Scan
Jurnal Teknologi Informasi dan Komunikasi

ISSN (Print): 1978-0087
ISSN (Online): 2686-6099
Volume 19 Nomor 3, Oktober 2024
Halaman: 124-132

Penerapan Metode *K-Means* dalam Pengelompokan Jumlah Kasus Penderita *Covid-19* di Dunia

Muhamad Raihan Ramadhani Isworo*, Anya Ningrum Nur'afifah, Kesya Nursyahada, Ananda Azra Razali
Program Studi Infomatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Indonesia

Diterima: Juli, 2024 | Revisi: September, 2024 | Diterbitkan: Oktober 2024

DOI: https://doi.org/10.33005/scan.v19i3.5029

ABTSRAK

Penelitian untuk mengelompokkan jumlah kasus COVID-19 di dunia dengan menggunakan metode K-Means Klustering, yaitu teknik dalam data mining yang sangat efektif untuk mendeteksi pola dalam dataset yang besar. Data yang dipergunakan diambil dari situs Worldometers pada 13 Oktober 2024, mencakup 231 negara di seluruh dunia. Untuk menentukan jumlah kluster yang paling tepat, studi ini menerapkan metode Elbow, dengan perhitungan yang didasarkan pada nilai Sum of Squared Errors (SSE). Hasil analisis memperlihatkan penurunan SSE yang paling mencolok dari k=1 ke k=2, serta penurunan signifikan yang berlanjut hingga k=3. Setelah k=3, penurunan SSE mulai melambat, yang menunjukkan bahwa tiga kluster adalah jumlah yang paling tepat untuk pengelompokan ini. Selanjutnya, evaluasi kualitas pengelompokan menggunakan Silhouette Score menghasilkan nilai 0.9186, yang menunjukkan bahwa hasil klustering sangat baik, dengan objek-objek dalam satu kluster memiliki kemiripan yang tinggi dan terpisah dengan jelas dari kluster lain.

Kata Kunci: *K-Means Klustering, Elbow Method, Sum of Squared Errors (SSE), Data Mining.*

The Application of the K-Means Method in Clustering the Number of Covid-19 Cases Worldwide

ABSTRACT

This study focuses on clustering the number of Covid-19 cases worldwide using the K-Means clustering method, a highly effective data mining technique for detecting patterns in large datasets. The data was sourced from the Worldometers website on October 13, 2024, covering 231 countries globally. To determine the optimal number of clusters, the study employed the Elbow Method, with calculations based on the Sum of Squared Errors (SSE). The analysis revealed a significant drop in SSE from k=1 to k=2, followed by another notable decrease up to k=3. Beyond k=3, the reduction in SSE slowed down, indicating that three clusters were the most appropriate for this dataset. Furthermore, the clustering quality evaluation using the Silhouette Score resulted in a value of 0.9186, signifying excellent clustering results, with objects within each cluster showing high similarity and clear separation from other clusters.

Keywords: K-Means Clustering, Elbow Method, Sum of Squared Errors (SSE), Data Mining.

***Corresponding Author:**
Email : 21081010106@student.upnjatim.ac.id
Alamat : Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur 60294

