Perbandingan Algoritma *K-Means* dan DBSCAN untuk Pengelompokan Data Penyebaran Covid-19 Seluruh Kecamatan di Provinsi Jawa Barat

Bayu Biantara
Universitas Buana Perjuangan Karawang
Karawang, Indonesia
if18.bayubiantara@mhs.ubpkarawang.ac.id

Tatang Rohana
Universitas Buana Perjuangan Karawang
Karawang, Indonesia
tatang.rohana@ubpkarawang.ac.id

Ayu Ratna Juwita
Universitas Buana Perjuangan Karawang
Karawang, Indonesia
ayurj@ubpkarawang.ac.id

***Abstract*— Virus *Covid-19* ditemukan pertama kali di Wuhan, Tiongkok. Virus *Covid-19* menyebar secara cepat, hampir seluruh negara yang ada di dunia. WHO memutuskan sebagai *Public Health Emergency of International Concern* (KKMMD/PHEIC) pada tanggal 30 Januari 2020**[1]**. *Clustering* merupakan salah satu substansi *Data Mining* untuk pengelompokkan suatu data. Terdapat beberapa Algoritma dalam *clustering* diantaranya Algoritma *K-Means* dan Algoritma DBSCAN. Tujuan dari penelitian untuk melakukan perbandingan Algoritma yang terbaik antara Algoritma *K-Means* dan DBSCAN dalam pengelompokan data penyebaran *Covid-19* seluruh kecamatan di provinsi Jawa Barat. Dari hasil penelitian validitas *cluster* antara Algoritma *K-Means* dan Algoritma DBSCAN menghasilkan Algoritma *K-Means* lebih optimal dari Algoritma DBSCAN karena memiliki nilai DBI terbaik dibandingkan Algoritma DBSCAN. Nilai DBI Algoritma *K-Means* diperoleh dengan nilai 0,4328 pada k=5, sedangkan Algoritma DBSCAN diperoleh nilai DBI pada nilai Eps 0,09 dan MinPts 3 yaitu sebesar 0,6706.**

Kata kunci — clustering, algoritma K-Means, algoritma dbscan

I. PENDAHULUAN

Pada bulan Maret tahun 2020 penyakit Virus Corona atau *Covid-19* melanda negara Indonesia. Virus ini menular sangat cepat, hampir seluruh negara yang ada didunia tak terkecuali Indonesia . Virus ini ditemukan pertama kali di Wuhan, Tiongkok. Virus corona atau *Covid-19* merupakan virus yang menyerang bagian pernafasan yang diakibatkan oleh virus *SARS-CoV-2* yang merupakan keluarga besar dari Coronavirus [2]. Virus ini menular dengan cara kontak langsung antar sesama manusia yang dikeluarkan penderita berupa droplet (percikan kecil) pada saat berbicara, batuk, atau bersin. Keberadaan *Covid-19* ini menyebabkan kegiatan sosial maupun ekonomi pada masyarakat mengalami kelumpuhan. Beberapa wilayah yang memiliki tingkat penyebaran paling tinggi ditetapkan sebagai daerah zona merah. *Covid-19* ini meluas ke 34 provinsi, salah satu diantaranya yaitu provinsi Jawa Barat [3]. Provinsi jawa barat merupakan daerah yang memiliki penduduk terbesar pada tahun 2018 yang berjumlah 48.683.700 [4].

