PENGKLASTERAN TERHADAP NEGARA-NEGARA DENGAN JUMLAH KASUS COVID 19 TERBESAR

Muhammad Ibnu Choldun Rachmatullah[1]
[1]Program Studi D-III Sistem Informasi, Politeknik Pos Indonesia
e-mail: [1]ibnucholdun@poltekpos.ac.id

ABSTRAK

Pandemi Covid 19 telah menyebar di seluruh negara di dunia. Berdasarkan data pada tanggal 11 Juli 2021 telah menyebabkan 187.393.771 orang terinfeksi di seluruh dunia. Paper ini bertujuan untuk mengelompokkan negara dengan jumlah kasus terbesar berdasar karakteristik kesamaan dan perbedaan dengan mempertimbangkan atribut jumlah kasus per 1 juta penduduk, jumlah kematian per 1 juta penduduk, dan jumlah tes per 1 juta penduduk. Metode pengklasteran yang digunakan adalah K-Means, sedangkan penentuan jumlah klaster optimal dengan menerapkan kriteria Elbow. Dari penerapan kriteria Elbow ini dari hasil eksperimen diperoleh jumlah klaster optimal=5, sehingga akan membagi 30 negara dengan jumlah kasus terbesar ke dalam lima kelompok/kategori. Dari pengelompokan ini dapat diketahui karakteristik 30 negara dengan jumlah kasus Covid 19 terbesar, berdasar jumlah kasus dan jumlah pengetesan per 1 juta penduduk.

***Kata Kunci**: Covid 19, pengklasteran, K-Means, kriteria Elbow*

1. PENDAHULUAN

Pandemi dan epidemi dapat menyebabkan banyak kematian hanya dalam beberapa hari. Dengan meningkatnya laju pertumbuhan penduduk, maka laju penyakit menular juga semakin meningkat. COVID-19 telah menyebabkan 187.393.771 orang terinfeksi dengan 4.042.198 kematian di seluruh dunia (berdasarkan data dari https://www.worldometers.info/coronavirus/ pada tanggal 11 Juli 2021). Pandemi menyebabkan gangguan dalam pembangunan ekonomi, yang mengakibatkan kekurangan bahan makanan pokok, inflasi, penurunan Produk Domestik Bruto (PDB) dan ancaman terhadap kehidupan. Misalnya, pandemi yang serius dapat mengurangi PDB sebesar 3-4% [1, 2].

Sejak awal COVID-19, para peneliti dan dokter berusaha untuk mengurangi penyebaran COVID-19. Carrillo dkk. [3], menggunakan pembelajaran mesin tanpa pengawasan untuk mengklasifikasikan 155 negara yang memiliki profil COVID-19 yang serupa. Clustering dilakukan untuk kasus terkonfirmasi COVID-19. Prevalensi penyakit, populasi pria, indeks kualitas udara, metrik sosial ekonomi dan indikator sistem kesehatan digunakan sebagai variabel fitur. Klaster-klaster yang terbentuk memberikan wawasan tentang persamaan dan perbedaan antar negara dalam hal dampak COVID-19. Model ini gagal membuat stratifikasi negara berdasarkan tingkat kematian COVID-19. Karya lain oleh Farseev et al. [4], mencakup faktor ekonomi dan kesehatan yang serupa untuk penyebaran COVID-19. Studi ini mengungkap hubungan signifikan antara COVID-19 dan statistik nasional lainnya. Ini mengidentifikasi empat kelompok berdasarkan indikator ekonomi dan sistem kesehatan negara. Stojkoski dkk. [5], menyajikan analisis determinan sosial ekonomi COVID-19. Ini menentukan faktor sosial ekonomi, perawatan kesehatan, demografi dan lingkungan yang sedikit banyak terlibat dalam penyebaran COVID-19. Aliran kerja oleh Zarikas et al. [6], adalah pengenalan algoritma pengelompokan yang dirancang khusus untuk pengelompokan negara berdasarkan kasus aktif COVID-19, kasus aktif per populasi dan per wilayah mengikuti konsep analisis hierarkis. Hasilnya mengarah pada analisis bahwa negara-negara yang menghadapi dampak serupa COVID-19, memiliki faktor sosial, ekonomi, dan lainnya yang sama [7][8].