PENDAHULUAN

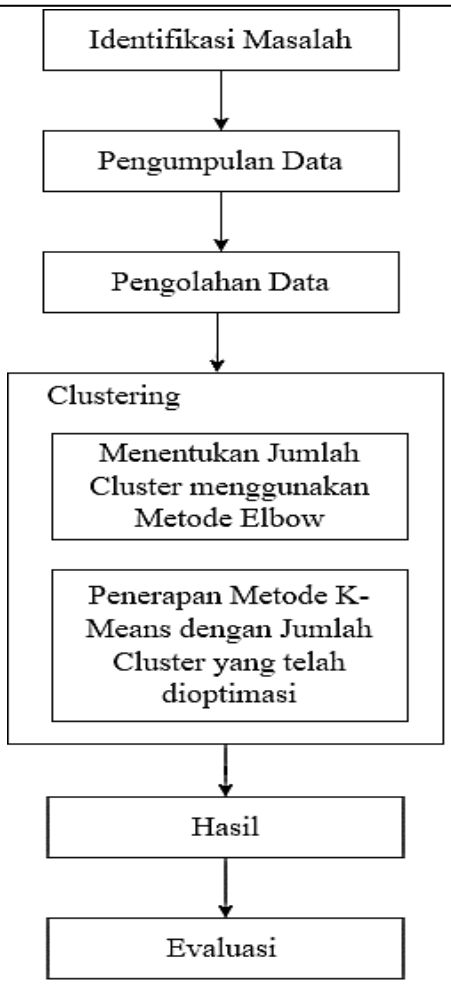
Pandemi *COVID-19* telah menjadi salah satu krisis kesehatan terbesar dalam sejarah modern, dengan dampak yang signifikan di seluruh dunia, termasuk di kawasan Asia. Virus *SARS-CoV-2*, penyebab utama penyakit *COVID-19*, menginfeksi sistem pernapasan manusia dan sudah menimbulkan jutaan kematian sejak pertama kali ditemukan [1]. dengan bertambahnya jumlah kasus yang tercatat, tantangan bagi sistem kesehatan global semakin rumit, menciptakan kebutuhan mendesak untuk menemukan strategi penanganan yang berbasis data dan efektif.

Pada 12 Maret 2020, Organisasi Kesehatan Dunia (*WHO*) secara resmi menyatakan *COVID-19* sebagai pandemi global, yang semakin mendesak perlunya merumuskan kebijakan penanggulangan yang tepat [2]. Dalam konteks ini, mengenali tren penyebaran dan pengaruh *COVID-19*, baik dari segi jumlah kasus maupun tingkat fatalitas, menjadi sangat krusial. Studi sebelumnya, seperti yang dilakukan Noviyanto (2020) dalam jurnalnya yang berjudul Penerapan *Data Mining* untuk Mengelompokkan Jumlah Kematian Penderita *COVID-19* Berdasarkan Negara di Benua Asia, menekankan pada tingginya angka kematian di negara-negara Asia. Noviyanto menerapkan teknik *data mining* untuk menganalisis pengelompokan angka kematian di negara-negara Asia, yang memberikan pemahaman tentang pola penyebaran virus di kawasan tersebut.

Namun, walaupun penelitian itu memberikan kontribusi yang signifikan, area geografinya hanya terbatas pada kawasan Asia. Dengan demikian, studi ini berupaya untuk memperluas jangkauan analisisnya secara global, dengan menitikberatkan pada pengelompokan negara-negara di seluruh dunia berdasar jumlah kasus *COVID-19* yang tercatat per 13 Oktober 2024. Melalui penerapan metode *K-Means Klustering*, studi ini bertujuan untuk mengelompokkan negara-negara berdasarkan tingkat keparahan pandemi mereka, yang diukur dengan total jumlah kasus *COVID-19* yang dilaporkan.

Agar mendapatkan hasil yang lebih tepat dan optimal, jumlah kluster dalam pengelompokan ini akan ditentukan dengan menggunakan Metode *Elbow*, yang dianalisis melalui perhitungan *Sum of Squared Errors (SSE).* Dengan demikian, studi ini tidak hanya mengelompokkan negara-negara ke dalam kategori dengan jumlah kasus tinggi, sedang, atau rendah, tetapi juga memberikan informasi lebih mendetail tentang pola penyebaran *COVID-19* di seluruh dunia.Hasil dari pengelompokan ini diharapkan mampu memberikan pemahaman yang lebih jelas mengenai tingkat keseriusan pandemi di berbagai negara.