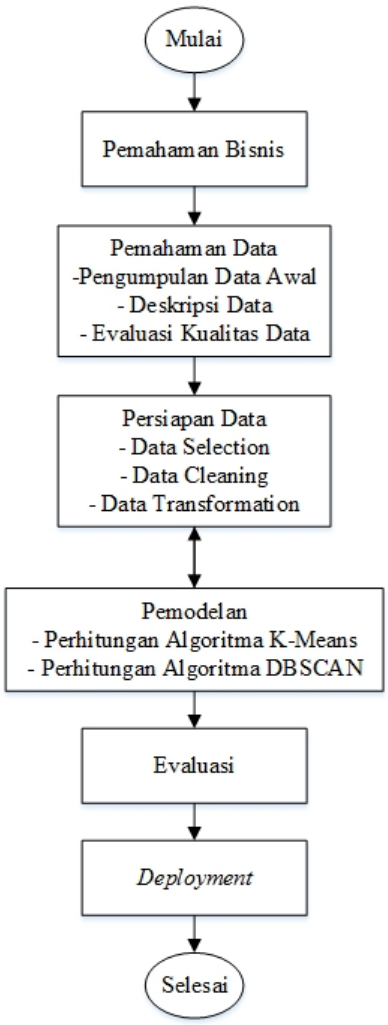
Seiring dengan adanya penyebaran Covid-19, pada penelitian ini melakukan pengelompokan penyebaran Covid-19 seluruh Kecamatan di Jawa Barat menggunakan Data Mining. Data Mining adalah serangkaian Knowledge Discovery Database (KDD) dimana sebuah cara pengolahan data dalam jumlah besar untuk mendapatkan pola yang tersimpan [6]. Dimana hasil dari pola yang ditemukan dapat diaplikasikan untuk pengambilan suatu keputusan. Metode yang dilakukan dalam Data Mining untuk pengelompokan sejumlah data disebut dengan Clustering. Clustering merupakan proses pengelompokkan kedalam kelas dengan objek didalamnya mempunyai kesamaan [7].

Adapun penelitian dalam pengelompokan data menggunakan metode *clustering* diantaranya oleh Rimelda Adha dkk (2021) tentang perbandingan antara Algoritma DBSCAN dan *K-Means* dalam pengelompokan kasus *Covid-19* di dunia. Hasil penelitian yang diperoleh bahwa Algoritma *K-Means* lebih optimal dari Algoritma DBSCAN dalam pengelompokan penyebaran *Covid-19*. Dimana Algoritma *K-Means* mempunyai nilai *Silhouette Index* (SI) terbaik terdapat pada percobaan k = 8 dengan nilai sebesar 0,6902 [8]. Penelitian yang kedua dilakukan oleh Zulia Imami Alfianti (2021) mengenai penerapan Algoritma *K-Means* dalam pengelompokan wilayah penyebaran *Covid-19* di Karawang. Hasil penelitian yang dilakukan terdapat 50% wilayah termasuk dalam penyebaran rendah , 33,3% persen wilayah termasuk dalam penyebaran sedang, dan 16,7% wilayah termasuk dalam penyebaran tinggi [9]. Penelitian lainnya oleh Nana Nurhaliza dan Mustakim (2021) mengenai penerapan Algoritma DBSCAN dalam pengelompokan data kasus *Covid-19* di dunia. Hasil penelitian yang dilakukan bersumber pada *Silhouette Index* (SI) dengan nilai Eps 0,2 dan MinPts 0,3 ditemukan cluster tertinggi pada percobaan ke-21 sebesar 0,3624 [10].

Berdasarkan riwayat penelitian yang telah dilakukan sebelumnya, penelitian ini bertujuan melakukan perbandingan Algoritma *K-Means* dan DBSCAN dalam pengelompokan data *Covid-19* seluruh Kecamatan di Jawa Barat. Metode validasi yang digunakan yaitu Davies-Bouldin Index (DBI) untuk memperhitungkan nilai *cluster* optimal sebagai penentu ketepatan Algoritma yang digunakan.

II. METODOLOGI PENELITIAN

Adapun metode penelitian yang digunakan dengan pendekatan model *Data Mining* yakni *Cross-Industry Standard Process Data Mining* (CRISP-DM) yang merupakan suatu model proses penyelesaian masalah pada *Data Mining* dengan beberapa tahapan analisis[11]. Berikut prosedur penelitian ditunjukkan gambar 1 berikut.



Gambar 1 Prosedur Penelitian

A. Pemahaman Bisnis

Pada tahap ini menentukan tujuan dari penelitian yaitu menerapkan Algoritma *K-Means* dan DBSCAN dalam pengelompokan data penyebaran Covid-19 seluruh Kecamatan di Jawa Barat. Tujuan lainnya untuk mengetahui Algoritma terbaik antara *K-Means* dan DBSCAN dalam pengelompokan data Covid-19 seluruh Kecamatan di Jawa Barat.

