

Penerapan Data Mining dalam Mengelompokkan Jumlah Kematian Penderita COVID-19 Berdasarkan Negara di Benua Asia

Noviyanto

Universitas Gunadarma
e-mail: viyan@staff.gunadarma.ac.id

Abstrak - COVID-19 disebabkan oleh SARS-CoV-2, yaitu virus jenis baru dari coronavirus (kelompok virus yang menginfeksi sistem pernapasan). Infeksi virus Corona bisa menyebabkan infeksi pernapasan ringan sampai sedang, seperti flu, atau infeksi sistem pernapasan dan paru-paru, seperti pneumonia. Maraknya penyebaran penyakit yang diakibatkan oleh virus COVID-19 yang telah ditetapkan sebagai pandemi oleh WHO pada tanggal 12 Maret 2020, akibat virus COVID-19 banyak pasien yang terjangkit mengalami kematian. Dalam mengelompokkan jumlah kematian penderita Covid-19 menggunakan teknik data mining metode k-means clustering. Data diambil dari link <https://www.worldometers.info/coronavirus/>. Hasil dari penelitian ini adalah cluster jumlah kematian penderita Covid-19 kedalam 3 cluster. terdapat 4 negara dengan cluster tingkat tinggi yaitu: Turki, Iran, India dan China dengan cluster tingkat sedang sebanyak 4 negara yaitu : Pakistan, Indonesia, Jepang, dan Piliphina dan dengan cluster rendah adalah 41 negara lainnya.

Kata kunci: Data mining, Clustering, K-Means, Covid-19

Abstract - COVID-19 is caused by SARS-CoV-2, a new type of virus from coronavirus (a group of viruses that infects the respiratory system). Corona virus infections can cause mild to moderate respiratory infections, such as flu, or respiratory and lung system infections, such as pneumonia. The rise of the spread of diseases caused by the COVID-19 virus which has been established as a pandemic by WHO on March 12, 2020, due to the COVID-19 virus many patients who have contracted have died. In classifying the number of deaths of patients with Covid-19 using data mining techniques k-means clustering method. Data is taken from the link <https://www.worldometers.info/coronavirus/>. The results of this study are the cluster number of deaths of patients with Covid-19 into 3 clusters. there are 4 countries with high level clusters, namely: Turkey, Iran, India and China with medium level clusters as many as 4 countries namely: Pakistan, Indonesia, Japan, and the Philippines and with low clusters are 41 other countries.

Keywords: Data mining, Clustering, K-Means, Covid-19

PENDAHULUAN

Data mining adalah proses menganalisa data dari perspektif yang berbeda dan menyimpulkannya menjadi informasi-informasi penting yang dapat dipakai untuk meningkatkan keuntungan, memperkecil biaya pengeluaran, atau bahkan keduanya. Secara teknis, data mining dapat disebut sebagai proses untuk menemukan korelasi atau pola dari ratusan atau ribuan field dari sebuah relasional database yang besar (Angga Ginanjar Mabur, 2012).

Pengelompokan Data Mining dibagi menjadi beberapa kelompok yaitu :

a. Deskripsi

Deskripsi merupakan cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data yang dimiliki.

b. Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variable target estimasi lebih ke arah numerik

daripada ke arah kategori. Model yang dibangun menggunakan record lengkap yang menyediakan nilai variable target sebagai nilai prediksi.

c. Prediksi

Prediksi menerka sebuah nilai yang belum diketahui dan juga memperkirakan nilai untuk masa mendatang.

d. Klasifikasi

Dalam klasifikasi terdapat target variable kategori, misal penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu tinggi, sedang, dan rendah.

e. Pengklasteran

Merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan.

f. Asosiasi

Asosiasi bertugas menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum

disebut analisis keranjang belanja (Maulana & Fajrin, 2018).

