Seminar Nasional Sains Data 2025 (SENADA 2025)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Penerapan *Streaming K-Means Clustering* Menggunakan Sparklyr untuk Penentuan Nilai K Terbaik pada Data US *Covid* Surat Kabar NY Times

Rafi Fadhlillah[1], Muhammad Fathir Fadillah[2], Mujadid Choirus Surya[3], Muhammad Farhan[4], Catherine Sinaga[5], Luluk Muthoharoh[6], Ardika Satria[7], Rizty Maulida Badri[8]

[1,2,3,4,5,6,7] *Program Studi Sains Data, Fakultas Sains, Institut Teknologi Sumatera*
[1]rafi.121450143@student.itera.ac.id
[2]muhammad.121450098@student.itera.ac.id
[3]mujadid.121450015@student.itera.ac.id
[4]muhammad.121450044@student.itera.ac.id
[5]catherine.121450072@student.itera.ac.id
[6]luluk.muthoharoh@sd.itera.ac.id
[7]ardika.satria@sd.itera.ac.id
[8] *Jurusan Fisika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan*
[8]riztymaulidabadri@gmail.com
Corresponding author email: luluk.muthoharoh@sd.itera.ac.id

Abstract: *Big data processing is a major focus in modern data analysis, especially in the context of the COVID-19 pandemic. Researchers propose an approach to determine the best K value in the Streaming K-Means Clustering algorithm using Sparklyr in the Apache Spark environment. Researchers apply this approach to COVID-19 data from the NY Times Newspaper, with the development of the Silhouette Coefficient evaluation method. Experimental results show that this approach is able to provide reliable and analytically relevant clustering results. This makes an important contribution to the understanding of the dynamics of the COVID-19 pandemic, strengthening the potential of real-time data analysis from news sources such as the NY Times Newspaper in providing valuable insights for stakeholders.*
Keywords: *Apache Spark, COVID-19, Sparklyr, Streaming K-Means Clustering Algorithm*

Abstrak: Pemrosesan data besar menjadi fokus utama dalam analisis data modern, terutama dalam konteks pandemi *COVID-19*. Peneliti mengusulkan pendekatan untuk menentukan nilai K terbaik dalam algoritma Streaming *K-Means Clustering* dengan menggunakan Sparklyr di lingkungan Apache Spark. Peneliti menerapkan pendekatan ini pada data *COVID-19* dari Surat Kabar NY Times, dengan pengembangan metode evaluasi *Silhouette Coefficient*. Hasil eksperimen menunjukkan bahwa pendekatan ini mampu memberikan hasil klasterisasi yang andal dan relevan secara analitis. Hal ini memberikan kontribusi penting untuk pemahaman dinamika pandemi *COVID-19*, memperkuat potensi analisis data *real-time* dari sumber-sumber berita seperti Surat Kabar NY Times dalam memberikan wawasan berharga bagi pemangku kepentingan.
Kata kunci: Algoritma Streaming K-Means Clustering, Apache Spark, COVID-19, Sparklyr

I. PENDAHULUAN

Dalam era *Big Data*, data berperan sebagai aset strategis dalam mendukung pengambilan keputusan berbasis fakta. Salah satu fenomena global yang memperlihatkan urgensi pemanfaatan data secara optimal adalah pandemi *COVID-19*. Sejak awal penyebarannya, *COVID-19* telah menginfeksi ratusan juta orang dan menyebabkan jutaan kematian di seluruh dunia. Di Amerika Serikat sendiri, hingga akhir 2022, tercatat lebih dari 100 juta kasus positif dan lebih dari satu juta kematian akibat virus ini [1]. Data harian mengenai jumlah kasus dan kematian akibat *COVID-19* yang dikumpulkan dan dipublikasikan oleh media seperti *The New York Times* (NYT) menjadi salah satu sumber informasi penting untuk memahami dinamika penyebaran pandemi. Data ini bersifat *real-time* dan terus diperbarui, sehingga sangat relevan untuk dianalisis secara cepat dan efisien guna mendukung kebijakan pencegahan dan pengendalian wabah secara responsif.

Meskipun ketersediaan data semakin melimpah, tantangan utama yang masih dihadapi masyarakat dan pembuat kebijakan adalah bagaimana mengelola dan menganalisis data tersebut secara tepat

waktu[2], [3]. Banyak data yang tersedia dalam format *streaming* dan berdimensi besar belum dimanfaatkan secara optimal karena keterbatasan metode analisis konvensional yang tidak mampu menangani data secara *real-time*[4]. Selain itu, pemahaman mengenai persebaran geografis dan tingkat keparahan kasus di berbagai wilayah belum terpetakan dengan baik secara otomatis. Ketidakmampuan dalam mengidentifikasi wilayah dengan tingkat penyebaran tinggi secara cepat dapat menghambat efektivitas intervensi kebijakan kesehatan masyarakat[5], [6].

