Kualitas & Kuantitas (2022) 56:1283–1291 https://
doi.org/10.1007/s11135-021-01176-w

Penerapan klasterisasi K-means untuk klasterisasi provinsi di Indonesia dari risiko pandemi COVID-19 berdasarkan data COVID-19

Dahlan Abdullah[1] · S. Susilo[2] · Ansari Saleh Ahmar[3] · R. Rusli[4] · Rahmat Hidayat[5]

Diterima: 24 Mei 2021 / Diterbitkan online: 3 Juni 2021

Abstrak

Penelitian ini dilakukan dengan tujuan untuk mengelompokkan provinsi-provinsi di Indonesia dari risiko pandemi COVID-19 berdasarkan data coronavirus disease 2019 (COVID-19). Pengelompokan ini berdasarkan data yang diperoleh dari Satuan Tugas Percepatan Penanganan Covid-19 Indonesia (SATGAS COVID-19) pada 19 April 2020. Provinsi di Indonesia dikelompokkan berdasarkan data kasus terkonfirmasi, meninggal, dan sembuh COVID-19. Ini dilakukan dengan menggunakan metode K-Means Clustering. Clustering menghasilkan 3 grup provinsi. Hasil klasterisasi provinsi diharapkan dapat memberikan masukan kepada pemerintah dalam membuat kebijakan terkait pembatasan kegiatan masyarakat atau kebijakan lainnya dalam mengatasi penyebaran COVID-19. Pengelompokan Provinsi berdasarkan kasus COVID-19 di Indonesia merupakan upaya untuk menentukan kedekatan atau kemiripan suatu provinsi berdasarkan kasus terkonfirmasi, sembuh, dan meninggal. Berdasarkan hasil penelitian ini, terdapat 3 klaster provinsi.

Kata kunciCOVID-19 · Pengelompokan · Pengelompokan K-means

* Ansari Saleh Ahmar
ansarisaleh@unm.ac.id

Dahlan Abdullah
dahlan@unimal.ac.id

R. Rusli
rusli.siman@unm.ac.id

Rahmat Hidayat
rahmat@pnp.ac.id

[1] Jurusan Teknologi Informasi, Fakultas Teknik, Universitas Malikussaleh, Lhokseumawe, Indonesia

[2] Jurusan Pendidikan Biologi, Universitas Muhammadiyah Prof. Dr. Hamka, Jakarta, Indonesia

[3] Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Makassar, Makassar, Indonesia

[4] Jurusan Matematika, Fakultas MIPA, Universitas Negeri Makassar, Makassar, Indonesia

[5] Departemen Teknologi Informasi, Politeknik Negeri Padang, Padang, Indonesia

1. Perkenalan

Coronavirus disease 2019 (COVID-19) merupakan penyakit menular yang saat ini beredar di seluruh dunia (Ahmar dan Rusli2020; Atuahene dkk.2020; Gupta dkk.2020). COVID-19 pertama kali dilaporkan di kota Wuhan, Provinsi Hubei, Cina pada bulan Desember 2019. COVID-19 adalah penyakit menular yang disebabkan oleh coronavirus yang baru ditemukan—severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)—yang pertama kali diidentifikasi di Wuhan (Ahmar dan Boj2020; Azarafza dkk.2021). Kasus COVID-19 pertama di Indonesia terdeteksi pada 2 Maret 2020 di Jakarta. Seiring waktu, pandemi telah menyebar ke berbagai provinsi di Indonesia. Hingga 19 April 2020, lebih dari 6575 kasus COVID-19 telah dilaporkan di 34 provinsi di Indonesia. Pada 19 April 2020, 6.575 kasus terkonfirmasi, 686 sembuh, dan 582 meninggal di Indonesia. Berdasarkan data COVID-19 dari Worldometer, terakhir diperbarui: 20 April 2020, 07:53 GMT, Indonesia memiliki kasus terkonfirmasi COVID-19 tertinggi di antara negara-negara anggota Perhimpunan Bangsa-Bangsa Asia Tenggara (ASEAN) (Worldometer2020).

