0.00000	1.00000	1.00000	1
SUMUT	0.00000	1.00000	1.00000	1
Bali	0.00000	1.00000	1.00000	1
Sulut	0.00000	0.50000	1.00000	1
Sumsel	0.00000	0.50000	1.00000	1
Lampung	0.00000	0.50000	0.50000	1
Kalteng	0.00000	0.50000	0.50000	1
Papua	0.00000	0.00000	0.50000	1
P. Barat	0.00000	0.00000	0.50000	1
NTT	1.00000	0.50000	0.00000	0
Kep.Babel	1.00000	0.50000	0.00000	0
Sulteng	1.00000	0.00000	0.00000	0
Bengkulu	1.00000	0.00000	0.00000	0
Kalut	1.00000	0.00000	0.00000	0
NTB	1.00000	0.00000	0.00000	0
Mal. Utara	1.00000	0.00000	0.00000	0
Sulbar	1.00000	0.00000	0.00000	0
SULTENG	0.50000	0.50000	0.00000	0
Aceh	0.50000	0.00000	0.00000	0
Jambi	0.50000	0.00000	0.00000	0
Maluku	0.50000	0.00000	0.00000	0
Gorontalo	0.50000	0.00000	0.00000	0

Sumber: Hasil Penelitian (2022)

Setelah pola ditemukan dalam data, langkah selanjutnya adalah menampilkan data dari 34 provinsi Indonesia dijadikan perhitungan k-means visualisasi jumlah cluster dari provinsi dari masing masing cluster terdapat pada gambar 3.



Sumber: Hasil Penelitian (2022)
 Gambar 3 kasus Cluster

Tabel 3 Informasi berdasarkan kasus cluster

Cluster	Keterangan
0	Nusa Tenggara Timur, Kepulauan Bangka Belitung, Sulawesi Tenggara, Bengkulu, Kalimantan Utara, Nusa Tenggara Barat, Maluku Utara, Sulawesi Barat, Sulawesi Tengah, Aceh, Jambi, Maluku, Gorontalo. Cluster ini memiliki total kasus tergolong rendah dengan jumlah 13 provinsi
1	Sumatera, Sumatera Barat, Kalimantan Selatan, Kepulauan Riau, Kalimantan Barat, Jawa Barat, DKI Jakarta, Banten, Daerah Istimewa Yogyakarta, Sumatera Utara, Bali, Sulawesi Utara, Sumatera Selatan, Lampung, Kalimantan Tengah, Papua dan Papua Barat. Cluster ini memiliki total kasus paling tinggi dengan jumlah 16 provinsi.
2	Jawa Timur, Kalimantan Timur, Sulawesi Selatan, Jawa Tengah, Riau. Cluster ini memiliki total kasus sedang dengan jumlah 5 provinsi.

Sumber: Hasil Penelitian (2022)

4. Evaluation

Menentukan jumlah cluster paling optimal, menggunakan evaluasi cluster dengan *Sum of Square Error* (SSE). Dari hasil Pengujian SSE yang mengalami penurunan paling besar pada k=3 dengan nilai SSE adalah 228913736548657.56. Menghasilkan nilai jarak saling membandingkan sesuai ketentuan. Nilai SSE dilihat pada tabel 4.

Tabel 4 Evaluation

Jumlah K	Nilai SSE
k: 1	2540576359776802.0
k: 2	631995018294042.8
k: 3	228913736548657.56
k: 4	109531299055175.58
k: 5	75318916274398.88
k: 6	52264312560320.89
k: 7	34884210231182.723

Sumber: Hasil Penelitian (2022)

5. Pembahasan

Dari hasil penelitian didapatkan bahwa data kasus Covid 19 di Indonesia yang terdapat 34 provinsi dari tanggal 1 Maret sampai dengan 11 Oktober 2020, cluster 0 merupakan cluster dengan jumlah 13 provinsi dikategorikan sebagai kasus tergolong rendah, cluster 1 merupakan cluster dengan jumlah 16 provinsi dikategorikan sebagai kasus paling tinggi, cluster 2 merupakan cluster dengan jumlah 5 provinsi dikategorikan sebagai kasus sedang.

IV. KESIMPULAN

Berdasarkan data yang dihasilkan yang diperoleh dengan menggunakan algoritma K-Means untuk mengelompokkan evaluasi data kasus Covid-19 di Indonesia dari beberapa jumlah kasus. Menggunakan metode elbow yang menghasilkan 3 cluster. Jumlah cluster yang ditentukan dengan metode elbow belum tentu menunjukkan kualitas hasil clustering yang baik. Nilai evaluasi menggunakan sum of squared error (SSE), yang memberikan nilai optimal K:3, yaitu 228913736548657.56.