Hasil dari pengelompokan ini diharapkan mampu memberikan pemahaman yang lebih jelas mengenai tingkat keseriusan pandemi di berbagai negara. Dengan pemahaman ini, diharapkan pemerintah serta lembaga kesehatan dapat merumuskan kebijakan penanganan yang lebih tepat dan berdasarkan informasi yang ada. Dengan analisis ini, diharapkan lahir pemahaman baru yang bermanfaat dalam menangani pandemi secara internasional dengan pendekatan yang lebih terstruktur dan berdasarkan fakta.



Gambar 1. Diagram alir kerangka kerja.

METODE PENELITIAN

Dalam bagian ini, penulis akan menjelaskan tahapan-tahapan terstruktur yang diterapkan dalam penelitian ini untuk mencapai target yang telah ditentukan. Metodologi yang digunakan dirancang secara sistematis agar proses penelitian dapat berjalan dengan lancar dan dapat menjadi acuan bagi peneliti berikutnya. Dengan menerapkan metodologi ini, diharapkan hasil yang diperoleh sesuai dengan sasaran penelitian dan dapat dilaksanakan dengan lebih efisien. Proses penelitian ini mencakup beberapa langkah utama yang saling terkait.

Proses Alur diagram alir pada Gambar 1 memiliki penjelasan sebagai berikut.

Identifikasi Masalah

Langkah awal adalah mengenali permasalahan utama, yakni bertambahnya jumlah kasus COVID-19 dan angka kematian akibat virus tersebut. Studi sebelumnya lebih berorientasi pada negara-negara di Asia. Studi ini selanjutnya diperluas untuk meneliti dampak COVID-19 secara internasional. Diharapkan bahwa ini dapat memberikan pemahaman yang lebih mendalam mengenai penyebaran dan efek dari pandemi.

Pengumpulan Data

Data untuk penelitian ini diperoleh dari situs web https://www.worldometers.info/coronavirus/ yang menyajikan data mengenai penyebaran COVID-19 di seluruh dunia. Data yang diperoleh mencakup jumlah total kasus (Total Cases) yang dilaporkan di 231 negara per 13 Oktober 2024. Pengumpulan data ini dilaksanakan secara daring dan melibatkan atribut utama yaitu jumlah kasus yang tercatat di setiap negara.

Pengolahan Data

Pada tahap ini dilakukan pembersihan data (data cleaning) untuk mengatasi masalah seperti data yang hilang (missing values) atau duplikasi data. Lalu data dilakukan normalisasi data untuk menyelaraskan skala data agar data memiliki tentang nilai yang sama [3].

Penentuan Jumlah Kluster Menggunakan Elbow Method

Setelah data disiapkan, langkah selanjutnya adalah menetapkan jumlah kluster yang ideal untuk pengelompokan dengan menggunakan metode Elbow Method. Pada tahap ini, perhitungan Sum of Squared Errors (SSE) dilakukan untuk setiap jumlah kluster yang diuji. SSE yang lebih kecil menunjukkan kelompok yang lebih efektif [4].

Penerapan Metode K-Means

Setelah mendapatkan jumlah kluster yang ideal, metode K-Means Klustering diterapkan pada data untuk mengategorikan negara-negara berdasarkan jumlah kasus COVID-19 yang tercatat [5].

Hasil dan Visualisasi

Setelah proses pengelompokan selesai, hasilnya akan disajikan dalam bentuk visual, seperti diagram atau peta, agar lebih mudah memahami distribusi negara berdasar jumlah kasus COVID-19. Visualisasi ini bertujuan untuk menyajikan gambaran yang jelas tentang pola penyebaran pandemi di seluruh dunia.

