

IMPLEMENTASI ALGORITMA DBSCAN(DENSITY BASED SPATIAL CLUSTERING ALGORITHM WITH NOISE) PADA COVID19 DI INDONESIA

Irma Fatmawati ¹⁾, Isnandar Slamet ²⁾, dan Sri Subanti ³⁾

^{1,2,3} Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Sebelas Maret

¹ Email korespondensi: fatmawatiirma@student.uns.ac.id

² email : isnandarslamet@staff.uns.ac.id

³ email : sri_subanti@yahoo.co.id

ABSTRAK

Corona virus merupakan keluarga besar virus yang menyebabkan penyakit ringan hingga berat seperti *common cold* atau pilek dan penyakit serius seperti MERS (*Middle East Respiratory Syndrome*) dan SARS (*Severe Accute Respiratory Syndrome*). Penyakit akibat corona virus dikenal dengan COVID-19. Penularan yang mudah yaitu melalui cairan tubuh (*droplet*) orang terinfeksi menyebabkan penyebaran virus ini terbilang cepat. Di Indonesia kasus COVID-19 juga menyebar dengan cepat di berbagai provinsi, tentunya hal tersebut menyebabkan keresahan masyarakat, tingkat kematian yang tinggi dan berbagai dampak lainnya. Dalam penelitian ini akan dilakukan pengelompokan kasus COVID-19 di Indonesia dengan menggunakan algoritma DBSCAN (*Density-Based Spatial Clustering Algorithm with Noise*) yaitu salah satu algoritma pengelompokan yang didasarkan pada kepadatan data dimana dibutuhkan dua inisial parameter yaitu *epsilon* dan *minPts* yang tepat untuk membuat kelompok yang terbentuk valid. Kemudian alat yang digunakan untuk validasi kelompok yaitu Indeks Validitas Silhouette. Indeks Validitas Silhouette merupakan salah satu teknik yang digunakan untuk menentukan seberapa baik suatu obyek di dalam kelompok. Data yang digunakan dalam penelitian berupa data COVID-19 di Indonesia dengan jumlah 5.378 data terhitung dari 1 Maret sampai 28 Agustus 2020. Data tersebut diambil dari kaggle dengan tiga variabel yaitu total kasus, total kesembuhan, dan total kematian. Berdasarkan hasil analisis diperoleh dua kelompok dan satu noise dalam pengelompokan menggunakan algoritma DBSCAN. Kemudian dilakukan validasi kelompok dengan indeks validitas Silhouette dan didapatkan sebesar 0.92%. Nilai tersebut berarti obyek terletak di dalam kelompok yang tepat.

Kata kunci: DBSCAN, COVID-19, pengelompokan

1. Pendahuluan

Virus corona merupakan virus RNA (*Ribonukleat Acid*) strain tunggal positif berkapsul dan tidak bersegmen. Memiliki protein S yang berperan dalam penempelan dan masuknya virus ke dalam sel *host*. Virus ini dapat masuk ke tubuh dan menginfeksi manusia melalui cairan tubuh (*droplet*) yang dikeluarkan oleh orang-orang yang sudah terinfeksi. Orang-orang yang terinfeksi COVID-19 memiliki kategori gejala ringan, sedang, maupun berat. Gejala klinis utama ditandai demam dengan suhu

di atas 38°C, batuk, kesulitan bernapas, diare, dan gejala saluran napas lain (Kementerian Kesehatan Republik Indonesia, 2020).

Awalnya penyakit ini ditemukan di Wuhan, Cina pada akhir bulan Desember 2019 dengan laporan sejumlah kasus pneumonia yang tidak biasa. Banyak warga Cina yang tertular penyakit ini terhitung pada 9 Februari 2020 terdapat korban baru virus corona sebanyak 811 orang dan orang-orang yang tertular sebanyak 37.198. Kasus COVID-19 ini menyebar tidak hanya di Cina saja, namun juga ke berbagai

negara. Sehingga pada 12 Maret 2020 WHO (*World Health Organisation*) mengumumkan bahwa virus corona sebagai pandemi. Pandemi merupakan penyebaran penyakit baru ke seluruh dunia. Berdasarkan data John Hopkins per 11 Maret 2020 virus corona telah menginfeksi 121.564 orang di 118 negara.