Evaluasi perkembangan kasus COVID-19 per provinsi menjadi salah satu dasar pemantauan perkembangan kasus COVID-19 di Indonesia. Namun, hingga saat ini belum ada pengelompokan provinsi berdasarkan kasus terkonfirmasi, kesembuhan, dan kematian yang dilakukan pada data tersebut. Algoritma pengelompokan K-means adalah teknik populer tanpa pengawasan yang digunakan untuk mengidentifikasi kesamaan antara objek berdasarkan vektor jarak yang cocok untuk kumpulan data kecil (Sreedhar et al.2017). Teknik ini menurut definisi adalah semacam algoritma cluster, dan memiliki beberapa keunggulan termasuk singkat, efisiensi dan kecepatan (Li dan Haiyan2012). Sedangkan tujuan dari analisis klaster adalah (1) menyelidiki struktur yang mendasari data, (2) klasifikasi: untuk menentukan tingkat kesamaan antara titik-titik data dan (3) kompresi: suatu metode untuk mengatur dan meringkas data ke dalam kelompok-kelompok yang dapat dimengerti (Govender dan Sivakumar 2020).

Armstrong, dkk. (2012) mengatakan bahwa algoritme K-means sangat membantu dalam mengelompokkan populasi klien pemulihan yang heterogen menjadi subkelompok yang lebih homogen dan K-means menawarkan pandangan yang lebih baik tentang karakteristik dan kebutuhan pelamar, yang dapat mengarah pada pilihan rehabilitasi yang lebih bertarget untuk orang-orang yang dirawat di rumah. Hal ini senada dengan Kusrini (2015) bahwa K-means clustering digunakan karena jumlah cluster yang dibutuhkan untuk kategorisasi item telah ditentukan dan selain itu, Fotouhi & Montazeri-Gh (2013) mengatakan bahwa pengelompokan K-means membutuhkan komputasi yang lebih sedikit daripada proses SAPM, yang menguntungkan kemampuan metode untuk pengelompokan lalu lintas yang akurat. Selanjutnya, Al-Wakeel dan Wu (2016) menunjukkan bahwa untuk profil beban berkorelasi kuat, disarankan sejumlah klaster terbatas.

Dengan menggunakan metode data mining seperti klasterisasi K-means, dimungkinkan untuk menemukan karakteristik utama dari masing-masing provinsi potensial yang dapat digunakan dalam upaya untuk memprediksi kasus COVID-19 di masa mendatang berdasarkan kesamaan data provinsi.

2 Metode dan Analisis Statistik

Penelitian ini dilakukan dengan menggunakan data yang diperoleh pada tanggal 19 April 2020 dari situs web Gugus Tugas Percepatan COVID-19 Indonesia (https://covid19.go.id/peta-sebaran). Analisis data menggunakan metode K-Means Clustering sebagai teknik untuk melakukan pengelompokan data. Selanjutnya, prosedur klasifikasi data didasarkan pada derajat keanggotaan masing-masing komponen (Ahmar et.al.,2018). Analisis ini dilakukan dengan menggunakan R Software versi 3.6.3. seperti yang dijelaskan di situs web (https://uc-r.github.io/) dan penelitian ini menggunakan R Software versi 3.6.3.

Langkah-langkah penelitian dilakukan sebagai berikut:

(1) Data kasus terkonfirmasi, sembuh, dan meninggal diperoleh dari situs web Indonesia COVID-19 (https://covid19.go.id/peta-sebaran).

(2) Data ini diekstraksi menjadi 3 bagian yang meliputi dikonfirmasi, pulih, dan kematian menurut provinsi yang berbeda.

(3) Apabila ada data yang lebih dominan dibandingkan dengan yang lain maka himpunan tersebut dijadikan 1 kelompok dan dikeluarkan dari proses analisis.

(4) Paket-paket berikut telah Diinstal dan dijalankan; rapiverse (versi 1.3.0), cluster (versi 2.1.0), dan factoextra (versi 1.0.7) dari R Software versi 3.6.3.

