Penggunaan Metode K-Means dan K-Means++ Sebagai Clustering Data Covid-19 di Pulau Jawa

Nursatio Nugroho[1], Faisal Dharma Adhinata[2*]

[1] Teknik Informatika, Institut Teknologi Telkom Purwokerto, Indonesia
[2] Rekayasa Perangkat Lunak, Institut Teknologi Telkom Purwokerto, Indonesia
Email: [1]18102208@ittelkom-pwt.ac.id, [2*] faisal@ittelkom-pwt.ac.id

(Naskah masuk: 1 Agu 2022, direvisi: 13 Sep 2022, 2 Okt 2022, diterima: 11 Okt 2022)

Abstrak

Virus Corona (*Covid-19*) merupakan penyakit menular yang dapat ditularkan antara hewan dan manusia. Pada akhir Desember 2019, virus itu teridentifikasi di Provinsi Wuhan, Cina. Saat ini, seluruh dunia sedang berjuang, mencegah, dan akhirnya menaklukkan penyebaran virus corona. Penelitian ini bertujuan untuk mengklaster data penyebaran *Covid-19* di setiap kabupaten di Pulau Jawa sehingga menghasilkan klaster zona yang harus dilaksanakan PPKM berdasarkan kasus positif, vaksin dosis pertama, dan dosis kedua. vaksin. Metode *K-Means* digunakan dengan cara menentukan jumlah *cluster* (K), mengatur pusat *cluster* secara arbitrer, mengelompokkan data ke dalam *cluster* dengan jarak terpendek, menghitung pusat *cluster*, dan mengulangi langkah 2-4 sampai tidak ada data yang berpindah ke lokasi yang berbeda. gugus. *K-Means++* digunakan dengan cara memilih secara acak nilai k pertama dari pusat *cluster* pertama titik data, mengelompokkan data berdasarkan jarak minimum ke *centroid*, memperbarui nilai titik *centroid* dengan menentukan rata-rata setiap *cluster*, dan ulangi langkah 2 dan 3 sampai tidak ada yang bergerak. Berdasarkan jumlah kasus positif, sembuh, dan meninggal, kasus tersebut dikategorikan. Setelah dilakukan pengelompokan dan mendapatkan klaster pada masing-masing kelompok, setiap klaster akan dievaluasi kualitasnya menggunakan koefisien siluet untuk memilih yang terbaik. Hasil kajian tersebut diharapkan dapat mengungkap sejauh mana penyebaran virus *Covid-19* di setiap kabupaten/kota di Pulau Jawa, serta *cluster* dengan skor *Silhouette Coefficient* tertinggi. Untuk hasil pengujian menggunakan *Silhouette Coefficient*, metode *K-Means* K=3 menghasilkan 0,825, K=4 menghasilkan 0,873, dan K=5 menghasilkan 0,862; untuk metode *K-Means++*, k=3 menghasilkan 0,822, K=4 menghasilkan 0,865, dan K = 5 menghasilkan 0,882. Hasil penelitian menunjukkan bahwa *K-Means++* lebih unggul dalam memberikan informasi sejauh mana penyebaran virus *Covid-19*, dan uji *Silhouette Coefficient* digunakan untuk menentukan kualitas *cluster* yang optimal.

Kata Kunci: *K-Means*, *K-Means++*, *Clustering*, *Covid-19*, Vaksin Pertama, Vaksin Kedua, *Silhouette Coefficient*.

The Use of K-Means and K-Means++ Methods as a Covid-19 Data Clustering in Java Island

Abstract

Corona virus (Covid-19) is an infectious disease that can be transmitted between animals and humans. At the end of December 2019, the virus was identified in Wuhan Province, China. Right now, the whole world is fighting, preventing, and finally conquering the spread of the coronavirus. This study aims to cluster data on the spread of Covid-19 in every district on the island of Java so as to produce zone clusters that must be implemented by PPKM based on positive cases, the first dose of vaccine, and the second dose. vaccine. The K-Means method is used by determining the number of clusters (K), setting the cluster center arbitrarily, grouping data into clusters with the shortest distance, calculating the cluster center, and repeating steps 2-4 until no data moves to a different location. group. K-Means++ is used by randomly selecting the first k value from the center of the first cluster of data points, grouping the data based on the minimum distance to the centroid, updating the centroid point value by determining the average of each cluster, and repeating steps 2 and 3 until nothing moves. Based on the number of positive cases, recovered, and died, the cases are categorized. After grouping and getting clusters in each group, each cluster will be evaluated for its quality using Silhouette Coefficients to choose the best. The results of the study are expected to reveal the extent of the spread of the Covid-19 virus in every district/city on the island of Java, as well as the cluster with the highest Silhouette Coefficient score. For the test results using the Silhoette Coefficien, the K-Means method K=3 produces 0.825, K=4

produces 0.873, and K=5 produces 0.862; for the K-Means++ method, k=3 produces 0.822, K=4 produces 0.865, and K = 5 produces 0.882. The results showed that K-Means++ was superior in providing information on the extent of the spread of the Covid-19 virus, and the Silhouette Coefficient test was used to determine the optimal cluster quality..

Keywords: *K-Means, K-Means++, Clustering, Covid-19, First Vaccine, Second Vaccine, Silhouette Coefficient.*

I. PENDAHULUAN

Virus Corona (*Covid-19*) adalah penyakit menular yang dapat menyerang manusia dan hewan. Virus ini ditemukan sekitar akhir Desember 2019 di Provinsi Wuhan, Cina. Saat ini, seluruh dunia sedang memerangi, menghentikan, dan mengalahkan penyebaran virus corona. Berdasarkan statistik *World Health Organization* hingga 27 Oktober 2021, jumlah kasus *Covid-19* di Indonesia sebanyak 4.241.849 dengan 143.249 kasus kematian [1]. Akibat pandemi *Covid-19* ini mempengaruhi ekonomi dunia tidak hanya Indonesia saja.