V. REFERENSI

- Abdullah, D., Susilo, S., Ahmar, A. S., Rusli, R., & Hidayat, R. (2022). The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. *Quality and Quantity*, 56(3), 1283–1291. <https://doi.org/10.1007/s11135-021-01176-w>
- Adhitama, R., Burhanuddin, A., & Ananda, R. (2020). Penentuan Jumlah Cluster Ideal Smk Di Jawa Tengah Dengan Metode X-Means Clustering Dan K-Means Clusterin. *JIKO (Jurnal Informatika Dan Komputer)*, 3(1), 1–5. <https://doi.org/10.33387/jiko.v3i1.1635>
- Cahyo Wiguno, T., & Nataliani, Y. (2022). Penerapan k-Means Clustering Berdasarkan Analisis RFM Terhadap Segmentasi Pembeli untuk Meningkatkan Strategi CRM. *Jurnal Media Informatika Budidarma*, 6, 1871–1881. <https://doi.org/10.30865/mib.v6i4.4472>
- Ekasetya, V. A., & Jananto, A. (2020). Klusterisasi Optimal Dengan Elbow Method Untuk Pengelompokan Data Kecelakaan Lalu Lintas Di Kota Semarang. *Jurnal Dinamika Informatika*, 12(1), 20–28. <https://doi.org/10.35315/informatika.v12i1.8159>
- Hardiani, T. (2022). Analisis Clustering Kasus Covid 19 di Indonesia Menggunakan Algoritma K-Means. *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, 11(2), 156–165. <https://doi.org/10.23887/janapati.v11i2.45376>
- Isy Karima Fauzia, Budi Arif Dermawan, & Tesa Nur Padilah. (2020). Penerapan K-Means Clustering pada Penyakit Infeksi Saluran Pernapasan Akut (ISPA) di Kabupaten Karawang. *Jurnal Sistem Dan Informatika (JSI)*, 15(1), 81–87. <https://doi.org/10.30864/jsi.v15i1.350>
- Negari, N., & Eryando, T. (2021). Analisis Penerimaan Sistem Informasi Pencatatan dan Pelaporan Kasus Acceptance Model (TAM) di UPT Puskesmas Cipadung Kota Bandung Analysis of Receiving Information System Recording and Reporting of COVID-19 Case (Silacak Application Version 1.2.5). *Jurnal Bikfokes (Biostatistik, Kependudukan, Dan Informatikan Kesehatan)*, 19. <http://dx.doi.org/10.51181/bikfokes.v1i3.5297>
- Purwayoga, V. (2021). Optimasi Jumlah Cluster pada Algoritma K-Means untuk Evaluasi Kinerja Dosen. *Jurnal Informatika Universitas Pamulang*, 6(1), 118. <https://doi.org/10.32493/informatika.v6i1.9522>
- Putu, N., Merliana, E., & Santoso, A. J. (n.d.). *Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means*. ISBN: 978-979-3649-81-8
- Ruwandara, D., Jajuli, M., & Rizal, A. (2021). Analisis Algoritma K-Means Clustering Untuk Daerah Penyebaran Sampah di Kota Bekasi. *JOINS (Journal of Information System)*, 6(1), 56–63. <https://doi.org/10.33633/joins.v6i1.4085>
- Sindi, S., Ningse, W. R. O., Sihombing, I. A., R.H.Zer, F. I., & Hartama, D. (2020). Analisis Algoritma K-Medoids Clustering Dalam Pengelompokan Penyebaran Covid-19 Di Indonesia. *Jurnal Teknologi Informasi*, 4(1), 166–173. <https://doi.org/10.36294/jurti.v4i1.1296>
- Wiguna, H., Nugraha, Y., Rizka R, F., Andika, A., Kanggrawan, J. I., & Suherman, A. L.

- (2020). Kebijakan Berbasis Data: Analisis dan Prediksi Penyebaran COVID-19 di Jakarta dengan Metode Autoregressive Integrated Moving Average (ARIMA). *Jurnal Sistem Cerdas*, 3(2), 74–83. <https://doi.org/10.37396/jsc.v3i2.76>
- Winarta, A., & Kurniawan, W. J. (2021). Optimasi cluster k-means menggunakan metode elbow pada data pengguna narkoba dengan pemrograman python. *Jurnal Teknik Informatika Kaputama (JTIK)*, 5(1), 113–119.
- Zulfa, N., Auliya, R. I., & Zaenal, A. (2021). Analisis Data Mining Untuk Clustering Kasus Covid-19 Di Provinsi Lampung Dengan Algoritma K-Means. *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, 2(2), 100. <http://jim.teknokrat.ac.id/index.php/JTSI>. <https://doi.org/10.33365/jtsi.v2i2.868>