Di antara berbagai metode analisis, algoritma clustering, seperti *K-Means*, telah terbukti efektif dalam mengungkap pola-pola tersembunyi dalam dataset [7], [8]. Namun, menentukan jumlah klaster yang optimal (nilai K) tetap menjadi tantangan utama dalam penggunaan algoritma ini. Penelitian ini bertujuan untuk menguji coba algoritma *K-Means Clustering* pada data *COVID-19* yang dilaporkan oleh NY Times menggunakan *Sparklyr*, sebuah kerangka kerja yang memfasilitasi pemrosesan data secara distribusi dan paralel melalui Apache Spark yang memungkinkan analisis yang cepat dan efisien pada dataset yang besar [7], [9]. Sejumlah penelitian sebelumnya telah membuktikan efektivitas algoritma clustering, khususnya *K-Means*, dalam mengelompokkan wilayah berdasarkan kemiripan karakteristik penyebaran penyakit [9], [10]. Namun, sebagian besar studi masih berbasis data statis dan tidak mempertimbangkan dinamika data yang terus berubah. Di sisi lain, teknologi *Big Data* seperti Apache Spark menawarkan kemampuan pemrosesan data secara terdistribusi dan paralel, sehingga mampu menangani data dalam skala besar secara efisien. Sparklyr, sebagai penghubung antara bahasa pemrograman R dan Apache Spark, memungkinkan integrasi metode analitik dengan performa tinggi. Oleh karena itu, kombinasi antara algoritma *K-Means* dan pemrosesan *Big Data* melalui Sparklyr menjadi pendekatan yang sangat menjanjikan untuk analisis data *COVID-19*.

Penelitian ini bertujuan untuk menguji coba algoritma *K-Means Clustering* pada data *streaming COVID-19* yang diperoleh dari *The New York Times*, dengan dukungan pemrosesan paralel menggunakan Sparklyr. Kebaruan penelitian ini terletak pada integrasi metode *streaming clustering*, yang memungkinkan model diperbarui secara dinamis seiring masuknya data baru. Model ini tidak hanya mengelompokkan wilayah berdasarkan jumlah kasus dan kematian secara *real-time*, tetapi juga melakukan evaluasi kualitas klaster menggunakan *Silhouette Coefficient* untuk menentukan jumlah klaster optimal. Penelitian ini tidak hanya menawarkan pendekatan yang cepat dan adaptif, tetapi juga mampu memberikan informasi yang relevan untuk mendukung pengambilan keputusan kebijakan kesehatan secara tepat waktu dan berbasis data. Dengan demikian, model dapat secara adaptif mengidentifikasi pola penyebaran baru dan menghasilkan informasi yang lebih aktual [11]. Pendekatan ini diharapkan dapat menghasilkan informasi yang lebih cepat, akurat, dan responsif terhadap perubahan pola penyebaran *COVID-19*, sehingga menjadi landasan yang kokoh bagi pengambilan keputusan kebijakan pencegahan dan penanggulangan pandemi secara efektif.

II. METODE

2.1. Data

Data *COVID-19* yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari sumber terpercaya, yaitu The Humanitarian Data Exchange (HDX). HDX menyediakan akses terbuka ke data kemanusiaan dari berbagai sumber, termasuk data *COVID-19*. Dalam studi ini, data yang digunakan berasal dari The New York Times *COVID-19* Dataset yang tersedia di HDX dan dapat diakses melalui tautan berikut: https://data.humdata.org/dataset/nyt-covid-19-data. Dataset ini awalnya disediakan oleh The New York Times melalui repositori GitHub di tautan https://github.com/nytimes/covid-19-data. Data tersebut berisi laporan harian mengenai total kasus dan total kematian *COVID-19* di berbagai wilayah Amerika Serikat, dimulai dengan kasus virus corona pertama yang dilaporkan di Negara Bagian Washington pada 21 Januari 2020, sehingga sangat cocok untuk analisis epidemiologi. Data tersebut kemudian dimasukkan ke dalam lingkungan

Apache Spark menggunakan Sparklyr, yang memungkinkan pemrosesan data secara distribusi dan paralel, yang sangat bermanfaat dalam menangani volume data yang besar dan kompleks [6].

2.2. Proses Preprocessing

Tahap awal dalam analisis data *COVID-19* adalah proses preprocessing, dengan fokus utama pada penanganan nilai kosong (missing values) pada kolom total kasus dan total kematian. Penanganan missing values penting agar hasil clustering yang dilakukan tetap akurat dan valid[12]–[14]. Dengan membersihkan data yang kosong, dataset menjadi konsisten dan siap digunakan dalam algoritma *K-Means*, sehingga setiap data memiliki informasi lengkap mengenai total kasus dan total kematian untuk membentuk pola klaster penyebaran *COVID-19* secara optimal[15].

2.3. Implementasi Streaming K-Means Clustering

Klasterisasi dikenal sebagai salah satu metode *unsupervised* yang sangat bermanfaat dalam analisis *Big Data*, terutama untuk mengelompokkan data tanpa label berdasarkan kemiripan fitur[16], [17]. Selain berfungsi sebagai alat analisis, klasterisasi juga sering digunakan sebagai tahap pra-pemrosesan untuk mereduksi dimensi data sehingga memudahkan algoritma pembelajaran selanjutnya [18], [19]. Pada penelitian ini, algoritma *K-Means Clustering* diimplementasikan menggunakan Sparklyr dalam lingkungan Apache Spark, yang mendukung pemrosesan data secara paralel dan terdistribusi sehingga dapat menangani volume data besar dengan efisien.