Angka kasus positif *Covid-19* sama dengan jumlah kasus yang dilaporkan di setiap kabupaten atau kota di Pulau Jawa. Pulau Jawa memiliki jumlah penduduk terbanyak, sehingga dapat mewakili setiap wilayah di Indonesia, dengan masing-masing kabupaten/kota memiliki ciri khasnya masing-masing. Akibatnya, kepadatan penduduk tiap kabupaten/kota berbeda-beda. Beberapa kabupaten/kota dengan skala kasus positif tertinggi akan membutuhkan perhatian khusus. Kota Jakarta Barat, Kota Jakarta Selatan, Kota Jakarta Utara, Kota Jakarta Pusat, Kab. Bogor, Kota Surabaya, dan Kota Jakarta Timur termasuk kabupaten/kota dengan prevalensi penularan virus corona yang tinggi. Hal ini menyulitkan pemerintah Indonesia untuk mendefinisikan contoh penanganan virus corona [2].

Selain itu, langkah pemerintah untuk membatasi penyebaran *Covid-19* antara lain penetapan Program Pembatasan Kegiatan Masyarakat (PPKM). Jika jumlah kasus positif bertambah, PPKM akan dilaksanakan dengan menggunakan data variabel yang dikumpulkan dari kasus positif, dosis vaksin pertama, dan dosis vaksin kedua. Selain itu, lonjakan kasus baru ini disebabkan oleh masih banyaknya penduduk yang menolak menggunakan masker saat bepergian. Penggunaan masker secara langsung dapat menghambat penularan infeksi karena masker memenuhi peraturan kesehatan [3].

Virus *Covid-19* telah menyerang berbagai daerah dalam jarak yang berdekatan. Warga di zona merah memiliki pilihan untuk berbaur tanpa harus jauh-jauh atau keluar kota. Mereka yang dinyatakan positif virus *Covid-19* dikarantina selama 14 hari dalam satu ruangan dengan pasien positif *Covid-19* lainnya [4], Penggunaan *K-Means Cluster Analysis* untuk mengorganisasikan data *Covid-19* merupakan salah satu pendekatan analisis *cluster non-hierarchical* yaitu strategi *cluster* dimana jumlah *cluster* diatur secara manual. Tujuan dari *cluster* ini adalah untuk membagi item yang diamati menjadi satu atau lebih kelompok tergantung pada karakteristiknya. Objek dengan kualitas yang sama akan ditempatkan dalam *cluster* yang sama, sedangkan objek dengan fitur yang berbeda akan dicampur dengan *cluster* lainnya [5].

Selain menggunakan Algoritma *K-Means* untuk mengelompokkan data *Covid-19*, penelitian ini juga menggunakan Algoritma *K-Means++*, yaitu mengorganisasikan data dengan menerapkan jumlah awal grup K dan mendasarkan setiap iterasi pada jarak terdekat. Iterasi adalah proses pengerjaan dari awal ke titik pusat baru dengan nilai yang sama dengan titik pusat sebelumnya; jika persyaratan ini belum terpenuhi, proses akan berlanjut sampai jumlah maksimum iterasi telah ditetapkan dari awal. Tujuan utama algoritma ini adalah untuk menempatkan titik-titik data ini sebagai pusat awal sejauh mungkin dari satu sama lain [6].

Data yang disajikan oleh *covid19.go.id* hanya agregat data per-provinsi, namun perlu informasi sebaran kasus *Covid-19* perkabupaten/kota. Penelitian ini bertujuan untuk mengelompokkan data sebaran *Covid-19* pada kabupaten/kota di pulau Jawa yang belum disajikan oleh *covid19.go.id* secara terperinci informasi ini sehingga dapat membantu pembentukan klaster zona berdasarkan karakteristik positif, vaksin dosis pertama, dan vaksin dosis kedua [7]. *Clustering* adalah teknik untuk mengkategorikan data berdasarkan kesamaan karakteristik di seluruh kelompok data. Metode *K-Means*, metode *K-Means++*, pendekatan *K-Modes*, *Hierarchical Clustering*, dan teknik lainnya adalah beberapa cara pengelompokan data. *K-Modes* lebih berguna untuk mengolah data kategorikal [8].

Penggunaan *K-Means* untuk klasterisasi data *Covid-19* merupakan salah satu pendekatan analisis *cluster non-hierarchical* yaitu strategi *cluster* dimana jumlah *cluster* diatur secara manual. Tujuan dari *cluster* ini adalah untuk membagi objek yang diamati menjadi satu atau lebih kelompok tergantung pada karakteristiknya. Objek dengan karakteristik yang sama akan ditempatkan dalam *cluster* yang sama, sedangkan objek dengan karakteristik yang berbeda akan dicampur dengan *cluster* lainnya [9]. Dalam kajian yang menganalisis *K-Means*, Siti Azizatus Sholihah mengelompokkan provinsi berdasarkan prevalensi *Covid-19* di setiap provinsi di Indonesia. Kesimpulan dari penelitian ini adalah rata-rata *Silhouette Coefficient* (SC) untuk k = 2 adalah 0,74 dengan 3 data tidak akurat karena 3 negatif. Namun *Silhouette Coefficient* = 0,74 termasuk dalam struktur kokoh. Provinsi Jawa Timur, Jawa Tengah, Jawa Barat, DKI Jakarta, Banten, dan Riau tergolong rawan penyebaran virus *Covid-19* berdasarkan teknik *K-Means*. Sementara 28 provinsi dinyatakan bebas dari risiko penularan *Covid-19*. Kekurangan pada metode *K-Means* yaitu pada pemilihan *centroid* awal.

Metode *K-Means++* memilih *centroid* awal yang dengan cara yang berbeda dengan *K-Means* [10].

Beberapa latar belakang diatas maka perlunya untuk informasi sebaran mengenai kasus *Covid-19* pada setiap kabupaten/kota di pulau Jawa dan perlu menguji *cluster* dari metode *K-Means* dan *K-Means++* untuk mencari *cluster* yang optimal dari kedua metode tersebut. Penelitian ini akan menggunakan variabel positif, vaksin dosis pertama, dan vaksin dosis kedua untuk mengetahui perkembangan kasus *Covid-19* dengan mengelompokkan zona penyebaran virus corona, yang dapat memberikan informasi status zona berdasarkan hasil *cluster* yang digunakan. Hasil dari *cluster* tersebut akan dievaluasi dengan menggunakan metode *Silhouette Coefficient* untuk menentukan *cluster* mana yang memiliki kualitas lebih baik.