B. Pemahaman Data

Pada tahap ini melakukan pengumpulan data penyebaran *Covid-19* bersumber dari situs resmi Pusat Informasi dan Koordinasi *COVID-19* Provinsi Jawa Barat dari tanggal 5 Agustus 2020 sampai dengan 5 November 2021 berdasarkan laporan harian Dinas Kesehatan Kota/Kabupaten di Jawa Barat.

C. Persiapan Data

Pada tahap persiapan data ini perlu dilakukan secara teliti. Dimana akan dilakukan proses pengolahan data awal dengan pemilihan variabel yang dianalisis dan melakukan perubahan pada beberapa variabel jika diperlukan. Sehingga menghasilkan *dataset final* yang siap untuk dilakukan pemodelan.

D. Pemodelan

Pada tahap ini menerapkan metode *Data Mining* ke *dataset* yang sudah disiapkan berguna untuk mencapai tujuan proyek yakni dengan menggunakan Algoritma *K-Means* dan Algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*).

1) Algoritma K-Means

Algoritma *K-Means* termasuk sebagai pengelompokan non hirarki dimana mampu membagi objek kedalam satu kelompok atau beberapa kelompok lainnya [9]. Algoritma *K-Means* adalah salah satu Algoritma yang dapat mengelompokkan suata data menjadi berbagai kelompok, dimana setiap kelompok mempunyai karakteristik yang berbeda dengan kelompok lain [9].

Berikut adalah tahapan-tahapan dalam Algoritma *K-Means*[12]:

1. Tentukan banyaknya *cluster* k.
2. Tentukan secara acak titik pusat (*centroid*) dari masing-masing *cluster*.
3. Pada setiap data di hitung jarak terdekat terhadap *centroid* menggunakan rumus *Euclidean Distance*.

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (1)$$

Keterangan :
D_e = *Euclidean Distance*
i = Banyak data
(x , y) = Titik data
(s , t) = Titik pusat

4. Kemudian bentuk *cluster* baru berdasarkan jarak minimum terhadap *cluster*.
5. Tentukan titik pusat (*centroid*) baru dengan rumus berikut :

$$\bar{v} = \frac{1}{N_i} \sum_{k=0}^{N_i} x_{kj} \quad (2)$$

Keterangan :
$\bar{v}_{ij}$ = Rata-rata titik pusat pada *cluster* ke-i untuk variabel ke-j
N_i = Banyaknya suatu anggota *cluster* ke-i
i , k = indikator dari *cluster*
j = indikator dari variabel
x_{kj} = Nilai data ke-k pada *cluster* tersebut untuk variabel ke-j

6. Ulangi tahap 3, 4 dan 5 sampai anggota *cluster* tidak beralih ke *cluster* lainnya maka iterasi tersebut berhenti[5] .

2) Algoritma Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) merupakan Algoritma dalam *clustering* untuk menentukan sampel inti berdasarkan kepadatan yang ditemukan oleh Ester Martin [10]. Dalam menentukan *cluster* terdapat dua parameter utama yaitu jumlah sampel minimal dan ε [8]. Setiap objek dalam radius area (*cluster*) harus memiliki sekurang-kurangnya berjumlah minimum data. Objek apa pun jika tidak sesuai dengan *cluster* disebut sebagai *Noise.*

Berikut tahapan-tahapan dari Algoritma DBSCAN sebagai berikut [8]:

1. Tentukan nilai Epsilon (Eps) dan MinPoints (MinPts).
2. Tentukan nilai p atau titik awal secara *random* atau acak.
3. Menghitung nilai Eps atau hitung jarak masing-masing titik yang memiliki kepadatan terhadap titik p dengan rumus *Euclidean Distance* berikut :

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (3)$$

Keterangan :
D_e = *Euclidean Distance*
i = Banyak data
(x , y) = Titik data
(s , t) = Titik pusat

4. Sebuah *cluster* terbentuk jika titik sudah mencukupi epsilon lebih dari minimum poin, maka titik tersebut sebagai titik pusat.
5. Ulangi tahap 3 dan 4 sampai semua titik dilakukan perhitungan. Lanjutkan ke titik lainnya ketika tidak terdapat titik yang memiliki kepadatan terhadap p atau titik awal.