Di Indonesia sendiri kasus COVID-19 pertama kali muncul pada 2 Maret 2020. Penyebaran yang mudah dan cepat mengakibatkan COVID-19 menyebar dengan cepat ke seluruh provinsi di Indonesia. Hanya perlu waktu 24 hari yaitu pada 26 Maret 2020 pemerintah Indonesia mengumumkan terdapat total kasus 893 terkonfirmasi, 78 kematian, dan 35 pasien sembuh dari total 27 provinsi. Seperti yang kita ketahui sendiri bahwasanya kasus COVID-19 ini terus bertambah setiap harinya, terhitung pada 23 September 2020 terdapat 252.920 kasus positif dan 9.837 kematian di Indonesia. Hingga saat ini, Indonesia sudah memasuki bulan ke tiga kondisi *new normal*, namun kasus COVID-19 belum mampu di tekan dengan optimal.

Selain ketakutan dan keresahan masyarakat yang tentunya dapat menyebabkan kesehatan mental bagi masyarakat, tingkat mortalitas yang tinggi, COVID-19 ini tentunya memberikan dampak yang besar terhadap berbagai sektor di Indonesia baik pendidikan, ekonomi, pariwisata, dan lainnya. Salah satu dampak dari virus corona yang paling berpengaruh dan berdampak buruk yaitu terhadap perekonomian Indonesia. Berdasarkan data yang diperoleh dari BPS (Badan Pusat Statistik) ekonomi Indonesia triwulan II 2020 terhadap triwulan II 2019 mengalami kontraksi pertumbuhan 5,32 % (*year on year*). Dari sisi produksi, lapangan usaha transportasi dan pergudangan mengalami kontraksi pertumbuhan tertinggi sebesar 30,84%. Sementara dari sisi pengeluaran, komponen ekspor barang dan jasa serta impor barang dan jasa mengalami kontraksi pertumbuhan masing-masing sebesar 11,66% dan 16,96%.

Pengelompokan merupakan teknik *unsupervised learning*, yaitu pembelajaran tidak dibimbing atau tidak disupervisi, pembelajaran tanpa data latih terlebih dahulu. Teknik pengelompokan adalah teknik mengelompokkan data berdasarkan kemiripannya menjadi beberapa kelompok. Setiap kelompok memiliki karakteristik khusus yang membedakan dengan kelompok yang lain (Aggarwala, 2015)

Metode yang akan digunakan yaitu DBSCAN (*Density-Based Spatial Pengelompokan*). *Density – Based Spatial Pengelompokan Algorithm With Noise* (DBSCAN) adalah algoritma pengelompokan yang didasarkan pada kepadatan (*density*) data. Konsep kepadatan yang dimaksud dalam DBSCAN adalah banyaknya data (*minPts*) yang berada dalam radius *Eps* (ϵ) dari setiap data.

Rembulan dkk., (2020) pernah melakukan penelitian mengenai analisis kelompok COVID-19 di Indonesia. Penelitian yang dilakukan menggunakan metode *K-Means software* SPSS. Dengan menggunakan jumlah kelompok $k=3$, maka dari tiga kelompok tersebut dilakukan analisis karakteristis, tingkat resiko, dan pergerakan setiap kelompok provinsi yang beresiko tinggi. Dalam penelitian tersebut, baru digunakan metode *K-Means* untuk menganalisis penyebaran kasus COVID-19 di Indonesia. Sehingga penelitian ini akan diimplementasikan algoritma DBSCAN untuk mengelompokkan kasus COVID-19 di Provinsi Indonesia.

