ANALISIS *CLUSTER* UNTUK PEMETAAN DATA KASUS *COVID* – 19 DI INDONESIA MENGGUNAKAN *K -MEANS*

Nurissaidah Ulinnuha[i]

Siti Azizatus Sholihah[ii]

[i] *Prodi Matematika FST, UINSA, nuris.ulinnuha@uinsby.ac.id*
[ii] *Mahasiswa Prodi Matematika FST, UINSA, h72217039@uinsby.ac.id*

ABSTRAK, Indonesia merupakan salah satu negara yang terjangkit virus *Covid – 19. Covid – 19* merupakan penyakit yang dapat menular yang ditandai dengan gejala pada bagian pernapasan. Oleh karena itu, di masa pandemi ini sangat penting untuk menghindari wilayah dengan persebaran *Covid – 19* yang tinggi. Pada penelitian ini dilakukan *clustering* penyebaran *Covid – 19* di Indonesia dengan menerapkan metode data mining. Pengelompokan dilakukan berdasarkan parameter jumlah pasien positif, sembuh, meninggal, suspect, probable, dan negatif. Salah satu cara untuk melihat perkembangan *Covid – 19* di Indonesia dapat menggunakan algoritma *K – Means* yang menggunakan beberapa kelompok. Data – data tanpa label diterima oleh algoritma *K – Means* ini. Penelitian ini menggunakan algoritma *K – Means* untuk menentukan bagaimana tingkat penyebaran Covid – 19 di setiap provinsi di Indonesia. Validasi *silhouette index* (SI) digunakan untuk menentukan *cluster* optimal. Hasil penelitian menunjukkan cluster optimal terletak pada k = 2 dengan nilai SC = 0,74 yang menunjukkan bahwa struktur *cluster* termasuk kuat. Berdasarkan hasil *cluster* optimal, didapat 2 kelompok yaitu kelompok rawan yang terdiri dari provinsi Banten, DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur, dan Riau, dan terakhir kelompok aman yang terdiri dari 28 provinsi lainnya.

***Kata Kunci**: Indonesia, Analisis cluster, K – Means, Covid – 19,* dan *silhouette coefficient*

1. PENDAHULUAN

Seluruh negara termasuk Indonesia saat ini digemparkan oleh suatu masalah yang diakibatkan virus *Severse Acute Respiratory Syndrome Coronavirus – 2* (SARS – CoV – 2)[1]. Permasalahan ini berawal dari kota Wuhan, Hubei, China pada akhir tahun 2019 yang membuat kepanikan karena sudah memakan banyak korban jiwa.

Pola penindakan *Covid – 19* seperti Pemberlakuan Sosial Berskala Besar (PSBB) dan *Social Distancing* belum terlaksana secara maksimal. Misalnya, kebanyakan masyarakat tidak memakai masker waktu keluar rumah. Penggunaan masker yang benar tidak hanya saat memakainya, tetapi masker harus sudah sesuai standart kesehatan.

Informasi sebaran dalam bentuk peta jauh lebih mudah untuk dipahami oleh masyarakat daripada hanya penyampaian informasi tabular atau jumlah kasus saja. Dengan penyajian data sebaran diharapkan dapat menjadi alat bantu dalam mengambil keputusan terkait, apakah wilayah perlu mengambil tindakan *lockdown* atau belum.

Pada penelitian ini, peneliti ingin melakukan *clustering* dengan menggunakan metode *K – Means* untuk melihat bagaimana perkembangan kasus *Covid – 19* di Indonesia, dan mengetahui daerah – daerah mana saja yang sekarang rawan *Covid – 19*. Algoritma *K – Means* ini dipakai karena algoritma ini mempunyai keakuratan yang validasi *clustering*nya cukup baik, dan relatif lebih praktis dalam mengelompokan data dengan kuantitas yang besar.

2. TINJAUAN PUSTAKA

Analisis *cluster* merupakan sebuah metode data mining yang mengelompokan suatu objek atau kasus menjadi kelompok – kelompok yang lebih kecil dimana setiap kelompok berisi objek yang mirip satu sama lain. Analisis *cluster* berguna untuk meringkas data dengan cara mengelompokan objek – objek berdasarkan kesamaan karakteristik.

Dalam pengelompokannya, digunakan suatu ukuran yang dapat menerangkan kedekatan antar data, yaitu ukuran jarak atau similaritas[2]. Ukuran jarak yang sering digunakan dalam analisis *cluster* adalah ukuran jarak euclidean karena mengukur jarak dari dua buah titik dalam dua dimensi, tiga dimensi bahkan lebih[3].