(5) Data yang diperoleh pada tahap 2 dimuat lebih lanjut pada Perangkat Lunak R.

```
perpustakaan("readxl")
data <- read_excel("C:\\datacovid19indonesia.xlsx")
```

(6) Persiapan Data:

(a) Baris adalah observasi, kolom adalah variabel.

(b) Nilai data yang hilang akan dihapus atau diperkirakan.

Untuk menghapus nilai yang hilang yang mungkin ada dalam data, ketik ini:

```
data <- na.omit(data)
```

(c) Data distandarisasi (yaitu diskalakan) untuk membuat variabel dapat dibandingkan. Untuk menskalakan/menstandarkan data menggunakan skala fungsi R:

```
data <- skala(data)
kepala (data)
```

(7) Pengukuran jarak clustering dilakukan dengan menggunakan jarak Euclidean.

```
euclidean <- get_dist(data)
fviz_dist(euclidean, gradien = daftar(rendah = "#00AFBB", mid = "putih", tinggi = "#FC4E07"))
```

(8) Proses analisis K-means dapat digambarkan sebagai berikut:

(a) Tentukan jumlah cluster (k) menggunakan cluster optimal. Tiga (3) cluster optimal yang paling populer digunakan, meliputi:

(1) Metode siku

```
set.seed(123)
fviz_nbclust(data, kmeans, metode = "wss")
```

(2) Metode siluet

```
set.seed(123)
fviz_nbclust(data, kmeans, metode = "siluet")
```

(3) Statistik kesenjangan

```
set.seed(123)
fviz_gap_stat(gap_stat)
```

Cluster yang optimal dilihat dari fungsi fviz_nbclust masing-masing metode. Selanjutnya nilai cluster yang optimal pada Metode Elbow adalah *k* nilai yang turun drastis pada grafik visualisasi sedangkan pada statistik Silhoute dan Gap muncul secara otomatis pada grafik.

(b) Mengekstrak hasil

Berdasarkan pendekatan metode cluster optimal pada langkah sebelumnya, maka akan diperoleh cluster yang optimal. Jumlah cluster digunakan untuk menghitung nilai clustering k-means.

Misalnya, pada tahap sebelumnya, nilai$k$=2 diperoleh.

```
set.seed(123)
endkmeans <- kmeans(data, 2, nstart = 25)
print(endkmeans)
```

Berdasarkan hasil tersebut akan diperoleh hasil k-means clustering. Hasil ini dapat divisualisasikan menggunakan kode:

```
fviz_cluster(endkmeans, data = data)
```

3 Hasil dan Pembahasan

Berdasarkan analisis statistik deskriptif (Tabel1) dari 34 provinsi di Indonesia, kasus terkonfirmasi terbanyak 3.032 kasus sembuh 234 kasus dan meninggal 287 kasus, provinsi tidak sembuh dan kasus meninggal. Rata-rata, jumlah kasus yang dikonfirmasi adalah 193 dengan standar deviasi 528.

Dalam Gambar.1, Jakarta jelas lebih banyak kasusnya sehingga provinsi tersebut menjadi episentrum data center sehingga Jakarta membentuk satu kelompok khusus dan tidak termasuk dalam proses klasterisasi data (Pamula et al.2011). Epicentrum berbasis di Jakarta karena merupakan ibu kota negara dan pusat perekonomian di Indonesia.

Selain itu, jumlah optimal$k$kelompok ditentukan menggunakan tiga (3) pendekatan yang paling umum digunakan yaitu Elbow, Silhouette, dan Gap Statistics. Hasilnya dapat dilihat pada Gambar.2a, b, dan c.

Berdasarkan Gambar.2, metode Elbow diperoleh optimal$k$pada$k$=2, metode Silhouette mendapatkan banyak cluster optimal pada$k$=2, dan statistik Gap yang diperoleh optimal$k$nilai untuk membentuk cluster di$k$=2.