Maraknya penyebaran penyakit yang diakibatkan oleh virus COVID-19 yang telah ditetapkan sebagai pandemi oleh WHO pada tanggal 12 Maret 2020, akibat virus COVID-19 banyak pasien yang terjangkit mengalami kematian. Data yang diperoleh dari link <https://www.worldometers.info/coronavirus/> menunjukkan angka kematian sebesar 329,731 dari seluruh negara-negara di benua Asia berdasarkan data update Mei 21, 2020, 03:12 GMT. Begitu juga di negara-negara benua Asia banyak yang mengalami kematian. Data kematian akibat penyebaran virus COVID-19 menunjukkan angka sebesar 25,851. Berdasarkan data yang di peroleh dapat dilakukan pengelompokan angka kematian penyebaran virus COVID-19 yang nanti dapat menjadi masukan kepada negara-negara di benua Asia agar meningkatkan kualitas penanganan pasien penderita virus COVID-19.

1. Rumusan Masalah

- a. Meningkatnya angka kematian akibat penyebaran virus COVID-19 di negara-negara Asia.
- b. Perlu adanya pengelompokan data kematian akibat penyebaran virus COVID-19.

2. Batasan Masalah

Supaya pembahasan pada penelitian ini tidak terlalu luas dan lebih fokus maka dibatasi hanya pada :

1. Penambahan data dari <https://www.worldometers.info/coronavirus/>
2. Pengelompokan data angka kematian akibat penyebaran virus COVID-19 di negara-negara Asia
3. Uji coba data yang digunakan menggunakan software RapidMiner .

4. Tujuan Penelitian

Penelitian ini bertujuan menghasilkan cluster data kematian akibat penyebaran virus COVID-19 berdasarkan negara-negara di Asia menggunakan teknik data mining metode K-Means Clustering.

a. Data Mining

Data mining adalah analisa terhadap data untuk menemukan hubungan yang jelas serta menyimpulkannya yang belum diketahui sebelumnya dengan cara terkini dipahami dan berguna bagi pemilik data tersebut. Data mining adalah metoda yang digunakan untuk mengekstraksi informasi prediktif tersembunyi pada database, ini adalah teknologi yang sangat potensial bagi perusahaan yang sangat potensial bagi perusahaan dalam memberdayakan data

warehouse. Secara garis besar, data mining dapat dikelompokkan menjadi 2 kategori utama, yaitu:

1. Deskriptive mining, yaitu proses untuk menemukan karakteristik penting dari data dalam satu basis data. Teknik data mining yang termasuk descriptive mining adalah clustering, asosiasi, dan sequential mining.
2. Predictive, yaitu proses untuk menemukan pola dari data dengan menggunakan beberapa variable lain di masa depan. Salah satu teknik yang terdapat dalam predictive mining adalah klasifikasi. Secara sederhana data mining biasa dikatakan sebagai proses penyaringan atau “menambang” pengetahuan dari sejumlah data yang besar. Istilah lain untuk data mining adalah Knowledge Discovery in Database (KDD) (Darmi & Setiawan, 2016).

b. Pengelompokan Data Mining

Algoritma Clustering mengelompokkan satu set dokumen ke dalam himpunan bagian atau kluster. Tujuan algoritma kluster adalah menciptakan kluster yang koheren secara internal, tetapi jelas berbeda satu sama lain. Dengan kata lain, dokumen dalam sebuah kluster harus semirip mungkin; dan dokumen dalam satu kluster harus berbeda mungkin dari dokumen dalam kluster lainnya.

Clustering merujuk pada pengelompokan dokumen, observasi atau kasus pada kelas yang objeknya mirip. Kluster adalah kumpulan dokumen yang mirip satu sama lain dan berbeda dengan dokumen pada kluster lain. Clustering berbeda dengan Clasification, pada clustering tidak ada target variabel untuk dikelompokkan. Algoritma clustering mencoba untuk membagi kumpulan data menjadi kluster yang anggotanya relatif sama, dimana kemiripan dokumen di kluster yang sama tinggi, dan kemiripan dokumen di kluster lain kecil. Dengan kata lain, seperti pada Gambar 1, Algoritma clustering mencoba membuat kluster dokumen yang between-clusters variation (BCV) lebih tinggi dibandingkan dengan within-cluster variation (WCV) (Robani & Widodo, 2016).