II. METODOLOGI PENELITIAN

A. Subjek dan Objek Penelitian

Subjek pada penelitian ini adalah pengelompokkan data pasien *Covid-19*. Sedangkan objek penelitiannya adalah data pasien *Covid-19* dengan pembaruan data pada tanggal 09 Febuari 2022 yang bersumber dari *web https://m.andrafarm.com/.*

B. Jenis Penelitian

Penelitian ini berdasarkan dari hasil riset dan data terbaru mengenai angka pasien yang sudah dinyatakan positif, jumlah vaksin dosis pertama dan vaksin dosis kedua pada setiap penduduk di pulau Jawa. Oleh karena itu penelitian ini termasuk jenis penelitian kuantitatif.

C. Sumber Data

Jenis data yang digunakan dalam penelitian ini adalah data sekunder. Data bisa didapatkan dari laman *web https://m.andrafarm.com/* dengan variabel yang dapat dilihat pada Tabel 1.

Tabel 1. Variabel Penelitian

Variabel	Pengertian
X1	Total orang yang sudah dinyatakan positif *Covid –19*
X2	Total orang yang sudahvaksin dosis pertama
X3	Total orang yang sudahvaksin dosis kedua

D. Normalisasi Data

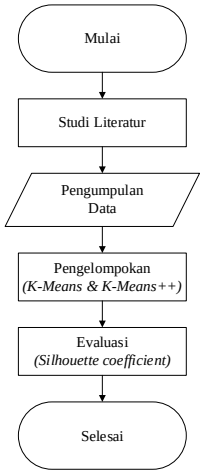
$$X^1 = \frac{X - X_{min}}{X_{max} X_{min}} \quad (1)$$

dimana,

$X1$ = Data i atribut yang akan dinormalisasikan.
$Xmin$ = Nilai i terkecil i atribut tersebut
$Xmax$ = Nilai tertinggi atribut tersebut.

E. Desain Penelitian

Prosedur penelitian ini ditunjukkan dalam Gambar 1. Berikut penjelasan dari Gambar 1:

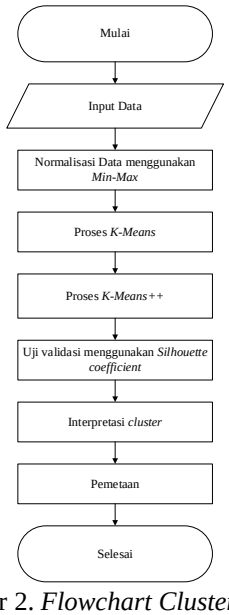


Gambar 1. *Flowchart* Penelitian

1. Studi Literatur
   Studi Literatur untuk mencari referensi pada saat melakukan penelitian.
2. Pengumpulan Data
   Pengumpulan data diambil dari sumber yang sudah ditentukan dari awal
3. Pengelompokan
   Pengelompokan data dilakukan menggunakan *K-Means* dan *K-Means++* untuk mendapatkan kabupaten/kota yang rawan dan aman dari kasus *Covid-19*.
4. Evaluasi
   Evaluasi untuk *K-Means* dan *K-Means++* menggunakan *Silhouette Coefficient* untuk mendapatkan *score* setiap *cluster* yang dibentuk.

F. *Flowchart Clustering*

Flowchart Clustering ini ditunjukkan dalam Gambar 2. Berikut penjelasan dari Gambar 2:



Gambar 2. *Flowchart Clustering*

1. Memasukkan data dalam bentuk tabel/matriks.
   Data yang bersumber dari data pasien *covid-19* yang diambil dari *website kawalcovid19.id*
2. Normalisasi data
   Dari data *Covid-19* diatas akan dinormalisasikan menggunakan *Min-Max* untuk mengurangi jarak antara nilai setiap variabel karena perbedaan skala.
3. Membuat *cluster* menggunakan metode non-herarki (*K-Means* dan *K- Means++*)
   a. Menentukan banyaknya nilai k yaitu banyaknya *cluster*. Nilai k akan diuji coba pada penelitian ini sebanyak 3 sampai 5 Nilai k disini akan berdampak pada hasil yang akan dihitung nanti, semakin banyak nilai k akan mengurangi nilai *error*.
   b. Memilih *centroid* (titik pusat) di setiap *cluster*
   c. Menghitung jarak setiap objek dan setiap *centroid* dengan persamaan berikut.

$$d_{(x,y)} = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \quad (2)$$

Penjelasan:
$d_{(y,x)}$ = Jarak data ke x ke pusat *cluster* y
x_i = Data x pada observasi ke-i
y_i = Titik pusat ke y observasi ke-i
n = Banyaknya observasi

$$V_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} X_{kj}\text{z} \quad (3)$$

Penjelasan:
V_{ij} = *Centroid* rata – rata pada *cluster* ke-I untuk variabel ke – j
Ni = Jumlah anggota *cluster* ke-i
i,k = Indeks dari *cluster*
j = Indeks variabel
Xkj = Nilai data ke – k variabel ke -j untuk *cluster* tersebut

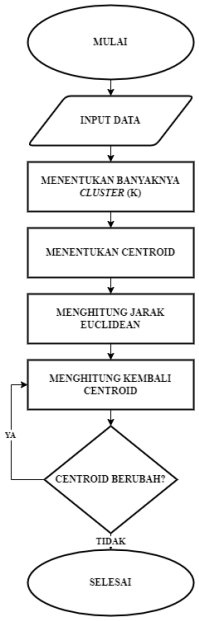
   d. Dihitung kembali nilai *centroid* untuk *cluster* yang baru terbentuk dengan persamaan diatas. Mengulangi langkah c dan d sampai tidak ada lagi perubahan posisi nilai *centroid*.

$$K = \frac{D(x)^2}{\sum_{x\epsilon X} D(x)^2} \quad (4)$$

Penjelasan:

$D(x)^2$ = Jarak *Euclidean distance*
$\sum_{x\epsilon X} D(x)^2$ = Jumlah jarak *Euclidean distance*

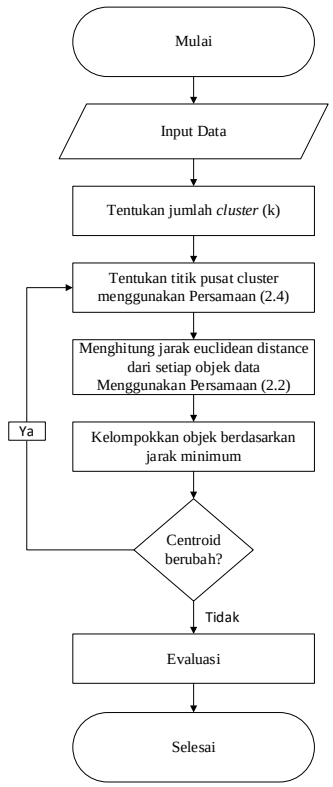
4. Hasil *Clustering* akan divalidasi dengan *Silhouette Coefficient*.
5. Tentukan kategori dari setiap *cluster* dengan melihat nilai *centroid*-nya. Membuat kategori untuk mengimplentasikan hasil *cluster* yang telah dihitung.
6. Dengan memeriksa anggota *cluster*, Anda dapat mewarnai kabupaten/kota. Mewarnai peta sesuai dengan kategori yang sudah ditentukan, warna merah untuk zona tinggi, warna kuning untuk zona sedang, dan warna hijau zona rendah.