Analisis klaster bertujuan untuk mengelompokkan/ mengkategorikan data berdasarkan kesamaan. Jadi pengklasteran ini mengelompokkan data yang mempunyai kesamaan yang tinggi ke dalam satu kategori, sehingga kategori yang sama dapat mencapai homogenitas maksimum dan meminimalkan heterogenitas, sedangkan kategori yang berbeda mencapai homogenitas maksimum dan heterogenitas minimum [9]. Algoritma analisis pengklasteran secara garis besar terbagi menjadi tiga, yaitu: algoritma yang bertujuan menemukan partisi yang optimal untuk membagi data menjadi sejumlah klaster tertentu; algoritma yang bertujuan menemukan metode hierarki struktur pengelompokan; dan algoritma yang bertujuan menemukan metode berdasarkan model probabilitas untuk pemodelan klasterl [10]. K-means adalah metode pengklasteran yang paling banyak digunakan sejauh ini, yang ide utamanya adalah secara bertahap mengoptimalkan hasil pengklasteran dan secara konstan mendistribusikan kembali target dataset ke setiap pusat clustering untuk mendapatkan solusi yang optimal; dan keunggulan terbesarnya terletak pada kesederhanaan, kecepatan, dan objektivitasnya, yang banyak digunakan di berbagai bidang penelitian seperti pengolahan data, pengenalan citra, analisis pasar, dan evaluasi risiko [11]. Algoritma pengklasteran K-means

adalah memilih sejumlah K data sebagai centroid awal setiap kategori dan membaginya menjadi K kategori sesuai dengan prinsip bahwa data akan masuk ke dalam kategori yang yang mempunyai jarak terkecil [12].

Pada paper ini akan dilakukan pengklasteran terhadap 30 negara yang mempunyai jumlah kasus Covid 19 terbesar. Metode pengklasteran yang digunakan adalah K-Means. Untuk pengklasteran atribut-atribut yang digunakan adalah jumlah kasus per 1 juta penduduk, jumlah kematian per 1 juta penduduk, dan jumlah tes per 1 juta penduduk dari 30 negara tersebut.

2. METODE PENELITIAN

Bagian ini dapat meliputi analisa, arsitektur, metode yang dipakai untuk menyelesaikan masalah, implementasi.

2.1. Data

Dataset yang digunakan adalah data Covid 19 yang ada di https://www.worldometers.info/coronavirus/ pada tanggal 11 Juli 2021. Dataset ini memiliki banyak fitur, namun dalam penelitian ini hanya akan dipilih atribut yang yang meliputi: nama negara, jumlah total kasus, jumlah kasus per 1 juta penduduk, jumlah kematian per satu juta penduduk, dan jumlah pengetesan per satu juta penduduk. Data yang digunakan hanya data dari 30 negara yang mempunyai jumlah kasus terbesar. Data yang digunakan dapat dilihat pada tabel 1.

Tabel 1. Data Covid 19 dari 30 Negara

No	Negara	Jumlah Kasus Positif	Jumlah Kasus Positif per 1 juta penduduk	Jumlah Kematian per 1 juta penduduk	Jumlah Tes per 1 juta penduduk
1	USA	34726111	104287	1870	1536607
2	India	30837222	22124	293	309130
3	Brazil	19069003	89064	2489	253170
4	France	5808383	88784	1702	1467901
5	Russia	5758300	39441	974	1060970
6	Turkey	5476294	64226	589	739129
7	UK	5089893	74576	1881	3261985
8	Argentina	4639098	101691	2159	385211
9	Colombia	4492537	87341	2183	406624
10	Italy	4269885	70728	2116	1215255
11	Spain	3937192	84176	1732	1141512
12	Germany	3743138	44530	1092	776218
13	Iran	3355786	39438	1007	287195
14	Poland	2880755	76202	1988	468835
15	Mexico	2586721	19850	1803	59338
16	Indonesia	2491006	9010	237	78191
17	Ukraine	2240753	51554	1210	253546
18	South Africa	2179297	36278	1068	230526
19	Peru	2078815	62164	5809	437868
20	Netherlands	1719120	100102	1034	857713
21	Czechia	1669182	155572	2827	2863320
22	Chile	1585160	82201	1751	910995

No	Negara	Jumlah Kasus Positif	Jumlah Kasus Positif per 1 juta penduduk	Jumlah Kematian per 1 juta penduduk	Jumlah Tes per 1 juta penduduk
23	Philippines	1467119	13209	232	139354
24	Iraq	1421746	34557	426	295816
25	Canada	1420278	37297	694	980247
26	Belgium	1093700	93951	2165	1362811
27	Sweden	1092540	107495	1437	1087335
28	Romania	1081210	56590	1791	525398
29	Bangladesh	1009315	6067	97	41662
30	Pakistan	973284	4321	100	66841

2.2. Standarisasi

Yang dimaksud standarisasi data dalam penelitian ini adalah melakukan normalisasi data, sehingga data asli yang sebelumnya mempunyai rentang atau skala yang berbeda-beda setelah diolah mempunyai rentang atau skala yang sama. Nilai-nilai atribut dari dataset yang mempunyai rentang berbeda harus dinormalisasi terlebih dahulu, sehingga masing-masing atribut mempunyai rentang yang sama. Normalisasi dapat dilakukan dengan normalisasi min-max(0-1) atau normalisasi Z-score.