2. Metode Penelitian

Penelitian ini merupakan studi kasus pada provinsi yang memiliki resiko tinggi kasus COVID-19. Data yang digunakan data sekunder yang diambil dari kaggle.com dengan jumlah data 5.378. Adapun variabel yang digunakan yaitu berupa jumlah kasus kematian, jumlah kasus sembuh, dan jumlah kasus aktif. Akan dilakukan pengaplikasian menggunakan algoritma DBSCAN. Berikut merupakan langkah- langkah dari penelitian yang akan dilakukan :

1.) *Preprocessing* data

Sebelum melakukan proses analisis, data mentah yang akan digunakan diolah terlebih dahulu agar siap untuk diproses. *Preprocessing* data adalah langkah awal dari pengecekan kualitas data. Tujuannya adalah untuk membuat data bersih, lengkap, dan siap diproses. Adapun tahapan untuk *preprocessing* data dalam penelitian ini adalah sebagai berikut :

- Mengunduh data pada alamat *website* : <https://www.kaggle.com/search> . Data yang diambil merupakan data harian kasus COVID-19 di Indonesia per provinsi dengan variabel jumlah kasus kematian, jumlah kasus sembuh, dan jumlah kasus aktif.
- Memastikan tidak terdapat *missing value* pada data penelitian.

- c. Melakukan standarisasi data agar data memiliki range yang sama, sehingga tidak menimbulkan variabel tertentu mendominasi variabel lainnya.
- 2.) Menggolongkan data dengan metode DBSCAN. Berikut langkah-langkahnya :
 - a. Inisialisasi parameter *minpts* dan *eps* berdasarkan point
 - b. Tentukan point awal atau p secara acak
 - c. Menghitung jarak semua point ke point p
 - d. Jika p adalah *core point* maka kelompok terbentuk
 - e. Jika p adalah border point tidak ada yang merupakan hubungan density reachable terhadap p dan DBSCAN akan mengunjungi titik selanjutnya dari data base
 - f. Lanjutkan proses sampai semua titik telah selesai
- 3.) Validasi Kluster
Tahap ini dilakukan untuk menguji tingkat validitas kelompok yang terbentuk, dimana digunakan nilai Indeks validitas *Silhouette*.
- 4.) Melakukan pemetaan terhadap hasil pengelompokan

3. Hasil dan Pembahasan

Perprocessing data

Tahap ini dilakukan seleksi data, pembersihan data, dan standarisasi data. Seleksi data dilakukan dengan menyeleksi data yang akan digunakan berdasarkan daerah yang akan diteliti dan mengambil tiga atribut dari dataset yang didapatkan dari *kaggle* yaitu total kasus, total kematian, dan total kesembuhan. Selanjutnya, pembersihan data dilakukan apabila ditemukan data yang kosong (*missing value*). Terdapat berbagai macam pengepresian *missing value* di dalam dataset, pada umumnya ditulis NA (*not available*) atau di representasikan dengan 0, 99, 9999, dan lainnya. Dalam data yang akan dilakukan penelitian tidak ditemukan adanya *missing value*, sehingga dilanjutkan ke langkah berikutnya yaitu kita menganalisis data dengan statistika deskriptif untuk mengetahui ringkasan secara umum data penelitian.

Statistika Deskriptif

Tabel 1. Analisis statistika deskriptif

	T.Kasus	T.Kematian	T.Kesembuhan
Min.	1	0	0
1 st Qu.	89	2	20
Median	267	9	131
Mean	1540	77.06	838.3
3 rd Qu	1223	49	513.8
Max.	37082	2302	29169

Berdasarkan tabel 1 tersebut dapat dilihat bahwa setiap variabel memiliki rentang nilai yang berbeda. Sehingga perlu dilakukan standarisasi data sebelum dilakukan analisis dengan tujuan agar data penelitian memiliki rentang yang sama sehingga tidak ada satu variabel yang mendominasi variabel lain serta agar *machine learning* lebih mudah dalam melakukan proses analisis. Berikut adalah analisis statistika deskriptif setelah dilakukan normalisasi data yang disajikan dalam tabel 2.