Proses analisis *cluster* meliputi[4] :

1. Menentukan ukuran kemiripan antara kedua objek.
   Proses ini mengukur seberapa jauh ada kesamaan antar objek
2. Melakukan proses standarisasi data (jika diperlukan)
   Standarisasi data dilakukan jika jarak nilai dari masing – masing variabel mengalami perbedaan skala. Salah satu metode yang digunakan untuk standarisasi data yaitu metode *Min – Max Normalization* dengan rumus[5] :

$$v' = \frac{v - min_A}{max_A - min_A} \quad (1,1)$$

   Dimana :
   v' = Data hasil normalisasi.
   v = Data asli.
   max_A = Nilai maksimal dari data.
   min_A = Nilai minimum dari data.
3. Melakukan peng*clusteran*.
   Pengelompokan dilakukan dengan melihat kararkteristik objek melalui perhitungan ukuran jarak.
4. Melakukan validasi *cluster*
   Cluster yang terbentuk kemudian diuji apakah hasilnya valid.

ASUMSI ANALISIS CLUSTER

Salah satu uji asumsi cluster yaitu tidak terdapat multikolinearitas. Multikolinearitas adalah hubungan linear yang sempurna atau pasti diantara beberapa atau semua variabel[6]. Salah satu cara indentifikasi multikolinearitas adalah menghitung nilai *Varians Inflantion Factor* (VIF) dengan rumus[2]:

$$VIF_i = \frac{1}{1 - R_i^2} \quad (1,2)$$

Dimana R_i^2 adalah nilai koefisien determinasi antar variabel Independen. Jika VIF>10 maka variabel tersebut terindikasi multikolinearitas. Untuk mengatasinya dapat dikeluarkan variabel yang memiliki nilai VIF>10.

K - MEANS

Algoritma K – Means merupakan salah satu algoritma partitional, karena K – Means didasarkan pada penentuan jumlah awal kelompok dengan mendefinisikan nilai centroid awalnya[2]. Metode ini berusaha membagi data kedalam kelompok sehingga data yang berkarakteristik sama dimasukkan kedalam satu kelompok sementara data yang berkarakteristik berbeda dimasukkan dedalam kelompok yang lain.

Tahapan – tahapan melakukan peng*cluster*an menggunakan K – Means yaitu[7] :

1. Menentukan banyaknya k, yaitu banyaknya *cluster*.
2. Menentukan titik pusat (centroid) secara acak ditahap pertama.
3. Menghitung jarak euclidean menggunakan rumus :

$$d_{(x,y)} = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2} \quad (1,3)$$

   Dimana :
   $d_{(x,y)}$= Jarak data 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
4. Menghitung kembali centroid dengan keanggotaan cluster yang terbentuk dengan menghitung nilai rata – rata dari semua data dalam dalam sebuah cluster.
5. Menghitung kembali setiap objek menggunakan centroid baru. Jika anggota cluster tidak mengalami perubahan lagi, maka proses *clustering* dinyatakan selesai.

KEKUATAN PEMBAGIAN CLUSTER

Untuk menentukan *cluster* yang optimal dapat menggunakan metode *silhouette coefficient.* Perhitungan *silhouette coefficient* bertujuan untuk mengetahui kualitas pengelompokan atau pengukuran ketepatan pengelompokan yang dapat dirumuskan[8] :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (1,4)$$

Dimana :
i = Objek yang akan diteliti
$a(i)$ = rata – rata jarak antar anggota dalam cluster
$b(i)$ = nilai minimum dari rata – rata jarak dari objek ke – i dengan objek yang berada di *cluster* lainnnya.
$s(i)$ = nilai silhouette index pada objek ke – i

Setelah mengetahui nilai *s(i)* dilanjut mencari nilai *silhouette coefficient* yaitu dengan mencari rata – rata nilai *s(i)*. Semakin tinggi nilai *s(i)* semakin terstruktur pula kekuatan *cluster*nya.

3. METODOLOGI

Data didapatkan dari laman web kawalcovid19.id dengan variabel yang diperhatikan adalah total orang yang positif covid – 19 (X_1), total orang yang sembuh covid – 19 (X_2), total orang yang meninggal covid – 19 (X_3), total orang yang tersuspect covid – 19 (X_4), total orang yang probable covid – 19 (X_5), total orang yang negatif covid – 19 (X_6).