Rumus dari normalisasi min-max adalah sebagai berikut:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \quad (1)$$

v'= data baru, v= data lama, min_A=nilai terkecil suatu atribut, max_A=nilai terbesar suatu atribut, , new_min_A=nilai terkecil suatu atribut yang baru(=0), new_max_A=nilai terbesar suatu atribut yang baru(=1).

2.3. Pengklasteran K-Means

Algoritma atau metode mengklasteran yang dipakai adalah pengklasteran K-Means dengan langkah-langkah sebagai berikut:

Algoritma pengklasteran dengan metoda K-Means [13]:

1. Tentukan jumlah klaster=k
2. Tentukan titik pusat secara acak sebanyak k
3. Hitung jarak masing-masing data ke pusat klaster dengan rumus:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}\{x_{ij} - x_{jk}\}^2} \quad (2)$$

4. Alokasikan data ke dalam klaster berdasarkan jarak terdekat
5. Hitung pusat klaster baru yang dihitung berdasarkan data yang ada di masing-masing klaster
6. Ulangi langkah 3 sampai 5, sampai tidak ada data yang berpindah klaster

{ Menghitung total within-cluster variation dengan rumus

$$tot.withiness = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k}\sum_{x_i \in C_k}(x_i - \mu_k)^2 \quad (3)$$

tot.withiness sering pula disebut dengan within-cluster sum of squares(wss).

di mana x_i= data ke i pada klaster tertentu, dan μ_k adalah pusat dari klaster tersebut}

2.4. Kriteria Elbow

Kriteria Elbow diterapkan untuk mencari jumlah klaster yang optimal, dengan langkah-langkah sebagai berikut:

Algoritma penerapan kriteria Elbow [14];

1. Jalankan algoritma pengklasteran(misal K-Means) untuk jumlah klaster(k) yang berbeda-beda, misal k dari 1 sampai dengan n
2. Untuk setiap k, hitung nilai wss
3. Gambarkan grafik wss berdasarkan jumlah klaster(k)

4. Lokasi di mana terlihat ada belokan/tikungan pada plot umumnya dianggap sebagai indikator jumlah cluster yang tepat.

4. HASIL DAN PEMBAHASAN

Data pada tabel 1 setelah dilakukan proses normalisasi dengan menggunakan persamaan (1) untuk atribut: jumlah kasus per 1 juta penduduk, jumlah kematian per 1 juta penduduk, dan jumlah tes per 1 juta penduduk dapat dilihat pada tabel 2.

Tabel 2. Data Covid 19 dari 30 Negara dengan atribut yang dinormalisasi

No	Negara	Jumlah Kasus Positif per 1 juta penduduk	Jumlah Kematian per 1 juta penduduk	Jumlah Tes per 1 juta penduduk	Klaster
1	USA	0.6609	0.3104	0.4642	5
2	India	0.1177	0.0343	0.0831	4
3	Brazil	0.5603	0.4188	0.0657	4
4	France	0.5584	0.2810	0.4429	5
5	Russia	0.2322	0.1535	0.3165	3
6	Turkey	0.3961	0.0861	0.2166	3
7	UK	0.4645	0.3123	1.0000	2
8	Argentina	0.6438	0.3610	0.1067	4
9	Colombia	0.5489	0.3652	0.1133	4
10	Italy	0.4391	0.3535	0.3644	5
11	Spain	0.5280	0.2862	0.3415	5
12	Germany	0.2658	0.1742	0.2281	3
13	Iran	0.2322	0.1593	0.0762	4
14	Poland	0.4752	0.3311	0.1326	4
15	Mexico	0.1027	0.2987	0.0055	1
16	Indonesia	0.0310	0.0245	0.0113	1
17	Ukraine	0.3123	0.1949	0.0658	4
18	South Africa	0.2113	0.1700	0.0586	1
19	Peru	0.3824	1.0000	0.1230	4
20	Netherlands	0.6333	0.1640	0.2534	3
21	Czechia	1.0000	0.4779	0.8762	2
22	Chile	0.5149	0.2896	0.2700	3
23	Philippines	0.0588	0.0236	0.0303	1
24	Iraq	0.1999	0.0576	0.0789	4
25	Canada	0.2180	0.1045	0.2915	3
26	Belgium	0.5926	0.3620	0.4103	5
27	Sweden	0.6821	0.2346	0.3247	3
28	Romania	0.3456	0.2966	0.1502	4
29	Bangladesh	0.0115	0.0000	0.0000	1