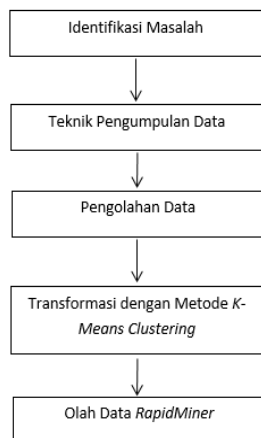
c. K-Means Clustering

K-means adalah salah satu bentuk pengelompokan yang paling sederhana. Prosedurnya sederhana dan mudah untuk mengklasifikasikan data yang diberikan melalui sejumlah cluster. Penentuan centroid dilakukan dengan cara mengambil data pertama sebagai centroid pertama, data kedua sebagai centroid kedua, dan seterusnya hingga jumlah centroid yang diperlukan. Langkah berikutnya adalah dengan menghitung jarak dari titik yang akan di cluster ke setiap centroid yang ada dan dikelompokkan sesuai dengan jarak terdekat kepada centroid-nya. Bila semua titik

sudah masuk kedalam pengelompokan maka langkah pertama selesai. Kemudian langkah berikutnya, kita perlu menghitung kembali k-centroid baru sebagai barycenters dari kelompok yang dihasilkan. Setelah memiliki k-centroid yang baru, pengelompokan di uji kembali terhadap k-centroid (Kristanto, A, & S, 2016).

METODOLOGI PENELITIAN

Pada rancangan penelitian ini penulis akan menguraikan metodologi dan kerangka penelitian kerja yang digunakan dalam menyelesaikan masalah penelitian. Metodologi penelitian ini digunakan secara sistematis agar mendapatkan alur kerja yang baik sehingga dapat digunakan sebagai pedoman untuk peneliti selanjutnya yang akan melaksanakan penelitian ini agar hasil yang dicapai tidak menyimpang dari tujuan yang di inginkan dan dapat terlaksana dengan lebih baik dan sesuai dengan tujuan yang telah ditetapkan. Berikut diagram alir model penelitian disajikan dalam rancangan *flowchart* dapat dilihat dari gambar 1.



Gambar 1. *Flowchart* Kerangka Kerja Penelitian

Proses alur *flowchart* pada gambar 1 adalah sebagai berikut :

1. Identifikasi Masalah
 Pengenalan suatu masalah, dan tahap awal dalam proses penelitian. Meningkatnya angka kematian akibat penyebaran virus COVID-19 di negara-negara di benua Asia.
2. Teknik Pengumpulan Data
 Data penelitian ini diperoleh dari website dengan situs url <https://www.worldometers.info/coronavirus/> mengenai data penyebaran virus COVID-19.
3. Pengolahan Data
 Pada tahapan ini, data diolah untuk mendapatkan hasil yang kemudian dapat dikelola ke tahap berikutnya, sehingga menghasilkan informasi yang tepat.
4. Transformasi dengan *K-Means Clustering*
 Transformasi data, data yang berjenis alfabet seperti nama negara terlebih dahulu harus

dilakukan proses inialisasi data terlebih dahulu ke dalam bentuk angka/*numerical*. Kemudian mengelompokkan data yang sudah ada dalam tiga kelompok yaitu tinggi, sedang, dan rendah dengan menggunakan metode *k-means clustering*.

5. Olah Data dengan *RapidMiner*

Dalam pengujian data menggunakan sebuah aplikasi *RapidMiner*. Dengan menggunakan *software RapidMiner* kita akan membandingkan bagaimana hasil pengolahan data secara manual dengan pengolahan data menggunakan sebuah *software*.

HASIL DAN PEMBAHASAN

1. Analisis Data

Analisis data merupakan cara yang digunakan untuk menjawab masalah penelitian melalui prosedur pengelolannya terutama masalah yang berkaitan dengan penelitian. Data yang diperoleh untuk penelitian ini adalah data kematian penderita Covid-19 berdasarkan Negara-Negara di Benua Asia pada tanggal 21 Mei 2019.