Metode *streaming* diterapkan agar proses clustering dapat dilakukan secara *real-time*, di mana jumlah klaster (K) dan posisi centroid dapat diperbarui otomatis setiap kali data baru masuk[20]. Hal ini penting untuk menjaga relevansi analisis terhadap dinamika data *COVID-19* yang terus berubah setiap hari. Langkah pertama dalam *K-Means* adalah menentukan centroid atau titik pusat dari setiap klaster berdasarkan rata-rata koordinat data yang tergabung dalam klaster tersebut[21], [22]. Pada Persamaan (1), K^i merepresentasikan centroid ke-*i*, yaitu titik pusat klaster. *M* adalah jumlah total titik data yang saat ini ditugaskan ke dalam klaster tersebut, menunjukkan seberapa banyak data yang termasuk dalam kelompok yang sedang dihitung centroidnya. Sementara itu, x_j adalah setiap titik data individual (ke-*j*) yang menjadi anggota dari klaster tersebut.

$$K^i = \frac{1}{M}\sum_{j=1}^{M} x_j \qquad (1)$$

Setelah centroid ditentukan, langkah berikutnya adalah mengalokasikan setiap titik data ke centroid terdekat dengan mengukur jaraknya menggunakan rumus Euclidean[23]. Pada persamaan (2), *d* merupakan jarak Euclidean antara titik data dan centroid, *D* adalah jumlah dimensi fitur (misalnya jumlah kasus dan kematian), x_j adalah nilai koordinat titik data pada dimensi ke-*j*, dan c_j adalah nilai koordinat centroid pada dimensi yang sama. Selisih tiap dimensi dikuadratkan untuk memastikan nilai positif dan memberi bobot lebih pada perbedaan besar, lalu dijumlahkan dan diakarkan untuk mendapatkan jarak sebenarnya. Titik data akan ditempatkan pada klaster dengan centroid yang memiliki jarak terkecil, sehingga membentuk kelompok data yang homogen.

$$d = \sqrt{\sum_{j=1}^{D}(x_j - c_j)^2} \qquad (2)$$

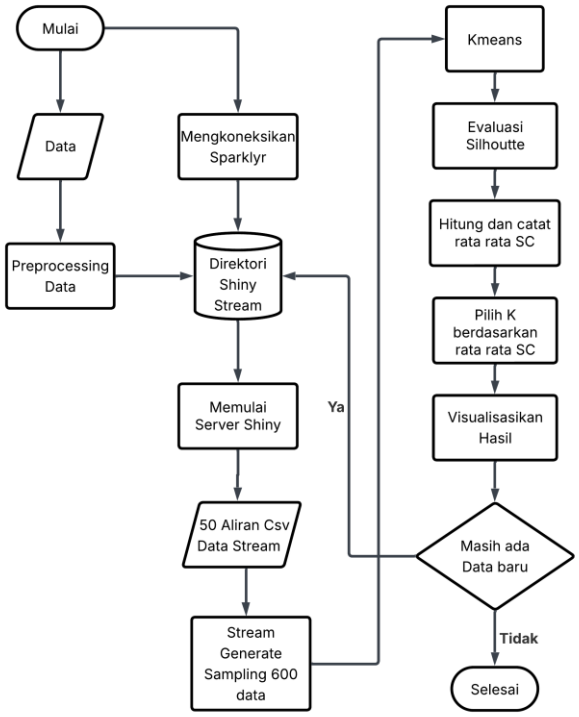
2.4. Penentuan Nilai K Terbaik dan Evaluasi Hasil

Metode evaluasi *Silhouette Coefficient* digunakan untuk menentukan nilai K yang paling tepat dalam proses clustering [9]. *Silhouette Coefficient* mengukur seberapa baik setiap titik data cocok dengan klaster yang ditetapkan. Nilai K dengan *Silhouette Coefficient* tertinggi dipilih sebagai nilai optimal, karena menunjukkan pembagian klaster yang paling jelas dan konsisten [24]. Nilai s(*i*) adalah *silhouette score* untuk sampel ke-i, berkisar dari -1 hingga +1, menunjukkan seberapa baik sampel cocok dalam klasternya. *a(i)* merepresentasikan rata-rata jarak antara sampel ke-*i* dan semua titik lain dalam klaster yang sama (mengukur kohesi). Sebaliknya, *b(i)* adalah rata-rata jarak antara sampel ke-*i* dan semua titik dalam klaster tetangga terdekat, yang berfungsi sebagai ukuran pemisahan. Pembagi *max(a(i),b(i))* sebagai faktor normalisasi untuk memastikan bahwa nilai *s(i)* berada dalam rentang yang konsisten, yaitu antara -1 dan +1. Nilai positif mendekati +1 menandakan bahwa titik data cocok dengan klasternya, sedangkan nilai negatif mendekati -1 menunjukkan kemungkinan salah klaster. Dengan metode ini, peneliti dapat memilih jumlah klaster *K* yang paling sesuai untuk merepresentasikan pola data secara akurat.