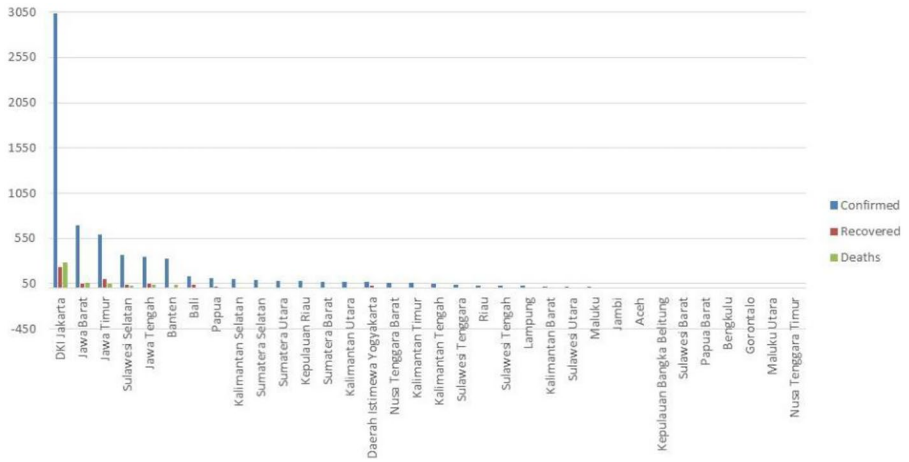
Oleh karena itu, berdasarkan hasil dari metode-metode tersebut, dapat disimpulkan bahwa yang optimal k nilai untuk membentuk cluster adalah 2. Selanjutnya hasil analisis Clustering menggunakan K-means dengan$k$=2 disajikan dalam Tabel2.

Seperti yang ditunjukkan pada Tabel2, terlihat bahwa Klaster 1 terdiri dari 5 provinsi dan Klaster 2 terdiri dari 28 provinsi. Jika digabungkan dengan Klaster DKI Jakarta, maka akan ada 3 klaster provinsi di Indonesia berdasarkan data COVID-19 (Gbr.3).

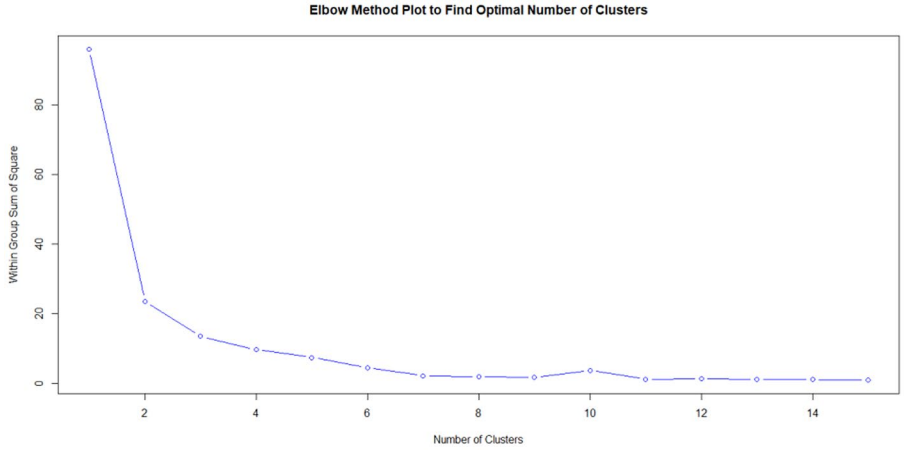
Penelitian ini sejalan dengan Zarikas, et.al. (2020), yang menunjukkan bahwa pengelompokan kasus aktif di suatu wilayah berguna untuk menarik kesimpulan tentang dampak penyakit yang menyebar dengan cepat di suatu wilayah. Selanjutnya, Azarafza dkk. (2021) menyatakan bahwa

Tabel 1 Statistik Deskriptif COVID-19 di Indonesia

Variabel	Observasi	Obs. dengan hilang data	Obs. tanpa yang hilang	Minimum	Maksimum	Mean	Std. penyimpangan data
Dikonfirmasi	34	0	34	1	3032	193	528
Pulih	34	0	34	0	234	20	43
Meninggal	34	0	34	0	287	17	50