Tabel 2.

Ringkasan data yang sudah di standarisasi

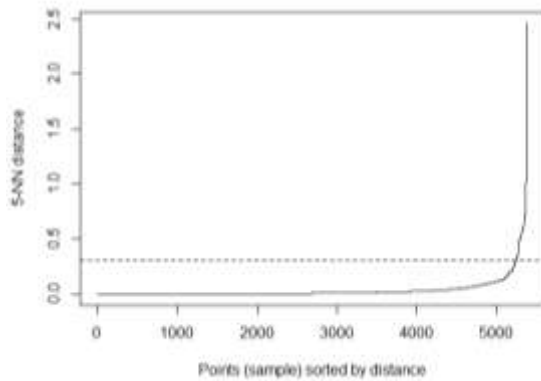
	Kasus	Kematian	Sembuh
Min.	-0.40724	-0.3530	-0.3377
Kuartil I	-0.38395	-0.3439	-0.3297
Median	-0.33685	-0.3118	-0.2850
Mean	0.00000	0.0000	0.0000
Kuartil II	-0.08394	-0.1285	-0.1308
Max.	9.40508	10.1935	11.4148

Berdasarkan analisis statistika deskriptif pada tabel 2 menunjukkan bahwa data telah memiliki skala yang sama, dimana setelah dilakukan standarisasi data nilai rata-rata dari ketiga variabel menjadi nol. Selanjutnya data siap untuk di analisis dengan menggunakan metode *Density Based Spatial Clustering Algorithm With Noise* (DBSCAN).

Penentuan *epsilon* dan *MinPts*

Langkah awal yang harus dilakukan untuk melakukan pengelompokan menggunakan algoritma DBSCAN adalah dengan menentukan nilai *eps* dan *minpts*. *eps* adalah nilai jarak antaritem yang menjadi dasar pembentukan *neighborhood* dari suatu titik, sedangkan *minpts* adalah minimal banyaknya item dalam cluster. Penentuan kedua parameter tersebut sangat berpengaruh dengan jumlah cluster yang dihasilkan sehingga diperlukan alat bantu untuk menemukan nilai *eps* yang tepat yaitu Plot

kNNdist. Knee pada plot tersebut menunjukkan nilai ϵ yang optimal. Berikut merupakan plot k-NN distance dengan $k=5$ untuk data set yang digunakan dalam penelitian ini.



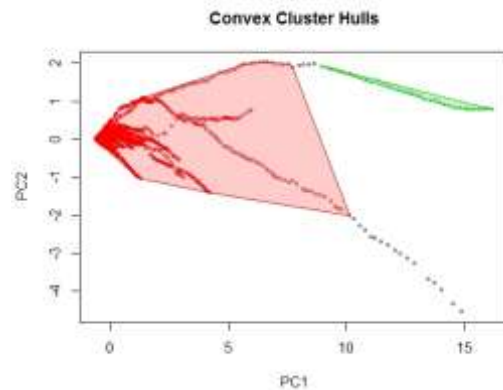
Gambar. 2 Plot kNN distance

Berdasarkan gambar 2 *plot kNN distance knee* terletak antara 0.2 hingga 0.5, sehingga perlu dilakukan kombinasi masukan ϵ antara 0.2 sampai dengan 0.5 dan $\minPts$ antara 1 sampai dengan 5. Hasil dari percobaan masukan ϵ dan $\minPts$ ditunjukkan pada tabel 3.