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \qquad (3)$$

2.5. Proses Penelitian

Proses penentuan jumlah optimal cluster dimulai dengan inisialisasi nilai awal k (jumlah cluster) dari 2 hingga 9. Selanjutnya, data stream dibaca dan diproses dengan menghapus nilai-nilai NA, memilih data relevan (kasus dan kematian), dan menghubungkannya ke Sparklyr untuk analisis data besar. Data yang telah dibersihkan kemudian disimpan ke file CSV. Pada tahap pembuatan cluster, data dibagi menjadi 50 aliran CSV, setiap aliran disampel sebanyak 600 baris (sesuai spesifikasi perangkat keras). Sampel CSV ini diproses dan dimasukkan kembali ke aliran CSV, lalu aplikasi Shiny UI dijalankan dengan server yang terkoneksi dengan data. *Silhouette Coefficient* dihitung untuk setiap cluster dan dicatat oleh aplikasi Shiny UI. Rerata *Silhouette Coefficient* dihitung untuk setiap nilai k, dan jumlah cluster yang menghasilkan rerata tertinggi dianggap sebagai jumlah optimal cluster. Proses ini memastikan cluster yang dihasilkan memiliki kualitas terbaik berdasarkan metrik *Silhouette Coefficient*. Langkah-langkah yang dilakukan dalam penelitian ini disajikan dalam bentuk diagram alir pada **Gambar 1** berikut:



Gambar 1. Flowchart Proses Penelitian

III. HASIL DAN PEMBAHASAN

3.1. Ingesti Data

Data *COVID-19* diambil dari file CSV yang berisi informasi terkait kasus dan kematian di berbagai wilayah di AS. Proses ingest melibatkan pembacaan file tersebut, pembersihan data dengan menghilangkan baris jika memiliki nilai kosong, dan penyimpanan kembali data yang telah dibersihkan untuk keperluan analisis lebih lanjut. Langkah ini memastikan data yang digunakan dalam analisis memiliki integritas dan kualitas yang baik.

3.2. Integrasi Spark Streaming

Library sparklyr digunakan untuk menghubungkan R beserta dplyr ke instansi Spark lokal. Ini memungkinkan aplikasi untuk melakukan *streaming* data dan *clustering* secara *real-time*. *Clustering K-means* dan perhitungan skor Silhouette dilakukan menggunakan Spark, yang mampu menangani pemrosesan data dalam skala besar dengan efisien. Integrasi ini memastikan bahwa aplikasi dapat bekerja dengan performa tinggi bahkan ketika dihadapkan dengan volume data yang besar. Kemampuan *streaming* dari Spark dimanfaatkan untuk memungkinkan aplikasi menangani input data yang kontinu dan memperbarui hasil klaster secara *real-time* [25]. Berikut adalah skor Silhouette hasil dari *streaming* dengan percobaan K2 sampai K9 seperti pada Tabel 1.

Untuk menentukan jumlah klaster optimal dalam proses *streaming clustering* terhadap data *COVID-19*, digunakan metode ***Silhouette Coefficient*** sebagai ukuran evaluasi. *Silhouette Coefficient* mengukur sejauh mana suatu objek mirip dengan klasternya sendiri (kohesi) dibandingkan dengan klaster lain (separasi). Nilai koefisien ini berada pada rentang -1 hingga 1, di mana nilai mendekati 1 menunjukkan klaster yang terbentuk sangat baik dan terpisah secara jelas [26]. Tabel 1 menunjukkan nilai rata-rata (*AVG*) dan standar deviasi (*STD*) dari *Silhouette Coefficient* untuk jumlah klaster K antara 2 hingga 9. Hasil terbaik diperoleh pada **K=2** dengan nilai Silhouette rata-rata sebesar **0,98478** dan standar deviasi **0,02291**, mengindikasikan bahwa dua

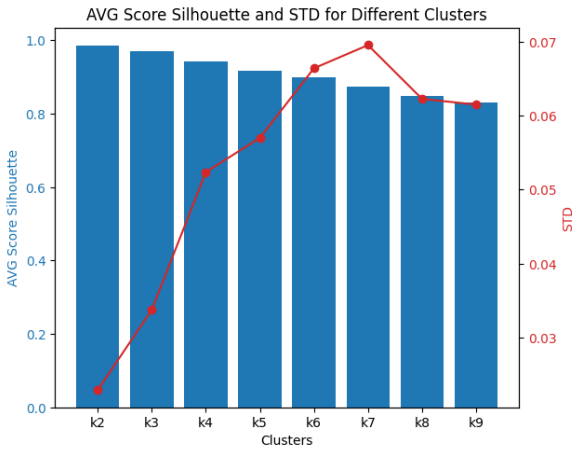
E-ISSN 2808-5841
P-ISSN 2808-7283

klaster memberikan pemisahan yang paling stabil dan kohesif. Nilai koefisien secara konsisten menurun seiring bertambahnya jumlah klaster, sementara nilai standar deviasi cenderung meningkat, menunjukkan menurunnya kualitas dan stabilitas klaster.