Evaluasi

Untuk mengevaluasi kualitas hasil klustering, digunakan metode Silhouette Score. Silhouette Score adalah metrik yang mengukur sejauh mana setiap data dalam satu kluster lebih mirip dengan data lain dalam kluster yang sama dibandingkan dengan data dalam kluster lain. Nilai Silhouette berkisar antara -1 hingga 1, dengan nilai yang lebih tinggi menunjukkan klustering yang lebih baik [6]. Evaluasi dilakukan dengan interpretasi sebagai berikut:

Tabel 1.
Interpretasi Evaluasi Silhouette Score

Interval Silhouette Score	Kategori Evaluasi	Interpretasi
0,7 < s ≤ 1	Sangat Baik	Klustering sangat jelas, data dalam kluster sangat mirip dan jauh dari kluster lain.
0.50 < s ≤ 0.70	Baik	Klustering baik, data dalam kluster cukup terpisah dengan jelas.
0.30 < s ≤ 0.50	Cukup Baik	Klustering cukup baik, tetapi ada beberapa overlap antar kluster.
0.00 < s ≤ 0.30	Kurang Baik	Klustering kurang jelas, data dalam kluster cenderung bercampur.
s < 0.00	Buruk	Klustering gagal, data salah penempatan dalam kluster yang tidak sesuai.

Sumber: Data Diolah

HASIL DAN PEMBAHASAN

Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data jumlah kasus COVID-19 dari berbagai negara di dunia yang diambil pada tanggal 13 Oktober 2024. Data yang digunakan dalam penelitian ini menggunakan data total kasus Covid 19 di seluruh dunia.

Pengolahan Data

Pada dataset yang digunakan memiliki 231 baris tanpa nilai yang hilang dan tidak ada duplikasi dalam data set. Selanjutnya dilakukan menormalisasikan data dengan rentang nilai 0 hingga 1 pada kolom Total Cases. Berikut merupakan sampel hasil normalisasi data :

Tabel 2.
Sample Data

No	Negara	Total Kasus
1	USA	111,820,082
2	India	45,035,393
3	France	40,138,560
4	Germany	38,828,995
5	Brazil	38,743,918

Sumber *: Website Worldmeters*

Tabel 3.
Hasil Normalisasi Data

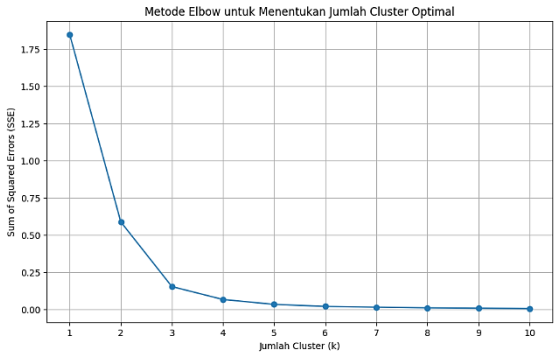
No	Negara	Total Kasus Asli	Hasil Normalisasi
1	USA	111820082.0	1.000000
2	India	45035393.0	0.402749
3	France	40138560.0	0.358957
4	Germany	38828995.0	0.347245
5	Brazil	38743918.0	0.346484

Sumber: Data Diolah

Tabel 4.
Hasil Perhitungan *SSE*

No	Jumlah Kluster (K)	Nilai SSE	Selisih SSE
1	1	1.847276	0.000000
2	2	0.587562	1.259714
3	3	0.152757	0.434805
4	4	0.066017	0.086741
5	5	0.033221	0.032795

Sumber: Data Diolah



Gambar 2. Grafik *Elbow Method*

Penentuan Jumlah Kluster Menggunakan Elbow Method

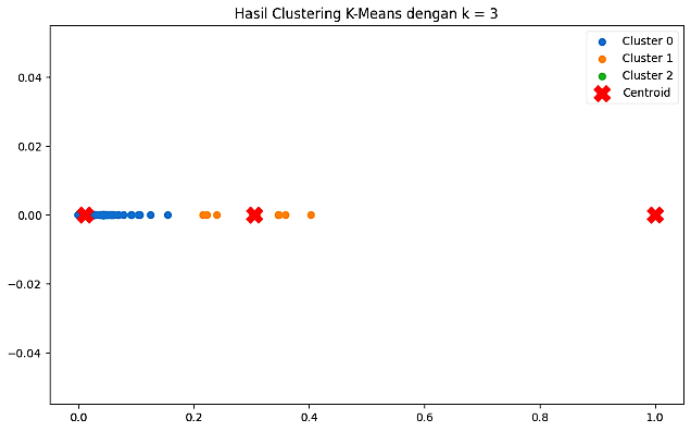
Nilai SSE (Sum of Squared Errors) menunjukkan total jarak kuadrat antara data dalam kluster dan centroid masing-masing. Semakin kecil nilai SSE, semakin baik kluster terbentuk. Selisih SSE menunjukkan pengurangan SSE saat menambah jumlah kluster. Penurunan besar di awal menunjukkan bahwa menambah kluster memberikan peningkatan signifikan dalam kualitas klusterisasi, tetapi setelah titik tertentu, penurunan menjadi kecil. Hasil perhitungan SSE (Sum of Squared Errors) yang akan digunakan untuk menentukan jumlah kluster K yang optimal :