No	Negara	Total Kematian	No	Negara	Total Kematian
1	Turkey	4,222	26	Uzbekistan	13
2	Iran	7,183	27	Tajikistan	41
3	India	3,434	28	Kyrgyzstan	14
4	China	4,634	29	Maldives	4
5	Saudi Arabia	339	30	Hong Kong	4
6	Pakistan	985	31	Sri Lanka	9
7	Qatar	16	32	Lebanon	26
8	Singapore	22	33	Cyprus	17
9	Bangladesh	386	34	Georgia	12
10	UAE	233	35	Jordan	9
11	Indonesia	1,242	36	Taiwan	7
12	Kuwait	124	37	Nepal	2
13	Israel	279	38	Palestine	2
14	Japan	768	39	Vietnam	0
15	Philippines	842	40	Myanmar	6
16	S. Korea	264	41	Yemen	30
17	Afghanistan	187	42	Brunei	1
18	Bahrain	12	43	Mongolia	0
19	Malaysia	114	44	Cambodia	0
20	Kazakhstan	35	45	Syria	3
21	Oman	30	46	Macao	0
22	Armenia	67	47	Timor-Leste	0
23	Iraq	134	48	Bhutan	0
24	Azerbaijan	43	49	Laos	0
25	Thailand	56			

2. Perhitungan Manual K-Means Clustering

- a. Menentukan jumlah data yang akan di cluster, dengan sampel data kematian pasien covid-19 Asia menurut negara dengan jumlah data sebanyak 49 Negara.
- b. Menentukan nilai k jumlah *cluster* sebanyak 3*cluster* yaitu *cluster* tinggi dan *cluster* rendah.
- c. Menentukan nilai *centroid* awal yang telah ditentukan secara random, *cluster* tinggi (C1), *cluster* sedang (C2), dan *cluster* rendah (C3).

Berikut adalah nilai *centroid* data awal pada iterasi 1:

Tabel 2. Centroid Awal Iterasi 1

C1	Maximum	985,000
C2	Average	105,239
C3	Minimum	0,000

- d. Setelah data nilai pusat *cluster* ditentukan, selanjutnya menghitung jarak setiap data terhadap pusat *cluster* yang dilakukan dengan titik pusan pada *cluster* pertama. Berikut adalah perhitungannya:

$$D (1.1) = \sqrt{(985,000 - 4,222)^2} = 980,778$$

Perhitungan dilakukan sampai D(1.49). Selanjutnya perhitunggan pada *cluster* 2 rata-rata dengan titik pusat centroid:

$$D (2.1) = \sqrt{(105,239 - 4,222)^2} = 101,017$$

Perhitungan dilakukan sampai D (2.49). Selanjutnya perhitunggan pada *cluster* 3 rata-rata dengan titik pusat centroid:

$$D (3.1) = \sqrt{(0,000 - 4,222)^2} = 4,222$$

Perhitungan dilakukan sampai D (3.49). Sehingga diperoleh hasil dari perhitungan data dengan titik pusat *centroid* setiap *cluster* sebagai berikut:

Tabel 3. Contoh Perhitungan Centroid Pada Setiap Cluster (Iterasi 1)

Negara Asal	Total Kematian	C1	C2	C3
Turkey	4,222	980,778	101,017	4,222
Iran	7,183	977,817	98,056	7,183
India	3,434	981,566	101,805	3,434
China	4,634	980,366	100,605	4,634
Saudi Arabia	339	646,000	233,761	339,000
Pakistan	985	0,000	879,761	985,000
Qatar	16	969,000	89,239	16,000
Singapore	22	963,000	83,239	22,000
Bangladesh	386	599,000	280,761	386,000
UAE	233	752,000	127,761	233,000

Lalu hitung jarak terdekat dari *cluster* dengan menggunakan Euclidean Distance, seperti hasil dibawah ini:

Tabel 4. Jarak Terdekat (Iterasi 1)

Jarak Terpendek	Hasil	C1	C2	C3
4,222	C3			1
7,183	C3			1
3,434	C3			1
4,634	C3			1

233,761	C2		1	
0,000	C1	1		
16,000	C3			1
22,000	C3			1
280,761	C2		1	
127,761	C2		1	

Sehingga hasil dari perhitungan jarak data ke titik pusat *cluster* pada iterasi 1 yaitu:

Tabel 5. Hasil Cluster Pada Iterasi 1

Cluster	Hasil
C1	3
C2	4
C3	5

- e. Selanjutnya lakukan kembali langkah 4 – 5 jika nilai hasil *centroid* dari iterasi 1 dengan nilai *centroid* selanjutnya bernilai sama ataupun nilai *centroid* sudah optimal dan posisi *cluster* tidak mengalami perubahan lagi maka proses iterasi berhenti. Akan tetapi jika posisi *cluster* masih berubah-ubah maka proses iterasi terus berlanjut sampai hasil bernilai sama.
- f. Menghitung titik pusat terbaru dengan menggunakan hasil dari setiap *cluster* pada iterasi 1. Berikut adalah contoh perhitungan titik pusat *cluster* baru pada iterasi 2:

$$C1 = \frac{985 + 768 + 842}{3} = 865,000$$

$$C2 = \frac{339 + 386 + 233 + 124 + 279 + 264 + 187 + 114 + 56}{9} = 198,455$$

$$C3 = \frac{4,222 + 7,183 + 3,434 + 4,634 + 16 + 22 + 339 + 16 + 22 + 1,242 + 30 + 43 + 13 + 1 + 14 + 4 + 4 + 9 + 26 + 17 + 12 + 9 + 7 + 2 + 2 + 0 + 6 + 30 + 1 + 0 + 0 + 3 + 0 + 0 + 0 + 0}{37} = 10,820$$

Setelah dilakukan perhitungan secara manual maka dihasilkan sebuah cluster jumlah kematian penderita covid-19.

Tabel 5. Hasil pengelompokkan Jumlah kematian penderita covid-19

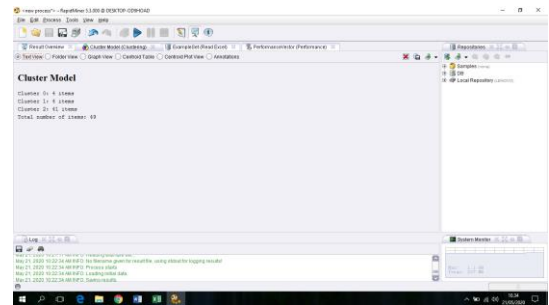
No	Negara	Hasil	C1	C2	C3
1	Turkey	C3	1		
2	Iran	C3	1		
3	India	C3	1		
4	China	C3	1		
5	Saudi Arabia	C2			1

6	Pakistan	C1		1	
7	Qatar	C3			1
8	Singapore	C3			1
9	Bangladesh	C2			1
10	UAE	C2			1
11	Indonesia	C3		1	
12	Kuwait	C3			1
13	Israel	C2			1
14	Japan	C1		1	
15	Philippines	C1		1	
16	S. Korea	C2			1
17	Afghanistan	C2			1
18	Bahrain	C3			1
19	Malaysia	C3			1
20	Kazakhstan	C3			1
21	Oman	C3			1
22	Armenia	C3			1
23	Iraq	C3			1
24	Azerbaijan	C3			1
25	Thailand	C3			1
26	Uzbekistan	C3			1
27	Tajikistan	C3			1
28	Kyrgyzstan	C3			1
29	Maldives	C3			1
30	Hong Kong	C3			1
31	Sri Lanka	C3			1
32	Lebanon	C3			1
33	Cyprus	C3			1
34	Georgia	C3			1
35	Jordan	C3			1
36	Taiwan	C3			1
37	Nepal	C3			1
38	Palestine	C3			1
39	Vietnam	C3			1
40	Myanmar	C3			1
41	Yemen	C3			1
42	Brunei	C3			1
43	Mongolia	C3			1
44	Cambodia	C3			1
45	Syria	C3			1
46	Macao	C3			1
47	Timor-Leste	C3			1
48	Bhutan	C3			1

49	Laos	C3			1
----	------	----	--	--	---

3. Tampilan RapidMiner

Untuk mendapatkan hasil pengelompokan pada tahap selanjutnya dilakukan dengan mengklik tanda panah biru yang terdapat di posisi tengah atas pada *toolbar*. Tahap ini akan menampilkan hasil akhir serta langkah terakhir dalam penggunaan tools rapidminer. Dapat dilihat pada gambar 1:



Gambar 1. Nilai Cluster Model

Keterangan :

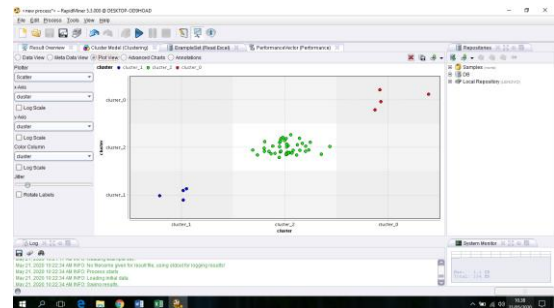
Jumlah Cluster 1 (Tinggi) berjumlah 4 Items

Jumlah Cluster 2 (Rendah) berjumlah 41 Items

Jumlah Cluster 0 (Sedang) berjumlah 4 Items

Jumlah keseluruhan items adalah 49.