Tabel 1. Hasil *streaming Silhouette Coefficient* untuk setiap K

K	AVG	STD
K2	0,98478	0,02291
K3	0,97006	0,03376
K4	0,94279	0,05238
K5	0,91634	0,05701
K6	0,89879	0,06642
K7	0,8749	0,06955
K8	0,84732	0,06227
K9	0,83073	0,06151

Gambar 2 mendukung hasil ini melalui visualisasi gabungan *bar chart* dan *line chart*. Grafik batang menunjukkan rata-rata Silhouette Score untuk setiap nilai K, sedangkan garis menunjukkan nilai standar deviasi. Hasilnya memperlihatkan pola penurunan performa clustering saat nilai K lebih dari 2. Dengan demikian, dapat disimpulkan bahwa **jumlah klaster optimal untuk data *streaming COVID-19* yang dianalisis adalah K=2**. Penggunaan *Silhouette Coefficient* sebagai metode evaluasi dalam konteks clustering data *streaming* telah didukung oleh berbagai penelitian sebelumnya, karena kemampuannya yang efisien dalam mengevaluasi kualitas klaster secara internal tanpa perlu pembanding eksternal [27]. Kombinasi algoritma *K-Means* dengan pendekatan *streaming* berbasis Apache Spark melalui Sparklyr juga memungkinkan pemrosesan data secara paralel dan *real-time*, sehingga sangat sesuai untuk menganalisis data berskala besar dan terus-menerus masuk seperti data *COVID-19* dari The New York Times [11].



Gambar 2. Plot Hasil *streaming Silhouette Coefficient* untuk setiap K

Pendekatan yang digunakan dalam penelitian ini memanfaatkan kekuatan Spark untuk pemrosesan data skala besar dan Shiny untuk visualisasi interaktif. Aplikasi ini memungkinkan

pengguna untuk memilih variabel yang akan dianalisis serta menentukan jumlah klaster yang diinginkan. Pada penelitian ini menggunakan variabel kasus pada sumbu x dan kematian pada sumbu y dengan k=2 yang merupakan k terbaik yang diperoleh dari hasil sebelumnya, hasilnya dapat dilihat pada Gambar 3. Gambar 3 menunjukkan visualisasi hasil klasterisasi data *COVID-19* menggunakan algoritma *K-Means Clustering* dengan jumlah klaster optimal yaitu K=2, sebagaimana ditentukan sebelumnya berdasarkan nilai tertinggi dari *Silhouette Coefficient*. Masing-masing plot merepresentasikan hasil clustering berdasarkan variabel jumlah kasus (*cases*) pada sumbu X dan jumlah kematian (*deaths*) pada sumbu Y, untuk data *streaming* yang terus diperbarui.

Setiap warna pada grafik mewakili satu klaster, dan titik-titik pada plot merepresentasikan wilayah atau waktu tertentu dalam dataset. Koordinat centroid dan nilai Silhouette yang ditampilkan di bawah masing-masing plot mengindikasikan evaluasi performa untuk setiap batch data *streaming*. Dapat diamati bahwa model mampu memisahkan wilayah dengan tingkat penyebaran dan kematian *COVID-19* yang tinggi dan rendah secara jelas ke dalam dua klaster. Pemisahan ini sangat penting dalam konteks epidemiologi, karena dapat membantu pemangku kebijakan dalam memprioritaskan intervensi kesehatan di wilayah dengan risiko tinggi.

Penggunaan algoritma *K-Means* dalam konteks data *streaming* ini mengacu pada pendekatan *online learning*, di mana centroid diperbarui secara dinamis setiap kali data baru masuk. Ini sesuai dengan prinsip *incremental clustering*, yang banyak digunakan dalam analisis data waktu nyata untuk menjaga adaptabilitas terhadap pola baru [28]. Lebih lanjut, pemanfaatan Apache Spark melalui antarmuka Sparklyr memungkinkan pemrosesan paralel dan efisien, sehingga sangat cocok untuk dataset berukuran besar seperti data *COVID-19* yang dikumpulkan harian dari seluruh Amerika Serikat [11]. Dengan pendekatan ini, model dapat terus mengidentifikasi dan memperbarui klaster berdasarkan data terbaru secara otomatis, tanpa perlu dilakukan pelatihan ulang secara manual. Ini mendukung praktik pengambilan keputusan berbasis data (*data-driven decision making*) yang responsif dan adaptif terhadap perubahan kondisi lapangan secara cepat.

Hasil *streaming K-Means* yang diamati pada gambar 3, terlihat bahwa titik pusat klaster (centroid) selalu mengalami perubahan untuk menyesuaikan setiap data *streaming* yang masuk dan diperbarui. Perubahan ini terjadi bahkan dengan sedikit pergeseran, yang menunjukkan responsivitas algoritma *clustering* terhadap data baru. Salah satu alasan utama dari pergeseran ini adalah perbedaan skala antara variabel X dan variabel Y, dimana nilai variabel X memiliki rentang yang sangat besar dibandingkan dengan variabel Y yang cenderung kecil. Hal ini mengakibatkan pusat klaster lebih sensitif terhadap perubahan kecil pada variabel Y, sementara variabel X memberikan kontribusi besar dalam penentuan posisi centroid.