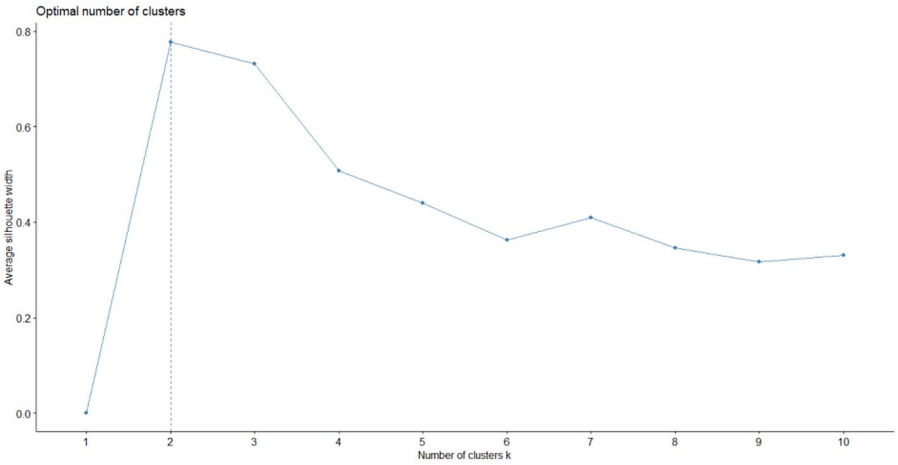
Gambar 1 Jumlah kasus COVID-19 tiap Provinsi di Indonesia



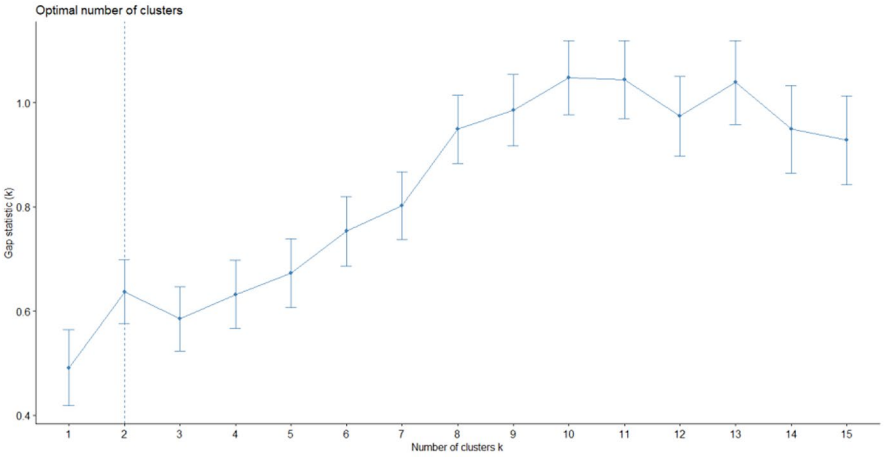
(a)

Gambar 2. Hasil dari**sebuah**Siku,**b**Siluet, dan**c**Gap Statistic untuk mencari k optimal

pola penularan infeksi antar provinsi diperkirakan dengan metode clustering. Oleh karena itu, berdasarkan pendapat tersebut dapat disimpulkan bahwa dengan melakukan klaster provinsi diberikan gambaran pola penyebaran penyakit dan solusi terkait pola persebaran tersebut.



(b)



(c)

Gambar 2.(lanjutan)