Sehingga dapat diketahui hasil pengelompokan rapidminer dapat dilihat pada gambar sebagai berikut :



Gambar 2. Hasil Pengelompokan

Berdasarkan pada gambar dapat dilihat bahwa kelompok tinggi memiliki 4 node, kelompok rendah memiliki 41 node, dan kelompok sedang 4 node.

KESIMPULAN

Adapun kesimpulan yang dapat di ambil pada tehnik data mining dalam clustering total kematian penderita Covid-19 berdasarkan negara-negara di Benua Asia dengan algoritma k-means ini adalah sebagai berikut:

1. Algoritma k-means clustering ini dapat membantu peneliti mengelompokkan clustering total kematian penderita Covid-19 berdasarkan negara-negara di Benua Asia. Dari data clustering total kematian penderita Covid-19 berdasarkan negara-negara di Benua Asia dapat di ketahui terdapat 4 negara

dengan *cluster* tingkat tinggi yaitu: Turki, Iran, India dan China dengan *cluster* tingkat sedang sebanyak 4 negara yaitu : Pakistan, Indonesia, Jepang, dan Piliphina dan dengan cluster rendah adalah 41 negara lainnya.

2. Aplikasi rapidminer ini dapat membantu peneliti mengelompokkan total kematian penderita Covid-19 berdasarkan negara di Benua Asia. Perhitungan manual algoritma *k-means clustering* dalam mengelompokkan total kematian penderita Covid-19 berdasarkan negara-negara di Benua Asia dan penerapan di *Rapidminer* menunjukkan hasil yang sama.

REFERENSI

- Angga Ginanjar Mabur, L. R. (2012). Penerapan Data Mining Untuk Memprediksi Kriteria Nasabah Kredit. *Jurnal Komputer Dan Informatika (KOMPUTA)*, 1(1), 53–57.
- Darmi, Y., & Setiawan, A. (2016). Penerapan Metode Clustering K-Means Dalam. *Y. Darmi, A. Setiawan*, 12(2), 148–157.
- Kristanto, N. H., A, A. C. L., & S, H. B. (2016). Implementasi K-Means Clustering untuk Pengelompokan Analisis Rasio Profitabilitas dalam Working Capital. *Juisi*, 02(01), 9–15.
- Maulana, A., & Fajrin, A. A. (2018). Penerapan Data Mining Untuk Analisis Pola Pembelian

Konsumen Dengan Algoritma Fp-Growth Pada Data Transaksi Penjualan Spare Part Motor. *Klik - Kumpulan Jurnal Ilmu Komputer*, 5(1), 27.

<https://doi.org/10.20527/klik.v5i1.100>

- Robani, M., & Widodo, A. (2016). Algoritma K-Means Clustering Untuk Pengelompokan Ayat Al Quran Pada Terjemahan Bahasa Indonesia. *Jurnal Sistem Informasi Bisnis*, 6(2), 164. <https://doi.org/10.21456/vol6iss2pp164-176>

PROFIL PENULIS

Noviyanto, menyelesaikan pendidikan Program Sarjana Teknik Informatika pada Fakultas Teknologi Industri Universitas Gunadarma, dan melanjutkan Program Pascasarjana Magister Manajemen Sistem Informasi (S-2) pada Program Pasca Sarjana Universitas Gunadarma. Saat ini Penulis menjabat sebagai Kepala Sub Bagian Sistem Informasi dan Kerja Sama, Lembaga Layanan Pendidikan Wilayah III Jakarta, Kementerian Pendidikan dan Kebudayaan. Penulis juga mengajar sebagai Dosen NIDK pada Fakultas Ilmu Komputer Universitas Gunadarma.