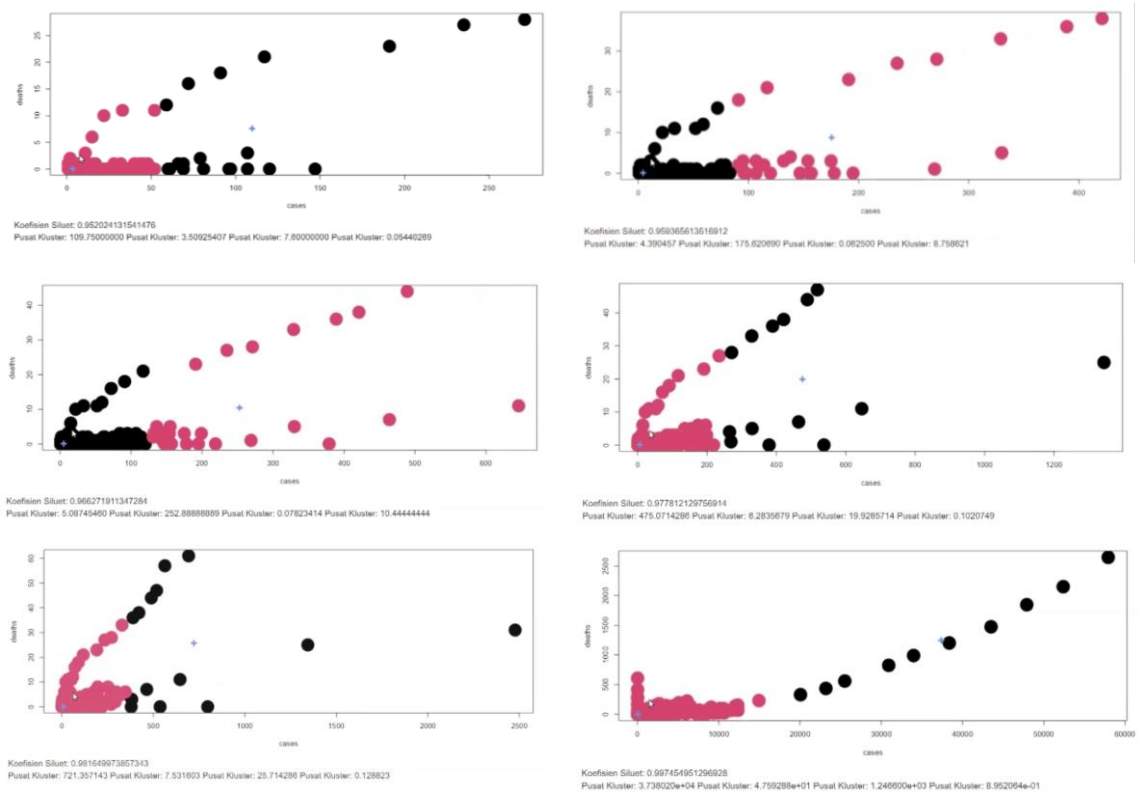
Model menghasilkan dua klaster utama. Klaster pertama terdiri dari wilayah-wilayah dengan jumlah kasus dan kematian relatif rendah, sedangkan klaster kedua mencakup wilayah dengan jumlah kasus dan kematian tinggi. Misalnya, negara bagian seperti Vermont (1.296 kematian), Alaska (1.633), dan Wyoming (2.340) termasuk dalam klaster pertama. Sebaliknya, wilayah seperti California (116.547 kematian), Texas (107.186), dan New York (86.570 termasuk NYC) tergolong ke dalam klaster kedua. Pola ini menunjukkan bahwa klasterisasi berhasil memetakan tingkat keparahan pandemi di setiap wilayah berdasarkan dua indikator epidemiologis utama. Studi oleh Sy, White, dan Nichols (2021) menunjukkan bahwa terdapat korelasi positif yang kuat antara **kepadatan penduduk dan tingkat infeksi COVID-19** di Amerika Serikat, di mana daerah dengan populasi padat mengalami beban pandemi lebih tinggi dibanding daerah rura[29].

Dari sisi geografis, klaster dengan tingkat kematian tinggi umumnya terletak di negara bagian yang padat penduduk, pusat bisnis, dan mobilitas tinggi, seperti California dan New York. Secara geografis, klaster kedua ini umumnya mencakup negara bagian dengan kepadatan penduduk tinggi atau wilayah metropolitan besar yang merupakan pusat aktivitas ekonomi dan sosial, sehingga

E-ISSN 2808-5841
P-ISSN 2808-7283

berpotensi mempercepat penyebaran virus, sehingga berdampak pada tingginya angka kematian. Temuan ini sejalan dengan data CDC per 14 Juni 2025, yang menunjukkan bahwa wilayah-wilayah tersebut merupakan kontributor utama kematian akibat *COVID-19* secara nasional [1]. Sebaliknya, negara bagian dengan populasi kecil dan kepadatan rendah lebih banyak masuk ke klaster berdampak ringan, memperkuat validitas model serta mengindikasikan korelasi yang kuat antara hasil pengelompokan dengan kondisi faktual yang dilaporkan oleh sumber resmi.