Prosedur Penelitian

Prosedur penelitian ini dijelaskan sebagai berikut:

1. Menginputkan data dalam bentuk matriks.
2. Menormalisasi data menggunakan persamaan 1,1
3. Melakukan uji multikolinearitas menggunakan persamaan 1,2
4. Membuat *cluster* dengan metode K – Means.
5. Validasi hasil clustering menggunakan *silhouette coefficient* pada persamaan 1,4
6. Melakukan interpretasi dengan melihat nilai centroid untuk menentukan kategori masing – masing *cluster*.
7. Melakukan pewarnaan pada peta dengan melihat anggota cluster.

4. PEMBAHASAN

Profile Data

Terdapat 34 data yang akan dimodel

Tabel 4.1 Sebaran data

No	Provinsi Asal	X_1	X_2	X_3	X_4	X_5	X_6
1	Aceh	8764	7149	358	2381	0	1729
2	Banten	18170	10397	425	12701	2486	10599
3	Bengkulu	3603	2597	117	1009	17	485
4	Bali	17593	16031	519	3927	802	22628
5	Jambi	3227	2417	55	2603	7	0
6	Lampung	6276	4317	274	3688	19	1979
7	Riau	24966	23104	583	80843	76	6615
8	Maluku	5722	4881	79	167	4	2116
9	Papua	13216	7088	147	3548	24	10504
Jumlah		**743.198**	**611.097**	**22.138**	**484.709**	**8.875**	**460.618**

Sumber data: kawalcovid.19.id

Normalisasi Data

Sebelum melakukan clustering, data dinormalisasikan menggunakan rumus *Min – Max Normalization*. Hasil normalisasi dapat dilihat pada Tabel 1.

Tabel 1. Indeks Pendugaan Distribusi

Provinsi Asal	X_1	X_2	X_3	X_4	X_5	X_6
Aceh	0,03743	0,03635	0,05707	0,02217	0	0,01122
Banten	0,08927	0,05621	0,06862	0,12089	1	0,08828
Bengkulu	0,00914	0,00853	0,01552	0,00904	0,00648	0,00315
Bali	0,08610	0,09065	0,08483	0,03695	0,32261	0,14689
Jambi	0,00707	0,00743	0,00483	0,02429	0,00282	0
Lampung	0,02385	0,01904	0,04259	0,03467	0,00764	0,01285
Riau	0,12665	0,13389	0,09586	0,77276	0,03057	0,04294
Maluku	0,02080	0,02004	0,00897	0,00099	0,00161	0,01374
Papua	0,06202	0,03598	0,02069	0,03333	0,00965	0,06819

Uji Multikolinearitas

Setelah mendapatkan data hasil normalisasi, langkah selanjutnya adalah melakukan uji multikolinearitas untuk mengetahui adanya variabel independen yang memiliki kesamaan karakteristik antar variabel independen lainnya. Cara menguji adanya multikolinearitas yaitu dengan cara mencari nilai VIF. Hasil uji multikolinearitas pertama dapat dilihat pada Tabel 2.

Tabel 2. Nilai VIF

Variabel	Nilai VIF
X_1	138,813190
X_2	155,371493
X_3	6, 049523
X_4	4,956888
X_5	1,648041
X_6	21,932782

Tabel 2 menunjukkan bahwa variabel X_1, X_2, dan X_6 terindikasi multikolinearitas. Maka variabel X_2 dikeluarkan terlebih dahulu, dikarenakan variabel X_2 memiliki angka yang cukup tinggi. Kemudian di uji multikolinearitas kembali, didapat bahwa variabel X_1dan X_6 terindikasi multikolinearitas dengan angka tertinggi pada variabel X_1. Setalah mengeluarkan variabel X_1 diperoleh hasil VIF yang dapat dilihat pada Tabel 3.

Tabel 3. Nilai VIF

Variabel	Nilai VIF
X_3	2,067071
X_4	2,769359
X_5	1,561552
X_6	2,537915

Tabel 3 menyimpulkan bahwa data dengan variabel X_3, X_4, X_5, X_6 tidak terindikasi multikolinearitas, sehingga data yang akan

digunakan hanya data hasil normalisasi yang tidak terindikasi multikolinearitas.

Silhouette Coefficient

Langkah selanjutnya yaitu menentukan nilai *silhouette coefficient* untuk mengetahui *cluster* yang optimal. Hasil silhouette coefficient dapat dilihat pada Tabel 4.