Pada tabel, terlihat bahwa penurunan SSE terbesar terjadi dari k = 1 ke k = 2 (selisih SSE = 1.259714) dan penurunan yang signifikan masih terlihat hingga k = 3 (selisih SSE = 0.434805). Setelah k = 3, penurunan nilai SSE menjadi lebih kecil dan lebih stabil.

Titik elbow terlihat sekitar k = 3 karena setelah itu, penurunan SSE tidak signifikan lagi. Maka, jumlah kluster optimal (k) yang dapat dipilih adalah 3.

Tabel 5.
Hasil *K-Means*

No	Negara	Total Kasus Asli	Kluster
1	USA	111820082.0	2
2	India	45035393.0	1
3	France	40138560.0	1
4	Germany	38828995.0	1
5	Brazil	38743918.0	1



Gambar 3. Visualisasi Kluster

Penerapan K-Means

Setelah menentukan jumlah kluster optimal menggunakan metode Elbow, K-Means diterapkan dengan jumlah kluster tersebut untuk mengelompokkan data. Jumlah kluster optimal dipilih berdasarkan titik siku pada grafik Elbow, yang menunjukkan penurunan SSE yang signifikan sebelum stabil. Dengan jumlah kluster ini, K-Means membagi data ke dalam kelompok yang memiliki karakteristik serupa, menghasilkan segmentasi yang optimal sesuai dengan pola distribusi data.

Hasil klusterisasi K-Means dengan 3 kluster mengelompokkan negara-negara berdasarkan jumlah kasus COVID-19 yang dilaporkan. Kluster 0 terdiri dari negara-negara dengan jumlah kasus menengah seperti Turkey, Spain, dan Australia, menunjukkan penyebaran yang cukup tinggi namun tidak ekstrem. Kluster 1 mencakup negara-negara dengan jumlah kasus tinggi seperti India, France, dan Germany, yang memiliki kasus lebih besar namun masih lebih rendah dibanding kluster tertinggi. Sementara itu, Kluster 2 hanya berisi Amerika Serikat, yang memiliki jumlah kasus sangat tinggi dan menjadi outlier dibandingkan negara lain.

Hasil dan Visualisasi

Visualisasi ini menampilkan hasil pengelompokan (klusterisasi) data menggunakan algoritma K-Means. Grafik yang dihasilkan berupa scatter plot dengan sumbu x mewakili 'Total Cases Normalisasi' dan sumbu y diatur ke nol untuk semua titik data agar mudah dilihat dalam satu garis.

Berikut penjelasan elemen-elemen penting dalam gambar.

- Titik-titik Data; setiap titik pada grafik mewakili suatu negara dengan 'Total Cases Normalisasi' sebagai posisinya pada sumbu x. Titik-titik ini diwarnai berdasarkan kluster yang ditetapkan oleh algoritma K-Means.
- Kluster; Titik-titik data yang memiliki warna sama termasuk dalam kluster yang sama. Ini menunjukkan bahwa negara-negara tersebut memiliki pola 'Total Cases' yang serupa. Jumlah kluster (k) ditentukan sebelumnya menggunakan metode Elbow.
- Centroid; Centroid setiap kluster diwakili oleh tanda 'X' berwarna merah. Centroid menandai pusat dari setiap kluster dan merupakan titik rata-rata dari semua titik data dalam kluster tersebut.