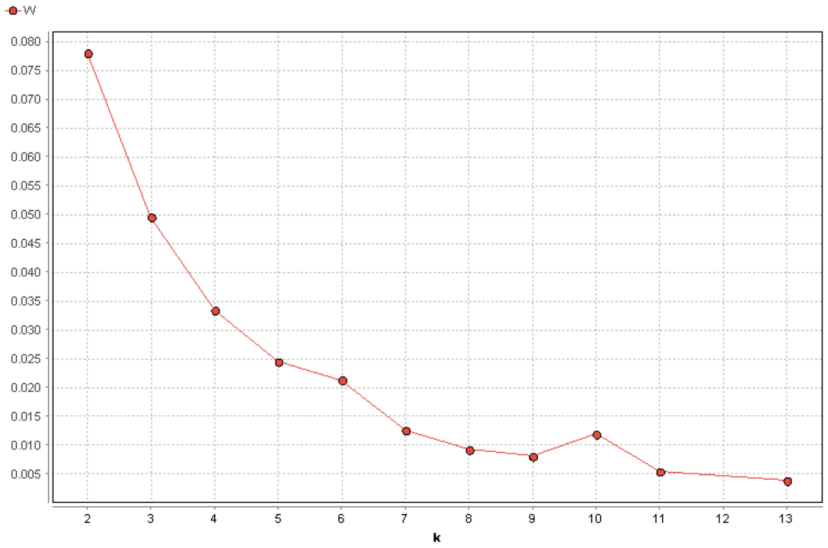
No	Negara	Jumlah Kasus Positif per 1 juta penduduk	Jumlah Kematian per 1 juta penduduk	Jumlah Tes per 1 juta penduduk	Klaster
30	Pakistan	0.0000	0.0005	0.0078	1

Pengklasteran K-Means akan dilakukan terhadap tiga atribut yang sudah dinormalisasi tersebut. Seperti telah disebutkan pada langkah-langkah algoritman K-Means, bahwa jumlah klaster harus ditentukan di awal. Pada eksperimen ini akan dilakukan pengklasteran dengan jumlah klaster dari 2 sampai 13, dan didapatkan nilai wss seperti pada tabel 3.

Tabel 3. Jumlah klaster(k) dan wss-nya(w)

k	wss
1	0.078
2	0.050
3	0.033
4	0.024
5	0.021
6	0.013
7	0.009
8	0.008
9	0.012
10	0.005
11	0.004
12	0.078
13	0.050

Dari tabel 3 akan disajikan dalam gambar 1 sebagai berikut:



Gambar 1. Penerapan kriteria Elbow

Dari penerapan kriteria Elbow seperti terlihat pada gambar 1, dapat disimpulkan jumlah klaster optimal adalah k=5 (pada gambar terlihat ada tikungan). Karena jumlah k optimal adalah 5, maka pengklasteran data yang

tersaji pada tabel 2, dilakukan pengklasteran dengan jumlah klaster adalah 5. Hasil dari pengklasteran dapat dilihat pada tabel 2 kolom ke 5. Karakteristik dari hasil pengklasteran untuk masing-masing klaster adalah sebagai berikut:

Tabel 3. Jumlah klaster(k) dan wss-nya(w)

Klaster	Negara	Karakteristik
1	Mexico, Indonesia, South Africa, Philippines, Bangladesh, Pakistan	jumlah pengetesan per 1 juta penduduk sangat rendah
2	UK, Czechia	jumlah pengetesan per 1 juta penduduk sangat tinggi
3	Turkey, Germany, Netherlands, Chile, Canada, Sweden	jumlah kasus per 1 juta penduduk rendah, jumlah tes per 1 juta penduduk rendah
4	India, Brazil, Argentina, Colombia, Iran, Poland, Ukraine, Peru, Iraq, Romania	jumlah kasus per 1 juta penduduk sedang, jumlah pengetesan rendah
5	USA, France, Italy, Spain, Belgium	jumlah kasus per 1 juta penduduk tinggi, jumlah pengetesan per 1 juta penduduk tinggi

Dari tabel 3 dapat disimpulkan dari 30 negara dengan jumlah kasus terbesar, mayoritas negara-negara yang melakukan jumlah tes per 1 juta penduduk tinggi adalah negara-negara di kawasan Eropa dan Amerika Utara, sedangkan mayoritas negara-negara di kawasan Asia dan Amerika Selatan, jumlah pengetesan per 1 juta penduduk masih rendah.