4. Kesimpulan

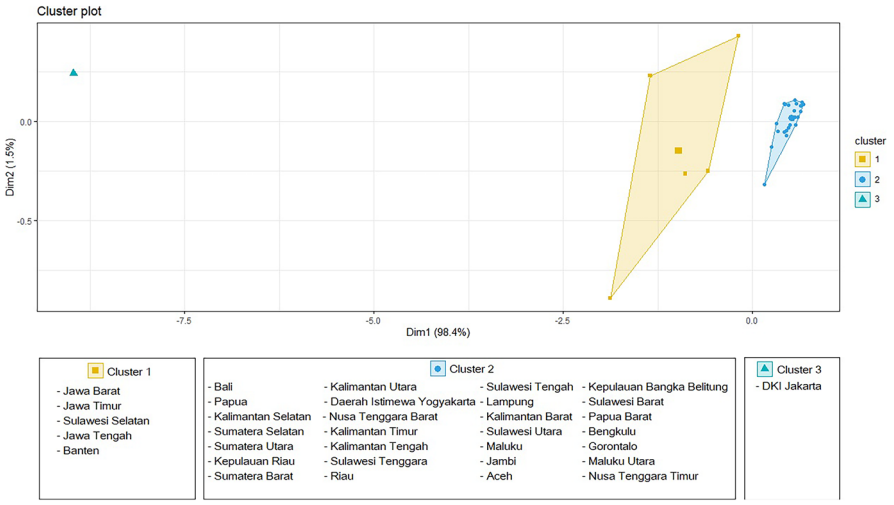
Pengelompokan/pengelompokan provinsi berdasarkan kasus COVID-19 di Indonesia merupakan upaya untuk menentukan kedekatan atau kemiripan suatu provinsi berdasarkan kasus terkonfirmasi, kasus sembuh, dan kasus meninggal. Berdasarkan hasil penelitian ini, terdapat 3 klaster provinsi yang masing-masing terdiri dari:**Cluster 1**(Jawa Barat, Jawa Timur, Sulawesi Selatan, Jawa Tengah); **Gugus 2**(Bali,Papua, Kalimantan Selatan, Sumatera Selatan, Sumatera Utara, Kepulauan Riau, Sumatera Barat, Kalimantan Utara, Daerah Istimewa Yogyakarta, Nusa Tenggara Barat, Kalimantan Timur, Kalimantan Tengah, Sulawesi Tenggara, Riau, Sulawesi Tengah,

Meja 2 Hasil clustering provinsi di Indonesia dengan K-Means clustering*

Propinsi	Gugus
Jawa Barat	1
Jawa Timur	1
Sulawesi Selatan	1
Jawa Tengah	1
Banten	1
Bali	2
Papua	2
Kalimantan Selatan	2
Sumatera Selatan	2
Sumatera Utara	2
Kepulauan Riau	2
Sumatera Barat	2
Kalimantan Utara	2
Daerah Istimewa Yogyakarta	2
Nusa Tenggara Barat	2
Kalimantan Timur	2
Kalimantan Tengah	2
Sulawesi Tenggara	2
Riau	2
Sulawesi Tengah	2
Lampung	2
Kalimantan Barat	2
Sulawesi Utara	2
Maluku	2
Jambi	2
Aceh	2
Kepulauan Bangka Belitung	2
Sulawesi Barat	2
Papua Barat	2
Bengkulu	2
gorontalo	2
Maluku Utara	2
Nusa Tenggara Timur	2

*Tidak termasuk Provinsi DKI Jakarta

Lampung, Kalimantan Barat, Sulawesi Utara, Maluku, Jambi, Aceh, Kepulauan Bangka Belitung, Sulawesi Barat, Papua Barat, Bengkulu, Gorontalo, Maluku Utara, Nusa Tenggara Timur); dan**Gugus 3**(DKI Jakarta). Hasil klaster provinsi diharapkan dapat memberikan masukan kepada pemerintah dalam membuat kebijakan terkait pembatasan kegiatan masyarakat atau kebijakan lainnya dalam mengatasi penyebaran COVID-19.



Gambar 3 Hasil Clustering Provinsi di Indonesia dengan K-Means Clustering

ucapan terima kasih Penulis ingin menjawab wasit atas saran mereka yang bermanfaat.

Kontribusi penulis Penulis berkontribusi pada naskah secara setara.

PendanaanPenulis menyatakan bahwa tidak ada dana untuk penelitian ini.

ketersediaan data Data dalam penelitian ini dapat diakses di: Website COVID-19 Indonesia (https://covid19.go.id/peta-sebaran) atau Harvard Dataverse (https://doi.org/10.7910/DVN/JUSYYX).