Evaluasi

Penerapan metode Elbow untuk menentukan jumlah kluster menghasilkan nilai k optimal yaitu 3 kluster. Untuk memvalidasi hasil klusterisasi ini, dilakukan evaluasi menggunakan metode Silhouette Coefficient untuk mengukur kualitas pengelompokan. Silhouette Score untuk jumlah kluster optimal (k = 3) diperoleh sebesar 0.9186, yang menunjukkan bahwa kualitas pengelompokan sangat baik. Nilai Silhouette Score berkisar antara -1 hingga 1, di mana nilai mendekati 1 menunjukkan bahwa objek dalam satu kluster memiliki kemiripan yang tinggi dengan objek lainnya dalam kluster yang sama (koheasi), serta terpisah dengan jelas dari kluster lainnya (pemisahan). Dengan nilai 0.9186, ini mengindikasikan bahwa hasil pengelompokan memiliki struktur yang baik dan sesuai dengan tujuan analisis. Oleh karena itu, hasil evaluasi ini mengonfirmasi bahwa jumlah kluster yang diperoleh melalui metode Elbow (k = 3) adalah pilihan yang tepat dan valid.

SIMPULAN

Berdasarkan hasil penelitian mengenai penerapan metode K-Means dalam pengelompokan jumlah kasus COVID-19 di dunia, dapat disimpulkan bahwa metode K-Means Klustering berhasil mengelompokkan negara-negara ke dalam tiga kluster, yaitu kluster dengan jumlah kasus rendah, kluster dengan jumlah kasus sedang, dan kluster dengan jumlah kasus tinggi. Pengelompokan ini memberikan gambaran yang lebih jelas mengenai tingkat keparahan pandemi di berbagai negara, sehingga membantu dalam memahami pola penyebaran COVID-19 secara global. Pemilihan jumlah kluster yang optimal menggunakan metode Elbow menunjukkan bahwa tiga kluster adalah jumlah yang paling sesuai. Hal ini didasarkan pada analisis nilai Sum of Squared Errors (SSE), di mana penurunan SSE terbesar terjadi dari k = 1 ke k = 2 (selisih SSE = 1.259714) dan penurunan yang signifikan masih terlihat hingga k = 3 (selisih SSE = 0.434805), sementara setelah k = 3, penurunan nilai SSE menjadi lebih kecil dan stabil. Meskipun dilakukan percobaan dengan jumlah kluster yang lebih sedikit, hasilnya tidak seefisien tiga kluster, karena terdapat negara dengan jumlah kasus yang sangat tinggi yang berbeda jauh dari kluster lainnya. Selain itu, proses normalisasi dan pembersihan data sangat penting untuk memastikan kualitas data yang digunakan, sehingga menghasilkan kluster yang lebih akurat dan relevan. Hasil pengelompokan ini diharapkan dapat memberikan wawasan yang bermanfaat bagi pemerintah dan lembaga kesehatan dalam menyusun strategi mitigasi dan penanganan pandemi secara lebih tepat, berdasarkan tingkat keparahan kasus di masing-masing negara.

DAFTAR PUSTAKA

[1]. Noviyanto, R. (2020). Penerapan data mining dalam mengelompokkan jumlah kematian penderita COVID-19 berdasarkan negara di benua Asia. Jurnal Teknik Informatika, 12(2), 45-60.

[2]. World Health Organization (WHO). (2020). COVID-19 Dashboard. Diakses melalui https://www.who.int

[3]. Tan, P. N., Steinbach, M., & Karpatne, A. (2020). Introduction to Data Mining (3rd ed.). Boston, MA: Pearson Education.

[4]. Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129-137.

[5]. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (pp. 281-297). Berkeley, CA: University of California Press.

[6]. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of kluster analysis. Journal of Computational and Graphical Statistics, 1(1), 53-65.