Tabel 3.
Hasil Percobaan

ϵ	$\minPts$	Jumlah kelompok	average silhouette
0.5	5	2	0.92
	4	2	0.90
	3	2	0.91
	2	3	0.91
	1	3	0.91
0.4	5	5	0.89
	4	4	0.90
	3	2	0.91
	2	2	0.91
	1	4	0.91
0.3	5	2	0.86
	4	3	0.88
	3	4	0.88
	2	7	0.88
	1	12	0.88
0.2	5	8	0.73
	4	9	0.73
	3	10	0.69
	2	18	0.69
	1	65	0.67

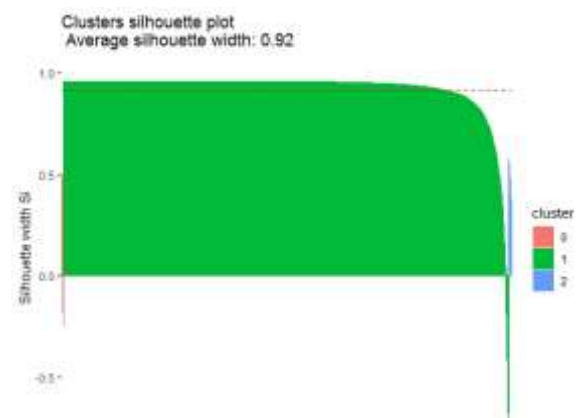
Dari tabel di atas dapat dilihat nilai average silhouette yang terbesar yaitu 0.92 dengan nilai ϵ sebesar 0.5 dan $\minPts$ sebesar 5, maka dari itu kita menggunakan nilai tersebut untuk inisialisasi kedua parameter. Didapatkan cluster seperti pada gambar di bawah ini



Gambar 3. convex cluster hull

Terlihat bahwa dari gambar terbentuk sebanyak dua kelompok dimana warna merah adalah kelompok satu yang terdiri dari sebanyak 5.321 obyek dan warna hijau merupakan kelompok dua dengan 37 obyek. Terdapat 20 obyek yang berwarna hitam, dimana obyek – obyek tersebut merupakan *noise* (data yang jumlah data di dalam radius kurang dari $\minPts$) .

Untuk mengetahui seberapa tepat obyek dimasukkan dalam kelompok dapat di lihat dari average silhouette.



Gambar 4. Average silhouette

Berdasarkan gambar di atas dapat dilihat bahwa didapatkan average silhouette sebesar 0.92 atau dengan kata lain mendekati nilai 1 yang berarti bahwa obyek terletak pada kelompok

yang tepat. Berikut merupakan ringkasan average silhouette untuk setiap kelompoknya.

Tabel 4.
Ringkasan hasil *Average Silhouette*

Kelompok	Ukuran	average silhouette
0	20	0.32
1	5321	0.92
2	37	0.48

Dari tabel 4 diperoleh 2 kelompok dengan jumlah anggota kelompok pertama sebanyak 5321 dengan *average silhouette* 0.92, kelompok 2 dengan anggota sebanyak 37 dengan *average silhouette* sebesar 0.48, dan *noise* dengan jumlah anggota sebanyak 20 dan *average silhouette* sebesar 0.32.

Berikut merupakan kelompok yang terbentuk, diambil dari 34 kelompok terakhir yang terbentuk. Dimana provinsi yang masuk di dalam kelompok 1 adalah Jawa Timur, provinsi yang menjadi *noise* adalah provinsi DKI Jakarta, dan provinsi yang lain masuk kedalam kelompok dua.

Tabel 5.
Hasil pengelompokan per provinsi

Lokasi	Kluster
Aceh	1
Bali	1
Banten	1
Bengkulu	1
DKI Jakarta	0
Daerah Istimewa Yogyakarta	1
Gorontalo	1
Jambi	1
Jawa Barat	2
Jawa Tengah	1
Jawa Timur	1
Kalimantan Barat	1
Kalimantan Selatan	1
Kalimantan Tengah	1
Kalimantan Timur	1
Kalimantan Utara	1
Kepulauan Bangka Belitung	1
Kepulauan Riau	1
Lampung	1
Maluku	1