E. Evaluasi

Pada tahap ini melakukan uji kualitas *cluster* yang terbentuk dari proses *clustering* yang telah dilakukan yakni menggunakan metode *Davies-Bouldin Index* (DBI).

Adapun langkah-langkah metode *Davies-Bouldin Index* (DBI) sebagai berikut [13] :

1. *Sum of Square Within-cluster* (SSW)
   Tahap pertama yang dilakukan yaitu mencari matrik kohesi dari *cluster* ke-i yang dimana menggunakan rumus dari SSW berikut ini.

$$SSW_i = \frac{1}{m_i}\sum_{j=i}^{m_i} d(x_j, c_i) \quad (4)$$

2. *Sum of Square Between-cluster* (SSB)
Untuk tahap kedua ini menentukan matrik separasi antar *cluster* yang ada dengan menggunakan rumus dari SSB sebagai berikut.

$$SSB_{i,j} = d(c_i, c_j) \quad (5)$$

3. *Ratio* (Rasio)
Untuk tahap ketiga ini mencari nilai rasio perbandingan dari *cluster* ke-i dan *cluster* ke-j. Dengan rumus sebagai berikut.

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (6)$$

4. *Davies-Bouldin Index*
Berikut rumus untuk mencari nilai DBI:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}(R_{i,j}) \quad (7)$$

Dari nilai DBI yang diperoleh, k adalah sebuah *cluster*. Untuk menentukan sebuah k (*cluster*) yang baik dapat dilihat berdasarkan nilai DBI yang terkecil dari pengelompokan menggunakan Algoritma *clustering* [13].

F. *Deployment*

Pada tahap *Deployment* merupakan tahap terakhir pada penelitian ini. Tahap ini melakukan penggambaran dari hasil penelitian yang dilakukan. Pengetahuan yang diperoleh dari *dataset* setelah dilakukan pengolahan menggunakan metode *Data Mining.*

III. HASIL DAN PEMBAHASAN

A. Data

Data pada penelitian ini berupa data penyebaran *Covid-19* yang ada di seluruh kecamatan pada provinsi Jawa Barat. Data yang diperoleh berjumlah 627 data kecamatan dari 27 kabupaten. Atribut yang diperoleh yaitu Kecamatan, Terkonfirmasi, Dalam Perawatan, Sembuh, dan Meninggal ditunjukkan tabel 1 berikut,