Gambar 3. *Flowchart K-Means*

G. *Flowchart K-Means++*

Adapun langkah–langkah pada algoritma *K-Means++* adalah sebagai berikut:



Gambar 4. *Flowchart K-Means++*

1. Memilih nilai K untuk titik *centroid* dengan persamaan. Setelah menentukan banyaknya nilai k dan dihitung menggunakan persamaan akan mengeluarkan hasil yang akan digunakan untuk titik *centroid* (titik pusat).
2. Mengelompokkan berdasarkan jarak minimun terhadap *centroid*. Dikelompokkan hasil dari nilai k yang sudah dihitung sebelumnya dengan jarak minimum terhadap *centroid* (titik pusat).
3. Memperbarui titik *centroid* dengan mencari rata – rata setiap *cluster* Setelah mengelompokkan nilai *centroid* (titik pusat) dengan jarak minimum, langkah ini mencari rata – rata dari setiap *cluster* untuk dijadikan nilai *centroid* (titik pusat).
4. Mengulangi langkah 2 dan 3 sampai semua objek tidak berpindah.
5. Menguji validasi menggunakan *Silhouette Coefficient*. Setelah nilai *centroid* semua tidak berpindah akan diuji menggunakan *Silhouette Coefficient* dan mendapatkan persentase nilai yang optimal.

H. *Silhouette Coefficient*

Adapun langkah untuk menghitung *Silhouette Coefficient* diawali dengan mencari jarak rata-rata data ke-i dengan semua data dalam *cluster* yang sama, disini anggap saja data ke-i terletak pada *cluster* A. Rumus dari a(i) terdapat di persamaan berikut.

$$a(i) = \frac{1}{|A|-1}\sum_{j\epsilon A, j\neq 1} d(i,j) \quad (5)$$

di mana,
A = total data di *cluster* A

Kemudian tentukan nilai b(i) yang merupakan nilai minimal dari rata-rata jarak data ke-i antara semua data pada setiap *cluster*. Sekarang mari kita perhatikan *cluster* lain selain A dan *cluster* C. Untuk menghitung jarak rata-rata data ke-i dengan semua data di *cluster* C, menggunakan persamaan berikut.

$$d(i,C) = \frac{1}{|C|}\sum_{j\epsilon C} d(i,j) \quad (6)$$

di mana,
C = total data di *cluster* C

Setelah menghitung d(i,C) untuk semua *cluster*, langkah selanjutnya untuk memilih jarak paling minimum sebagai nilai b(i), menggunakan persamaan berikut.

$$b(i) = \min_{C\neq A} d(i,j) \quad (7)$$

Cluster terbaik kedua untuk data ke-i, setelah *cluster* A, adalah *cluster* B, yang disebut sebagai tetangga dari data ke-i jika *cluster* B memiliki nilai jarak minimal, yaitu d(i,B) = b(i). *Silhouette Coefficient* ditentukan pada langkah terakhir setelah a(i) dan b(i) diketahui.

Jika nilai indeksi *Silhouette* mendekati 1 berarti proses *Clustering* berhasil, dan jika mendekati 0, berarti prosedur *Clustering* tidak berhasil. Berikut rumus *Silhouette Coefficient*, pada persamaan berikut.

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),(b(i)} \quad (8)$$

s(i) = nilai *silhouette coefficient* pada object ke-i
(i) = object yang akan diteliti
a(i) = rata-rata i jarak i antar anggota i dalam *cluster* yang sama.
b(i) = nilai minimal jarak rata-rata antara *item* ke-i dengan objek pada *cluster* lain.
Nilai s(i) berletak diantara -1 dan 1, di mana setiap nilai ditafsirkan sebagai berikut:
s(i) = 1 => data ke-i digolongkan dengan baik (dalam A)
s(i) = 0 => data ke-i terletak di tengah antara dua *cluster* (A dan B)
s(i) = -1 => data ke-i tergolong lemah (dekat ke *cluster* B daripada A)

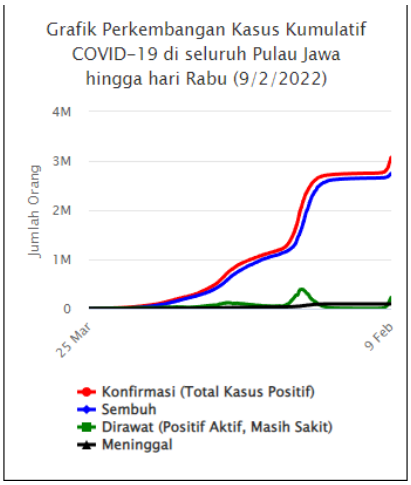
Tabel 2. *Silhouette Coefficient*

Range	Interpretasi
0.71 – 1.00	Struktur yang dihasilkan kuat
0.51 – 0.70	Struktur yang dihasilkan baik
0.26 – 0.50	Struktur yang dihasilkan lemah
≤ 0.25	Tidak terstruktur

III. HASIL DAN PEMBAHASAN

A. Penjelasan Data

Data dari M.Andrafarm merupakan kumpulan dataisekunder dari berbagai *website* resmi setiap provinsi, kabupaten, dan kota yang ada di Indonesia, bisa diakses secara gratis pada laman *web* https://m.andrafarm.com/. Data tersebut digunakan untuk analisis i *cluster* dengan objek kabupaten dan kota pada Pulau Jawa berdasarkan 3 variabel, yaitu kasus positif, vaksin pertama, dan, vaksin kedua.



Gambar 5. Grafik Penyebaran *Covid-19* Setiap Kabupaten dan Kota di Pulau Jawa

Gambar 5 menunjukkan epidemi virus *Covid-19* di Indonesia. Jika ternyata semakin hari semakin meningkat, maka perlu dilakukan analisa daerah mana saja yang rawan dan aman dari *Covid-19*.