Tabel 4. Hasil *silhouette coefficient*

Cluster	Nilai SC
2	0,74
3	0,48
4	0,35
5	0,35
6	0,37

Tabel 4 menyimpulkan bahwa *cluster* yang optimal terletak pada 2 *cluster* dengan nilai SC tertinggi yaitu sebesar 0,74. Nilai centroid pada percobaan 2 *cluster* dapat dilihat pada Tabel 5.

Tabel 5. Nilai Centroid 2 *cluster*

Pusat	X_3	X_4	X_5	X_6
C1	0,040289	0,041664	0,031203	0,0425483
C2	0,421753	0,574913	0,449382	0,2996306

Tabel 5 menyimpulkan bahwa 2 *cluster* membentuk 2 kelompok yang memiliki perbedaan , nilai *centroid* pertama lebih tinggi dibandingkan dengan *centroid* kedua pada setiap variabel sehingga *centroid* pertama disebut daerah rawan dan *centroid* kedua disebut daerah yang aman.

Pewarnaan Peta

Langkah terakhir pada penelitian ini adalah pewarnaan atau zonasi daerah rawan *covid – 19* di Indonesia. Dalam hal ini, akan diketahui provinsi yang tergolong dalam kelompok rawan, dan aman. Provinsi berwarna biru tergolong rawan penyebaran covid – 19 sedangkan Provinsi berwarna orange adalah kota yang rawan dari penyebaran virus covid - 19

Gambar 1. Hasil pemetaan penyebaran covid – 19 di Indonesia

Gambar 1 menyimpulkan bahwa pada bulan Desember, provinsi yang rawan akan covid – 19 terdiri dari 6 Provinsi yaitu Jawa Barat, Jawa Timur, Banten, Jawa Tengah, DKI Jakarta, dan Riau, sedangkan provinsi yang aman akan covid – 19 terdiri dari 28 provinsi lainnya.

5. KESIMPULAN

Pembahasan hasil menunjukkan bahwa faktor meninggal dan probable sebagai faktor yang mempengaruhi pola penyebaran virus covid – 19. Jumlah *cluster* optimal diperoleh pada 2 *cluster* dengan nilai silhouette coefficient sebesar 0,74 dengan 3 data yang tidak tepat. Dari 2 cluster tersebut didapat hasil bahwa provinsi Jawa Timur, Jawa Barat, Jawa Tengah, DKI Jakarta, Banten, dan Riau tergolong rawan covid – 19 dari pola penyebaran virus covid – 19, sedangkan 28 provinsi lainnya tergolong aman dari pola penyebaran virus covid – 19.

6. DAFTAR PUSTAKA

[1] W. H. Organization, "Laboratory testing for coronavirus disease 2019 Covid - 19 in a suspected human cases," *2 Maret*, 2020.

[2] A. N. Fathia, R. Rahmawati, and Tarno, "Analisis klaster kecamatan di kabupaten semarang berdasarkan potensi desa menggunakan metode ward dan single linkage," *Gaussian*, vol. 5, no. 4, pp. 801–810, 2016.

[3] U. Jannah, "Perbandingan jarak Euclid dengan jarak mahalanobis pada analisis cluster hirarki," 2010.

[4] S. Yulianto and K. H. Hidayatullah, "Analisis Klaster Untuk Pengelompokan Kabupaten/kota di Provinsi Jawa Tengah Berdasarkan Indikator Kesejahteraan Rakyat," *Statistika*, vol. 2, no. 1, pp. 56–63, 2014.

[5] Junaedi, Budianto, and Melani, "Data Transformation pada datamining," in *Prosiding Konferensi Nasional Inovasi dalam desain dan Teknologi - IDeaTech*, 2011, pp. 93–99.

[6] T. S. Madhulatha, "An Overview On Clustering Methods," *IOSR J. Eng.*, vol. 2, no. 4, pp. 719–725, 2012.

[7] N. Dwitri, J. A. Tampubolon, S. Prayoga, and P. P. P. A. N. W. F. I. R. H. Zer, "Penerapan Algoritma K-Means Dalam Menentukan Tingkat Penyebaran Pandemi Covid-19 di Indonesia," *J. Teknol. Inf.*, vol. 4, no. 1, pp. 128–132, 2020.

[8] S. A. diana Budiman, D. Safitri, and D. Ispriyanti, "Perbandingan Metode K-Means dan Metode DBSCAN pada Pengelompokan Rumah Kost Mahasiswa di Kelurahan Tembalang Semarang," *J. gaussian*, vol. 5, no. 4, pp. 757–762, 2016.