Faktor lain yang turut memperkuat pola ini adalah perbedaan akses dan kapasitas sistem kesehatan antarwilayah. Penelitian oleh Mishra et al. (2021) menyoroti bahwa meskipun beberapa negara bagian memiliki rumah sakit modern, kapasitas ICU dan tenaga medis tidak merata, sehingga daerah dengan kasus melonjak sering mengalami kekurangan layanan kritis. Dengan demikian, hasil klasterisasi berdasarkan data kasus dan kematian bukan hanya menunjukkan perbedaan statistik, tetapi mencerminkan ketimpangan dalam faktor struktural dan sosial yang mempengaruhi tingkat keparahan pandemi di setiap wilayah[30].



Gambar 3. Hasil *Streaming* Plot Clustering *K-Means* menggunakan K terbaik

IV. KESIMPULAN

Penelitian ini berhasil menerapkan algoritma ***Streaming K-Means Clustering*** menggunakan *Sparklyr* di platform *Apache Spark* untuk menganalisis data *COVID-19* dari The New York Times. Dengan bantuan aplikasi *Shiny*, pengguna dapat memilih variabel dan jumlah klaster secara interaktif. Hasil analisis menunjukkan bahwa jumlah klaster terbaik adalah **K=2**, karena memberikan nilai ***Silhouette Coefficient*** tertinggi dan standar deviasi terendah, yang menandakan pembagian klaster yang paling baik. Dua klaster yang terbentuk menggambarkan perbedaan tingkat penyebaran *COVID-19*. Klaster pertama terdiri dari wilayah dengan jumlah kasus dan kematian yang rendah, sementara klaster

kedua mencakup wilayah dengan jumlah kasus dan kematian yang tinggi. Penggunaan *Spark Streaming* memungkinkan analisis data dilakukan secara *real-time*, sehingga hasilnya bisa langsung menyesuaikan dengan kondisi terbaru. Hal ini sangat membantu bagi peneliti dan pembuat kebijakan dalam merespon perubahan situasi pandemi dengan cepat dan tepat.

REFERENSI

[1] C. for D. C. and Prevention, "COVID Data Tracker," *CDC*, 2023. https://covid.cdc.gov/covid-data-tracker/#datatracker-home (accessed May 30, 2025).

[2] E. Erwin, L. Judijanto, Y. Boari, and A. caezar to Tadampali, *PEMASARAN DIGITAL (Teori dan Implementasi)*, no. January. 2024. [Online]. Available: https://www.researchgate.net/publication/377638698

[3] B. J. Singh, A. Chakraborty, and R. Sehgal, "A systematic review of industrial wastewater management: Evaluating challenges and enablers," *J. Environ. Manage.*, vol. 348, no. October, p. 119230, 2023, doi: 10.1016/j.jenvman.2023.119230.

[4] M. Hosseinzadeh *et al.*, "Data cleansing mechanisms and approaches for big data analytics: a systematic study," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 1, pp. 99–111, 2023, doi: 10.1007/s12652-021-03590-2.

[5] A. Khorram-Manesh, K. Goniewicz, and F. M. Burkle, "Unleashing the global potential of public health: A framework for future pandemic response," *J. Infect. Public Health*, vol. 17, no. 1, pp. 82–95, 2024, doi: 10.1016/j.jiph.2023.10.038.

[6] D. Oliveira, A. Henriques, P. Nogueira, and A. Costa, "Impact of social prescribing intervention on people with type 2 diabetes mellitus in a primary healthcare context: a systematic literature review of effectiveness," *J. Public Heal.*, no. 0123456789, 2024, doi: 10.1007/s10389-024-02315-x.

[7] G. Rjoub, H. Elmekki, S. Islam, J. Bentahar, and R. Dssouli, "A hybrid swarm intelligence approach for optimizing Multimodal Large Language Models deployment in edge-cloud-based Federated Learning environments," *Comput. Commun.*, vol. 237, pp. 1–23, 2025, doi: 10.1016/j.comcom.2025.108152.

[8] S. Das and A. Abraham, *Data Mining and Knowledge Discovery Handbook*, no. July 2010. 2010. doi: 10.1007/978-0-387-09823-4.

[9] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[10] P. A. D. Thakare and S. M. Chaudhari, "Introducing a Hybrid Swarm Intelligence Based Technique for Document Clustering," *Int. J. Eng. Res. Appl. ISSN*, vol. 2, no. 6, pp. 1455–1459, 2012.

[11] M. Zaharia *et al.*, "Apache spark: A unified engine for big data processing," *Commun. ACM*, vol. 59, no. 11, pp. 56–65, 2016, doi: 10.1145/2934664.

[12] A. Khusaeri, "Penerapan Teknik Imputasi K-Means Terhadap Performa Hasil Klasifikasi Algoritma Naive Bayes," vol. 5, pp. 41–51, 2019.