B. Normalisasi Data

Sebelum pengelompokan, data dinormalisasi menggunakan perhitungan *Min – Max Normalization* yang ditunjukkan pada persamaan dibawah. Di bagian ini menggambarkan proses normalisasi salah satu objek dari *input* data variabel X1 di Kota Tanggerang Selatan dengan *max* = 203252 dan *min* = 1359. Di bawah ini adalah proses normalisasi untuk variabel X1 yang range-nya adalah Kota Tangerang Selatan.

Di bawah ini adalah proses normalisasi variabel X1 di Kota Tanggerang Selatan dengan *margin* 0 hingga 1.

$$X^1 = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (9)$$
$$= \frac{47364 - 1359}{203252 - 1359}$$
$$= \frac{46005}{201893}$$
$$= 0{,}227868$$

Dengan digunakan perhitungan akan mendapatkan data normalisasi yang bisa dilihat pada Tabel 3.

Tabel 3. Normalisasi Data

Kabupaten / Kota	X1	X2	X3
Kota Tangerang Selatan	0,227868	0,291192	0,316020
Kota Tangerang	0,222732	0,438048	0,447272
Kab. Tangerang	0,168634	0,617048	0,625351
Kota Cilegon	0,061929	0,071801	0,081838
Kab. Serang	0,042334	0,136434	0,145711
Kab. Lebak	0,038857	0,204924	0,168744
Kota Serang	0,035083	0,136434	0,132965
Kab. Pandeglang	0,027841	0,202095	0,145877
Kota Jakarta Timur	0,114531	0,761681	0,849941
Kota Jakarta Selatan	1,000000	0,903166	1,000000

Tabel 3 adalah hasil data yang dinormalisasi menggunakan rumus *Min-Max Normalization.*

C. Proses *K-Means*

Algoritma *K-Means* berusaha untuk mengklasifikasikan data ke dalam kelompok berdasarkan sifat bersama. Data yang digunakan adalah data yang dinormalisasi. Prosedur *K-Means* dibagi menjadi beberapa tahapan sebagai berikut:

1. Menentukan nilai k

Pada tahap ini, peneliti menguji berbagai ukuran *cluster*, dari tiga hingga lima, untuk menentukan ukuran *cluster* yang ideal. Sehubungan dengan perhitungan yang dimungkinkan oleh program *Python*, peneliti akan menguji 3 *cluster* dengan kelompok rawan, cukup rawan, dan aman. Sedangkan 4 *cluster* maka akan dikelompokkan berdasarkan sangat rawan, rawan, cukup rawan, dan aman. Untuk 5 *cluster* dikelompokkan berdasarkan sangat rawan, rawan, cukup rawan, sangat aman, dan aman. Yang terakhir 6 *cluster* dikelompokan berdasarkan sangat rawan, rawan, cukup rawan, sangat aman, aman, dan cukup aman.

2. Menentukan *centroid* awal dengan acak

Pada tahap ini peneliti mendapatkan *centroid* i awal secara acak yang bisa dilihat pada Tabel 4.

Tabel 4. *Centroid* Awal

Cluster	C1	C2	C3
3	0,59342077	0,74905316	0,76830423
4	0,05000647	0,17249937	0,1717042
5	0,14397775	0,48489708	0,54648138

3. Menghitung jarak menggunakan *Euclidean Distance*.

Langkah selanjutnya adalah menentukan jarak dari setiap titik data ke *centroid* setelah *centroid* ditentukan secara acak. Tabel 5-7 menampilkan hasil perhitungan jarak Euclidean dari percobaan 3, 4, 5 *cluster*.

Tabel 5. Data Hasil Perhitungan Jarak *Euclidean* Pada Percobaan 3 *Cluster*

Kabupaten / Kota	C1	C2	C3
Kota Tangerang Selatan	0,740153	0,257972	0,312525
Kota Tangerang	0,580686	0,419867	0,135053
Kab. Tangerang	0,467231	0,646135	0,155859
Kota Cilegon	1,101085	0,135492	0,627115
Kab. Serang	1,032773	0,045113	0,540717
Kab. Lebak	0,981371	0,034416	0,481789
Kota Serang	1,044365	0,054992	0,551616
Kab. Pandeglang	1,003224	0,045103	0,503933
Kota Jakarta Timur	0,485963	0,900725	0,411782
Kota Jakarta Selatan	0,492687	1,456858	1,055179

Tabel 6. Data Hasil Perhitungan Jarak *Euclidean* Pada Percobaan 4 *Cluster*

Kabupaten / Kota	C1	C2	C3	C4
Kota Tangerang Selatan	0,340098	0,345041	0,740153	0,179211
Kota Tangerang	0,511377	0,160983	0,580686	0,316684
Kab. Tangerang	0,739994	0,123033	0,467231	0,538278
Kota Cilegon	0,052065	0,664008	1,101085	0,241205
Kab. Serang	0,050340	0,577891	1,032773	0,152448
Kab. Lebak	0,115443	0,519256	0,981371	0,094669
Kota Serang	0,039545	0,588876	1,044365	0,162685
Kab. Pandeglang	0,101833	0,541446	1,003224	0,119022

Kota Jakarta Timur	0,994099	0,377057	0,485963	0,793478
Kota Jakarta Selatan	1,538339	1,026399	0,492687	1,365640

Tabel 7. Data Hasil Perhitungan Jarak *Euclidean* Pada Percobaan 5 *Cluster*

Kabupaten / Kota	C1	C2	C3	C4	C5
Kota Tangerang Selatan	0,340098	0,785977	0,177080	0,336859	0,836472
Kota Tangerang	0,511377	0,589337	0,312792	0,147558	0,717664
Kab. Tangerang	0,739994	0,344011	0,533964	0,124714	0,671348
Kota Cilegon	0,052065	1,120641	0,245545	0,665190	1,185028
Kab. Serang	0,050340	1,034579	0,156750	0,580667	1,130590
Kab. Lebak	0,115443	0,972210	0,098920	0,522730	1,088937
Kota Serang	0,039545	1,044820	0,167013	0,591848	1,142434

Hasil perhitungan jarak *Euclidean* pada uji coba tiga *cluster*, empat *cluster*, dan lima *cluster* yang memiliki hasil pengelompokan masing-masing data yang memiliki fitur yang sama ditunjukkan pada Tabel 5 sampai dengan Tabel 7. Tabel tersebut sampai pada kesimpulan bahwa data dikelompokkan atau dimasukkan ke dalam *cluster* 1 jika palingdekat dengan *centroid* 1, dan seterusnya.