5. KESIMPULAN

Paper ini bertujuan untuk melakukan pengklasteran K-Means terhadap 30 negara dengan jumlah kasus terbesar. Atribut yang digunakan untuk pengklasteran adalah: jumlah kasus per 1 juta penduduk, jumlah kematian per 1 juta penduduk, dan jumlah tes per 1 juta penduduk. Untuk memperoleh jumlah klaster optimal dilakukan penerapan kriteria Elbow. Hasil penerapan kriteria Elbow ini diperoleh jumlah klaster optimal adalah 5. Dengan menggunakan jumlah klaster=5, ke 30 negara terbagi dalam lima kelompok negara berdasar dua atribut yang menonjol yaitu jumlah kasus per 1 juta penduduk dan jumlah tes per 1 juta penduduk. Mayoritas negara-negara yang melakukan jumlah tes per 1 juta penduduk tinggi adalah negara-negara di kawasan Eropa dan Amerika Utara, sedangkan mayoritas negara-negara di kawasan Asia dan Amerika Selatan, jumlah pengetesan per 1 juta penduduk masih rendah.

DAFTAR PUSTAKA

[1] Prager, F., Wei, D. and A. Rose, A., 2017, Total economic consequences of an in-fluenza outbreak in the united states, *Risk Analysis*, vol 37, hal 4–19..

[2] Rizvi, S.A., Umair, M., Cheema, M.A., 2021, Clustering of Countries for COVID-19 Cases based on Disease Prevalence, *Health Systems and Environmental Indicators*. https://doi.org/10.1101/2021.02.15.21251762

[3] Carrillo-Larco, R.M., Castillo-Cara,M., 2020, Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach, *Wellcome Open Research 5*, vol 56.

[4] Farseev, Y.-Y., Chu-Farseeva, Q., and Yang, D. B. Loo, 2020, *Understanding economic and health factors impacting the spread of COVID-19 disease*.

[5] Stojkoski, V., Utkovski, Z., Jolakoski, P.,Tevdovski, D., and Kocarev, L., 2020, *The socio-economic determinants of the coronavirus disease (COVID-19) pandemic*.

[6]] Zarikas,V., Poulopoulos, S. G. , Gareiou, Z., and Zervas, E., 2020, Clustering analysis of countries using the COVID-19 cases dataset, *Data in Brief*, vol 31.

[7] Agrebi,S., and Larbi, A., 2020, Use of artificial intelligence in infectious diseases, *Artificial Intelligence in Precision Health*, hal 415–438.

[8] Carrillo-Larco, R.M, and Castillo-Cara, M., 2020, Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach, *Wellcome Open Research*, vol 5.

[9] Hossain, M.Z., Akhtar, M.N., R. B. Ahmad, R.B., and Rahman, M., 2019, A dynamic K-means clustering for data mining, *Indonesian Journal of Electrical Engineering and Computer Science*, vol 13, hal 521–526.

[10] Jothi, R., Mohanty, S.K., and Ojha, A., 2019, DK-means: a deterministic k-means clustering algorithm for gene expression analysis, *Pattern Analysis and Applications*, vol 22, hal 649–667.

[11] Shakeel, P.M., Baskar, S., Dhulipala, V. S. , and Jaber, M.M., 2018, Cloud based framework for diagnosis of diabetes mellitus using K-means clustering, *Health Information Science and Systems*, vol 6, hal 1–7.
[12] Slamet, C., Rahman, A., Ramdhani, M.A., and Darmalaksana, W., 2016, Clustering the verses of the Holy Qur'an using K-means algorithm, *Asian Journal of Information Technology*, vol 15, hal 5159–5162.
[13] Hancer, E. , Xue, B., and Zhang, M., 2020. A survey on feature selection approaches for clustering, *Artificial Intelligence Review*, vol 53(6), hal 4519–4545.
[14] Shmueli, G., Bruce, P.C., Yahav, I., and Patel, N.R., 2020, Data Mining for Business Analytics: Concepts, Techniques, and Applications. John Wiley & Sons, New Jersey.