[13] M. R. A. Prasetya, A. M. Priyatno, and Nurhaeni, "Penanganan Imputasi Missing Values pada Data Time Series dengan Menggunakan Metode Data Mining," *J. Inf. dan Teknol.*, vol. 5, no. 2, pp. 52–62, 2023, doi: 10.37034/jidt.v5i2.324.

[14] A. F. Nugraha, Y. Pristyanto, and I. Pratama, "Penanganan Missing Values Untuk Meningkatkan Kinerja Model Machine Learning Pada Data Telemarketing," *Pseudocode*, vol. 7, no. 2, pp. 165–171, 2020, doi: 10.33369/pseudocode.7.2.165-171.

[15] R. Aji Sasmoyo, "Sistem Clustering Daerah Rawan Kematian Anak Di Bawah Umur Di Wilayah Jawa Barat Menggunakan Algoritma K-Means," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 4, pp. 5635–5640, 2024, doi: 10.36040/jati.v8i4.9965.

[16] M. Riduwan, C. Fatichah, and A. Yuniarti, "Klasterisasi Dokumen Menggunakan Weighted K-Means Berdasarkan Relevansi Topik," *JUTI J. Ilm. Teknol. Inf.*, vol. 17, no. 2, p. 146, 2019, doi: 10.12962/j24068535.v17i2.a892.

[17] A. Abid and R. P. Setiawan, "Pemanfaatan Metode Clustering untuk Menganalisa Penduduk Kebumen yang Memiliki Keterampilan Teknologi Informasi dan Komunikasi (TIK)," *J. Data Sci. Theory Appl.*, vol. 2, no. 2, pp. 36–41, 2023.

[18] M. Alvarez-Garcia, R. Ibar-Alonso, and M. Arenas-Parra, "A comprehensive framework for explainable cluster analysis," *Inf. Sci. (Ny).*, vol. 663, no. July 2023, p. 120282, 2024, doi: 10.1016/j.ins.2024.120282.

[19] N. Rane, M. Paramesha, S. Choudhary, and J. Rane, "Machine Learning and Deep Learning for Big Data Analytics: a Review of Methods and Applications," *SSRN Electron. J.*, no. June, pp. 172–197, 2024, doi: 10.2139/ssrn.4835655.

[20] A. J. S. Ni Putu Eka Merliana, Ernawati, "PROSIDING SEMINAR NASIONAL MULTI DISIPLIN ILMU & CALL FOR PAPERS UNISBANK (SENDI_U) Kajian Multi Disiplin Ilmu untuk Mewujudkan Poros Maritim dalam Pembangunan Ekonomi Berbasis Kesejahteraan Rakyat ANALISA PENENTUAN JUMLAH CLUSTER TERBAIK PADA METODE K-ME," *Pros. Semin. Nas. MULTI DISIPLIN ILMU&CALL Pap. UNISBANK*, pp. 978–979, 2015, [Online]. Available: https://www.unisbank.ac.id/ojs/index.php/sendi_u/article/view/3333

[21] A. R. Wahidah *et al.*, "SISTEM PENDUKUNG ANALISA KEY PERFORMANCE INDICATOR (KPI) MENGGUNAKAN METODE DATA MINING BERBASIS," vol. 02, no. 03, pp. 151–158, 2022.

[22] A. Fitriani, E. Arfi, and A. Huda, "Penerapan Algoritma K-Means Clustering dalam Memetakan Produktivitas Lokasi Perkebunan Nanas PT Great Giant Pineapple," vol. 7, no. 2, pp. 215–231, 2024.

[23] A. Faisol, M. Orisa, and T. Informatika, "PENERAPAN K-MEANS CLUSTERING UNTUK PEMETAAN WILAYAH RAWAN BENCANA ALAM KOTA MALANG," vol. 8, no. 5, pp. 8560–8567, 2024.

[24] M. Vallath, *Oracle 10g RAC Grid, Services and Clustering*. Elsevier, 2006.

[25] R. Gubareva and R. P. Lopes, "Literature Review on the Smart City Resources Analysis with Big Data Methodologies," *SN Comput. Sci.*, vol. 5, no. 1, 2024, doi: 10.1007/s42979-023-02457-x.

[26] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[27] P.-N. Tan, M. Steinbach, Anuj Karpatne, and V. Kumar, *Introduction to Data Mining (Second Edition)*. Pearson Education, 2005.

[28] Charu C. Aggarwal, *Data Classification Algorithms and Applications*. New York: CRC PRESS, 2014.

[29] K. T. L. Sy, L. F. White, and B. E. Nichols, "Population density and basic reproductive number of COVID-19 across United States counties," *PLoS One*, vol. 16, no. 4 April, pp. 1–11, 2021, doi: 10.1371/journal.pone.0249271.

[30] V. Mishra *et al.*, "Health inequalities during COVID-19 and their effects on morbidity and mortality," *J. Healthc. Leadersh.*, vol. 13, pp. 19–26, 2021, doi: 10.2147/JHL.S270175.