4. Menentukan *centroid* baru

Langkah selanjutnya adalah menghitung *centroid* sekali lagi dengan mencari rata-rata dari setiap *cluster* yang identik. Hasil *centroid* baru ditunjukkan pada Tabel 8. Untuk menentukan *centroid* sekali lagi, yaitu menghasilkan *centroid* baru dengan rata-rata setiap *cluster*.

Tabel 8. *Centroid* Baru

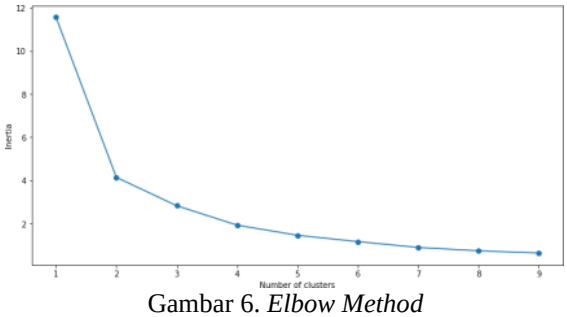
Uji Coba	Pusat	X_1	X_2	X_3
3	0	0,593421	0,749053	0,768304
	1	0,050006	0,172499	0,171704
	2	0,143978	0,484897	0,546481
4	0	0,033637	0,109380	0,104160
	1	0,154025	0,505072	0,576512
	2	0,593421	0,749053	0,768304
	3	0,068009	0,245369	0,249222
5	0	0,033637	0,109380	0,104160
	1	0,239609	0,858044	0,860356
	2	0,068200	0,247846	0,252898
	3	0,181742	0,504977	0,572227
	4	0,824970	0,686680	0,748159

Operasi *Clustering* telah berakhir karena *centroid* baru pada Tabel 8 sama dengan *centroid* lama pada Tabel 4.

D. Proses *K-Means++*

1. Menentukan nilai k

Pada tahap ini, peneliti melakukan tes 3 *cluster* sampai 5 *cluster* untuk mencari berapa jumlah *cluster* yang optimal. Mengenai perhitungan dibantu dengan program *python*, peneliti akan menguji 3 *cluster* dengan kelompok tinggi, sedang, dan rendah. Sedangkan 4 *cluster* maka akan dikelompokkanberdasarkan sangat tinggi, tinggi, sedang, dan rendah. Yang terakhir 5 *cluster* dikelompokkanberdasarkan sangat tinggi, tinggi, sedang, sangataman, dan aman. Berikut Gambar 6 percobaan *cluster* 1-10 menggunakan *Elbow Method.*



Gambar 6. *Elbow Method*

2. Menentukan *centroid* awal

Pada tahap ini peneliti mendapatkan *centroid* awal yang bisa dilihat pada Tabel 9.

Tabel 9. *Centroid* Awal

Uji Coba	C1	C2	C3
3	0,05000647	0,17249937	0,1717042
4	0,06819959	0,24784561	0,25289828
5	0,068009	0,24536864	0,24922218

3. Menghitung jarak menggunakan *Euclidean Distance*

Langkah selanjutnya adalah menentukan jarak dari setiap titik data ke *centroid* setelah nilai *centroid* yang sudah didapatkan. Tabel 10 menampilkan hasil perhitungan jarak *Euclidean* dari percobaan 3, 4, 5 *cluster*.

Tabel 10. Data Hasil Perhitungan Jarak *Euclidean* Pada Percobaan 3 *Cluster*

Kabupaten / Kota	C1	C2	C3
Kota Tangerang Selatan	0,257972	0,291335	0,717517
Kota Tangerang	0,419867	0,118030	0,547636
Kab. Tangerang	0,646135	0,176224	0,410869
Kota Cilegon	0,135492	0,605904	1,078398
Kab. Serang	0,045113	0,519813	1,006365
Kab. Lebak	0,034416	0,461186	0,952879
Kota Serang	0,054992	0,530779	1,017916
Kab. Pandeglang	0,045103	0,483469	0,974877
Kota Jakarta Timur	0,900725	0,433455	0,416539
Kota Jakarta Selatan	1,456858	1,066720	0,545188

Tabel 11. Data Hasil Perhitungan Jarak *Euclidean* Pada Percobaan 4 *Cluster*

Kabupaten / Kota	C1	C2	C3	C4
Kota Tangerang Selatan	0,177 080	0,356 396	0,340 098	0740 153
Kota Tangerang	0,312 792	0,170 222	0,511 377	0,580 686
Kab. Tangerang	0,533 964	0,111 107	0,739 994	0,467 231
Kota Cilegon	0,245 545	0,677 173	0,052 065	1,101 085
Kab. Serang	0,156 750	0,591 194	0,050 340	1,032 773
Kab. Lebak	0,098 920	0,532 545	0,115 443	0,981 371
Kota Serang	0,167 013	0,602 188	0,039 545	1,044 365
Kab. Pandeglang	0,123 189	0,554 722	0,101 833	1,003 224
Kota Jakarta Timur	0,789 072	0,365 056	0,994 099	0,485 963
Kota Jakarta Selatan	1,362 299	1,015 330	1,538 339	0,492 687

Tabel 12. Data Hasil Perhitungan Jarak *Euclidean* Pada Percobaan 5 *Cluster*

Kabupaten / Kota	C1	C2	C3	C4	C5
Kota Tangerang Selatan	0,179 211	0,753 725	0,340 098	0,785 977	0,321 684
Kota Tangerang	0,316 684	0,628 342	0,511 377	0,589 337	0,139 874
Kab. Tangerang	0,538 278	0,580 632	0,739 994	0,344 011	0,144 504
Kota Cilegon	0,241 205	1,106 228	0,052 065	1,120 641	0,641 180
Kab. Serang	0,152 448	1,048 832	0,050 340	1,034 579	0,555 443
Kab. Lebak	0,094 669	1,005 254	0,115 443	0,972 210	0,497 178
Kota Serang	0,162 685	1,060 660	0,039 545	1,044 820	0,566 510
Kab. Pandeglang	0,119 022	1,026 862	0,101 833	0,992 771	0,519 524
Kota Jakarta Timur	0,793 478	0,644 455	0,994 099	0,158 236	0,400 623
Kota Jakarta Selatan	1,365 640	0,448 132	1,538 339	0,774 423	1,037 842

Hasil perhitungan jarak *Euclidean* pada uji coba dua *cluster*, tiga *cluster*, dan empat *cluster* yang memiliki hasil pengelompokan masing-masing data yang memiliki fitur yang sama ditunjukkan pada Tabel 10 sampai dengan Tabel 12. Tabel tersebut sampai pada kesimpulan bahwa data dikelompokkan atau dimasukkan ke dalam *cluster* 1 jika paling dekat dengan *centroid* 1, dan seterusnya.