Lokasi	Kluster
Maluku Utara	1
Nusa Tenggara Barat	1
Nusa Tenggara Timur	1
Papua	1
Papua Barat	1
Riau	1
Sulawesi Barat	1
Sulawesi Selatan	1
Sulawesi Tengah	1
Sulawesi Tenggara	1
Sulawesi Utara	1
Sumatera Barat	1
Sumatera Selatan	1
Sumatera Utara	1



Gambar 5. Visualisasi hasil pengelompokan

Hasil pengelompokan tersebut kemudian dilakukan visualisasi dengan bantuan software geoda. Warna orange pekat menunjukkan noise yaitu provinsi DKI Jakarta, warna orange sedang menunjukkan kelompok dua yaitu provinsi Jawa Timur, dan warna orange muda menunjukkan kelompok 1 dengan provinsi Bangka Belitung, Gorontalo, Riau, Kepulauan Riau, Sulawesi Selatan, Kalimantan Barat, Aceh, Bengkulu, Lampung, Jawa Tengah, Nusa Tenggara Barat, Sumatera Utara, Jambi, Banten, Yogyakarta, Kalimantan Selatan, Sulawesi Utara, Sulawesi Barat, Maluku Utara, Kalimantan Tenggara, Nusa Tenggara Timur, Papua Barat, Jawa Barat. Karakteristik kelompok yang terbentuk yaitu sebagai berikut.

Tabel 6.
Karakteristik pengelompokan

Kelompok	Karakteristik
1	Jumlah kasus tergolong rendah hingga sedang Jumlah kematian tergolong rendah hingga sedang Jumlah kesembuhan tergolong

Kelompok	Karakteristik
	rendah hingga sedang
2	Jumlah kasus tergolong tinggi Jumlah kematian tergolong tinggi Jumlah kesembuhan tergolong tinggi

4. Kesimpulan

Berdasarkan hasil analisis pengelompokan menggunakan DBSCAN dalam studi kasus COVID-19 di Indonesia didapatkan didapatkan dua kelompok berdasarkan total kasus, total kesembuhan, dan total kematian dengan indeks validitas Silhouette sebesar 0.92% . Hal tersebut dapat di artikan bahwa obyek terletak dalam kelompok yang tepat. Untuk penelitian selanjutnya diharapkan dapat dilakukan pengelompokan dengan algoritma lainya dengan tujuan agar dapat dibandingkan hasilnya dengan algoritma DBSCAN (*Density Based Spatial Clustering Algorithm with Noise*).

5. Daftar Pustaka

- Aggarwala, C. C. 2015. *Data Mining*. Yorktown Height, New York: Springer International Publisher Switzerland
- Budiman, S. A., Safitri, D., & Ispriyanti, D. 2016. Perbandingan Metode K-Means dan Metode DBSCAN Pada Pengelompokan Rumah Kost Mahasiswa Di Kelurahan Tembalang Semarang. *Jurnal Gauss*. 5 : 757 – 762
- Johnson, R. A., & Wichern, D. W. 2007. *Applied Multivariate Statistical Analysis 6th Edition*. United States of America: Pearson Education In
- Kementerian Kesehatan Republik Indonesia. 2020. *Novel Coronavirus*. Tersedia pada : <https://www.kemkes.go.id/> Diakses pada 18 Juli 2020
- Rembulan, G. D., Wijaya, T., Palullungan, D., Alfina, K. N., & Qurthubi, M. 2020. Kebijakan Pemerintah Mengenai Corona Virus Disease (COVID-19) di Setiap Provinsi di Indonesia Berdasarkan Analisis Kluster. *Journal of Industrial Engineering and Management Systems*. 13 :74-86.
- Wang, Z., Qiang, W., Ke, H.A . 2020.*Handbook of 2019-nCoV*

Pneumonia Control and Prevention.Hubei Science and Technology Press. China.

Wolting, C., McGlade, C. J., & Tritchler, D. 2006. Kelompok Analysis of Protein Array Result Via Simillarity of Gene Ontlogy Abotation. *BMC Bioinformatics* , 1-13.doi: 10.1186/1471-2105-7-338