Deklarasi

Konflik kepentingan Penulis menyatakan bahwa tidak ada konflik kepentingan.

Referensi

Ahmar, AS, Boj, E.: Tanggal prediksi 200.000 kasus covid-19 di spanyol. J. Aplikasi. Sci. Ind. teknologi. pendidikan**2**(2), 188–193 (2020)

Ahmar, AS, Rusli, R.: Akankah kasus covid-19 di dunia mencapai 4 juta dengan pendekatan peramalan menggunakan su-air mata. JOIV: Int. J. Informasikan. melihat**4**(3), 159-161 (2020)

Ahmar, Ansari Saleh, Napitupulu, Darmawan, Rahim, Robbi, Hidayat, Rahmat, Sonatha, Yance, Azmi, Meri: Menggunakan k-means clustering untuk cluster provinsi di indonesia. J. Fisik: Konf. Ser.**1028**, 012006 (2018)

Al-Wakeel, A., Jianzhong, Wu.: Analisis klaster berbasis K-means dari pengukuran meteran pintar perumahan. Prosedur Energi**88**, 754–760 (2016)

Armstrong, JJ, Zhu, M., Hirdes, JP, Stolee, P.: Analisis klaster K-Means pengguna layanan rehabilitasi di sistem perawatan kesehatan rumah ontario: memeriksa heterogenitas populasi geriatri yang kompleks. Lengkungan. fisik Med. rehabilitasi.**93**(12), 2198–2205 (2012)

Atuahene, S., Kong, Y., Bentum-Micah, G.: Pandemi Covid-19, kerugian ekonomi dan pengelolaan sektor pendidikan usia. Bergalah. Ekonomi Kelola. pejantan**1**(2), 103–109 (2020)

Azarafza, M., Azarafza, M., Akgun, H.: Metode clustering untuk analisis pola penyebaran virus corona (covid-19) infeksi di iran. J. Aplikasi Sci. Ind. teknologi. pendidikan**3**(1), 1–6 (2021)

Fotouhi, A., Montazeri-Gh, M.: Tehran driving cycle development menggunakan metode k-means clustering. Scientia Iranica.**20**(2), 286–293 (2013)

Govender, P., Sivakumar, V.: Penerapan k-means dan teknik pengelompokan hierarkis untuk analisis polusi udara: Sebuah tinjauan (1980–2019). atmosfer. polusi. Res.**11**(1), 40–56 (2020)

Gupta, Ritik Ranjan, Arya, Ravi Kumar, Kumar, Jatin, Shubham, Tanay: Bagaimana covid-19 membawa pernikahan India-industri ding berhenti? JINAV: J. Menginformasikan. melihat**1**(2), 83–91 (2020)

Kusrini, K.: Pengelompokan barang retail dengan menggunakan K-means clustering. Komputer Procedia. Sci.**72**, 495–502 (2015)

Li, Y., Haiyan, Wu.: Metode pengelompokan berdasarkan algoritma k-means. fisik procedia**25**, 1104–1109 (2012)

Pamula, R. Deka, JK, Nandi, S.: Metode pendeteksian outlier berbasis clustering. Pada tahun 2011 Interna-Konferensi Nasional tentang Aplikasi Teknologi Informasi yang Muncul, halaman 253–256. IEEE, 2011. Sreedhar,

Chowdam, Kasiviswanath, Nagulalaply, Reddy, PakantiChenna: Mengelompokkan kumpulan data besar menggunakan k-means modifikasi inter dan intra clustering (km-i2c) di hadoop. J. Data Besar**4**(1), 27 (2017) Worldometer.

Pandemi virus corona COVID-19, 2020. URLhttp://worldometers.info/coronavirus. Terakhir diakses 20 April 2020.

Zarikas, Vasilios, Poulopoulos, Stavros G., Gareiou, Zoe, Zervas, Efthimios: Analisis pengelompokan negara mencoba menggunakan set data kasus covid-19. Data Singkat**31**, 105787 (2020)