4. Menentukan *centroid* baru

Langkah selanjutnya adalah menghitung *centroid* sekali lagi dengan mencari rata- rata dari setiap *cluster* yang identik. Hasil dari *centroid* baru ditunjukkan pada Tabel 13. Untuk menentukan *centroid* baru, yaitu untuk mendapatkan *centroid* baru dengan rata-rata setiap *cluster*.

Tabel 13. *Centroid* Baru

Uji Coba	Pusat	X_1	X_2	X_3
3	C1	0,593421	0,749053	0,768304
	C2	0,050006	0,172499	0,171704
	C3	0,143978	0,484897	0,546481
4	C1	0,158818	0,513718	0,585709
	C2	0,033637	0,109380	0,104160
	C3	0,593421	0,749053	0,768304
	C4	0,068200	0,247846	0,252898
5	C1	0,068009	0,245369	0,249222
	C2	0,068009	0,245369	0,249222
	C3	0,739057	0,670467	0,719675
	C4	0,033637	0,109380	0,104160
	C5	0,239609	0,858044	0,860356

E. *Silhouette Coefficient*

Setelah menyelesaikan tahapan pengerjaan *K-Means* dan *K-Means++* penguji selanjutnya akan menguji hasil dari pengerjaan kedua metode tersebut dan melihat hasil *cluster* terbaik *score* nya saat diuji menggunakan *Silhouette Coefficient* berikut hasil dari kedua metode bisa di lihat pada Tabel 14 dan 15.

Tabel 14. Hasil Pengujian Setiap *Cluster* Dari *K-Means*

K	3	4	5
SC	0,825	0,873	0,862

Tabel 15. Hasil Pengujian Setiap *Cluster* Dari *K-Means++*

K	3	4	5
SC	0,825	0,873	0,862

Tabel 14 dan 15 merupakan hasil pengujian antara metode *K-Means* dan *K-Means++* diantara kedua metode tersebut yang mendapatkan hasil pengujian tertinggi adalah *K-Means++* dengan *score* 0.882 merupakan Struktur yang dihasilkan kuat.

F. Hasil dan Penjelasan

Centroid akhir *cluster* yang ideal dapat diidentifikasi dalam analisis *Clustering* menggunakan *K-Means* dengan membandingkan pola pembentukan masing-masing *centroid* pada Tabel 5 dan 7 untuk *K-Means++* pembentukan masing – masing *centroid* pada Tabel 10 dan 12 dengan *centroid* akhir yang dipilih.

Tabel 16. Pola setiap *centroid K-Means++*

Pusat	$X1$	$X2$	$X3$	Kategori *Cluster*
C1	Rendah	Rendah	Rendah	Aman

C2	Sangat Tinggi	Sangat Tinggi	Sangat Tinggi	Sangat Rawan
C3	Sangat Rendah	Sangat Rendah	Sangat Rendah	Sangat Aman
C4	Tinggi	Tinggi	Tinggi	Rawan
C5	Cukup Rendah	Cukup Rendah	Cukup Rendah	Cukup Aman

Tabel 16 menyimpulkan bahwa C3 memiliki pola *centroid* Sangat Rendah, sehingga sangat aman untuk kasus positif *Covid-19*. Karena C1 memiliki pola *centroid* yang rendah, maka dianggap aman untuk pasien positif *Covid-19*. Karena pola *centroid* rendah C5, itu dianggap cukup aman untuk kasus positif *Covid-19*. Karena pola *centroid* C2 yang sangat tinggi, diklasifikasikan sebagai sangat sensitif terhadap kasus positif *Covid-19*. Karena pola *centroid* C4 yang tinggi, maka tergolong lebih berpeluang menemui kasus positif *Covid-19*.

Data nama – nama kabupaten dan kota yang telah dikelompokkan berdasarkan *cluster* yang sudah dikerjakan menggunakan *K-Means++* dengan *score* pengujian tertinggi dapat dilihat pada Tabel 17.

Tabel 17 menyimpulkan bahwa Kota Jakarta Barat, Kota Jakarta Selatan, Kota Jakarta Utara, Kota Jakarta Pusat merupakan kota yang sangat rawan akan Covid-19 dan Kab. Bogor, Kota Surabaya, Kota Jakarta Timur merupakan kabupaten dan kota yang rawan akan Covid-19 dikarenakan pada kabupaten/kota tersebut akan cepat dalam penyebarannya, sedangkan 110 kabupaten/kota lainnya dikategorikan aman akan Covid-19 sehingga tingkat penularan dari kabupaten/kota tersebut akan kecil.

Ada tindakan yang dapat dilakukan untuk meningkatkan kesadaran masyarakat, seperti 5M, di kabupaten/kota yang aman terhadap Covid-19 (Menjaga Jarak, Cuci Tangan,). Selain itu, setiap orang yang masuk dari transportasi yang sama harus diperiksa dan karantina 14 hari harus diberlakukan bagi mereka yang menimbulkan bahaya.

Provinsi rawan Covid-19 dapat mengambil tindakan pencegahan yang sama dengan daerah aman, selain meminta masyarakat menahan diri untuk tidak menghadiri pertemuan yang tidak penting dan hanya bepergian untuk alasan yang sangat diperlukan

Tabel 17. Hasil *Cluster* Kabupaten Dan Kota Di Pulau Jawa

Kelompok	Jumlah Anggota	Kabupaten/Kota
Sangat Aman	46	Kab. Bondowoso, Kota Blitar, Kab. Pandeglang, Kota Serang, Kab. Lumajang, Kab. Madiun, Kota Batu, Kota Madiun, Kab. Magetan, Kota Kediri, Kab. Banjarnegara, Kota Mojokerto, Kab. Serang, Kota Cilegon, Kab. Ngawi, Kab. Pacitan, Kab. Pamekasan, Kota Pasuruhan, Kab. Ponorogo, Kota Probolinggo, Kab. Sampang, Kab. Situbondo, Kab. Sumenep, Kab. Trenggalek, Kab. Kepulauan Seribu, Kab. Bangkalan, Kota Pekalongan, Kab. Pengandaran, Kab. Batang, Kab. Blora, Kota Banjar, Kab. Purwakarta, Kota Sukabumi, Kota Cirebon, Kab. Kudus, Kota Cimahi, Kota Magelang, Kab. Pekalongan, Kab. Rembang, Kota Salatiga, Kab. Purworejo, Kab. Kulon Progo, Kab. Gunung Kidul, Kota Tegal, Kab. Wonosobo, Kab. Temanggung
Cukup Aman	16	Kota Bekasi, Kab. Sidoarjo, Kab. Cilacap, Kota Tanggerang, Kota Bandung, Kab. Garut, Kota Semarang, Kab. Cianjur, Kab. Bekasi, Kab. Karawang, Kota Depok, Kab. Malang, Kab. Bandung, Kab. Cirebon, Kab. Sukabumi, Kab. Tanggerang
Aman	48	Kota Tanggerang Selatan, Kab. Sregen, Kab. Semarang, Kab. Purbalingga, Kab. Pemalang, Kab. Tuban, Kab. Pati, Kab. Magelang, Kab. Kendal, Kab. Kebumen, Kab. Karanganyar, Kab. Jepara, Kab. Grobogan, Kab. Demak, Kab. Brebes, Kab. Sukoharjo, Kota Surakarta, Kab. Tegal, Kab. Wonogiri, Kab. Probolinggo, Kab. Pasuruhan, Kab. Nganjuk, Kab. Mojokerto, Kota Malang, Kab. Lamongan, Kab. Kediri, Kab. Boyolali, Kab. Jombang, Kab. Gresik, Kab. Bojonegoro, Kab. Blitar, Kab. Banyuwangi, Kota Yogyakarta, Kab. Sleman, Kab. Bantul, Kab. Jember, Kab. Banyumas, Kab. Tulungagung, Kota Tasikmalaya, Kab. Indramayu, Kab. Kuningan, Kab. Majalengka, Kab. Ciamis, Kab. Subang, Kab. Lebak, Kab. Tasikmalaya, Kab. Sumedang
Sangat Rawan	4	Kota Jakarta Barat, Kota Jakarta Selatan, Kota Jakarta Utara, Kota Jakarta Pusat
Rawan	3	Kab. Bogor, Kota Surabaya, Kota Jakarta Timur

IV. KESIMPULAN

Metode *K-Means++*, sebagaimana ditunjukkan dari validasi *cluster* menggunakan *Silhouette*, menghasilkan hasil *cluster* yang lebih baik untuk *Clustering* nilai k lebih tinggi dengan rata-rata *Silhouette Coefficient* (SC) sebesar 0,882 pada k = 5. Sedangkan untuk *K-Means* akan lebih baik digunakan untuk *Clustering* nilai k yang rendah dengan *score* k = 3 sebesar 0,825, dan k = 4 sebesar 0,873.

Berdasarkan *Clustering* dengan metode *K-Means++*, untuk Kota Jakarta Barat, Kota Jakarta Selatan, Kota Jakarta Utara, Kota Jakarta Pusat tergolong sangat rawan oleh penyebaran *Covid-19* untuk Kab. Bogor, Kota Surabaya, Kota Jakarta Timur tergolong rawan oleh penyebaran covid – 19.

Untuk 110 kabupaten/kota lainnya dikategorikan aman dari kasus penyebaran *Covid-19*. Informasi tersebut akan berguna bagi masyarakat supaya lebih ketat dalam menjaga Kesehatan.

REFERENSI

[1] A. Solichin and K. Khairunnisa, "Klasterisasi Persebaran Virus Corona (covid-19) di DKI Jakarta Menggunakan Metode K-Means," *Fountain of Informatics Journal*, vol. 5, no. 2, p. 52, 2020.

[2] D. D. Darmansah, "Analisis Penyebaran Penularan Virus Covid-19 di Provinsi Jawa Barat Menggunakan Algoritma K-Means Clustering," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 8, no. 3, pp. 1188–1199, 2021.

[3] D. T. Utari, "Analisis Karakteristik Wilayah Transmisi Covid-19 Dengan Menggunakan Metode K-Means Clustering," *Jurnal Media Teknik dan Sistem Industri*, vol. 5, no. 1, p. 25, 2021.

[4] S. F. Mandang and B. N. Sari, "Penerapan K-Means Cluster Pada Daerah Penggunaan Teknologi di Indonesia," *JOINS (Journal of Information System)*, vol. 6, no. 1, pp. 131–138, 2021.

[5] D. Sari and Y. Sukestiyarno, "Analisis Cluster Dengan Metode K-Means Pada Persebaran Kasus COVID-19 Berdasarkan Provinsi di Indonesia", *prisma*, vol. 4, pp. 602-610, Feb. 2021.

[6] Sharon, S. Defit, and G. W. Nurcahyo, "Tingkat Efisiensi Penggunaan Resep Dokter Spesialis Menggunakan Metode K-Means Clustering," *Jurnal Informasi dan Teknologi*, pp. 121–127, 2021.

[7] D. A. Dewi and D. A. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette Pada Algoritma Clustering K-Medoids Dalam Pengelompokan Produksi Kerajinan Bali," *Matrix : Jurnal Manajemen Teknologi dan Informatika*, vol. 9, no. 3, pp. 102–109, 2019.

[8] F. M. Falahi, "Penerapan Metode Clustering Untuk Pengelompokan Mahasiswa Potensial Drop Out Menggunakan Algoritma K-Means ++," thesis, Universitas Islam Negeri Sunan Ampel, Surabaya, 2019.

[9] R. A. Indraputra and R. Fitriana, "K-Means Clustering Data Covid-19," *Jurnal Teknik Industri*, vol. 10, no. 3, pp. 275–282, 2020.

[10] N. Ulinnuha and S. A. Sholihah, "Analisis Cluster Untuk Pemetaan Data Kasus Covid-19 Di Indonesia Menggunakan K-Means," *Jurnal MSA (Matematika dan Statistika serta Aplikasinya)*, vol. 9, no. 2, 2021.