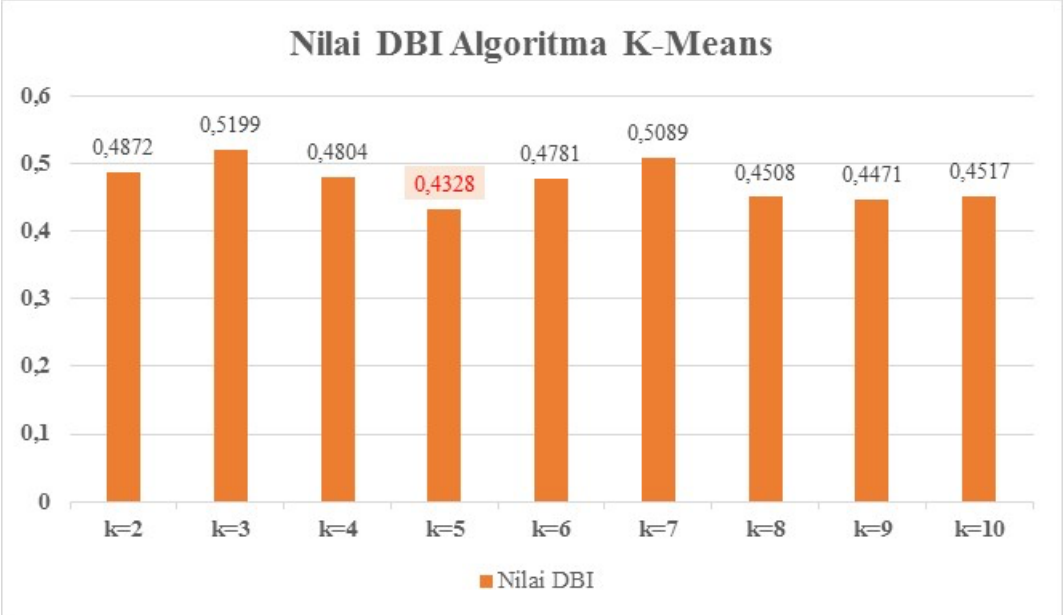
Tabel 1 *Dataset*

No	Kecamatan	Terkonfirmasi	Dalam Perawatan	Sembuh	Meninggal
1	BABAKAN MADANG	680	0	679	1
2	BOJONG GEDE	3.842	2	3.829	11
3	CARINGIN	458	0	451	7
4	CARIU	162	0	146	16
5	CIAMPEA	634	1	627	6
6	CIAWI	563	0	563	0
7	CIBINONG	5.627	2	5.614	11
8	CIBUNGBULANG	493	0	493	0
9	CIGOMBONG	770	0	766	4
10	CIGUDEG	244	2	242	
-	-	-		-	-
-	-	-		-	-
-	-	-		-	-
627	PURWAHARJA	675	0	661	14

Sumber : https://pikobar.jabarprov.go.id/distribution-case

B. Perhitungan Algoritma *K-Means*

Berdasarkan perhitungan Algoritma *K-Means* menunjukan 3 *cluster* yang terbentuk dengan C1 terdapat 540 Kecamatan, C2 terdapat 74 Kecamatan, dan C3 terdapat 13 data Kecamatan. Selanjutnya melakukan validitas terhadap *cluster* dengan metode *Davies-Bouldin Index* (DBI) dengan percobaan *cluster* dari k=2 hingga k=10. Nilai DBI dari perhitungan dengan Algoritma *K-Means* ditunjukkan gambar 2 berikut.

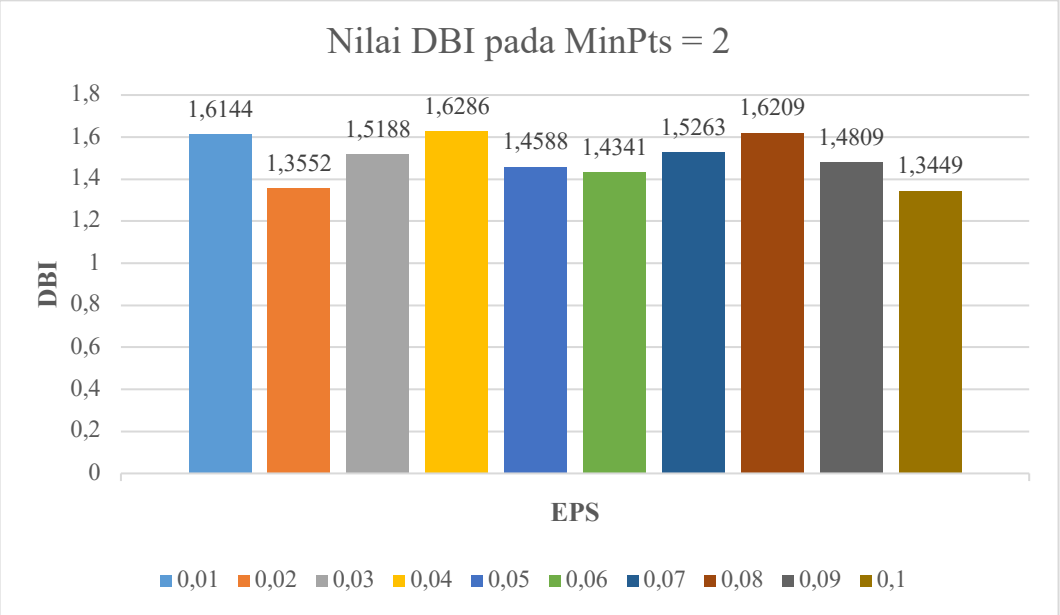


Gambar 2 Grafik nilai DBI Algoritma *K-Means*

Pada perhitungan Algoritma *K-Means* dengan percobaan *cluster* dari k=2 hingga k=10 diperoleh nilai DBI terbaik pada percobaan k=5 dengan nilai 0,4328.

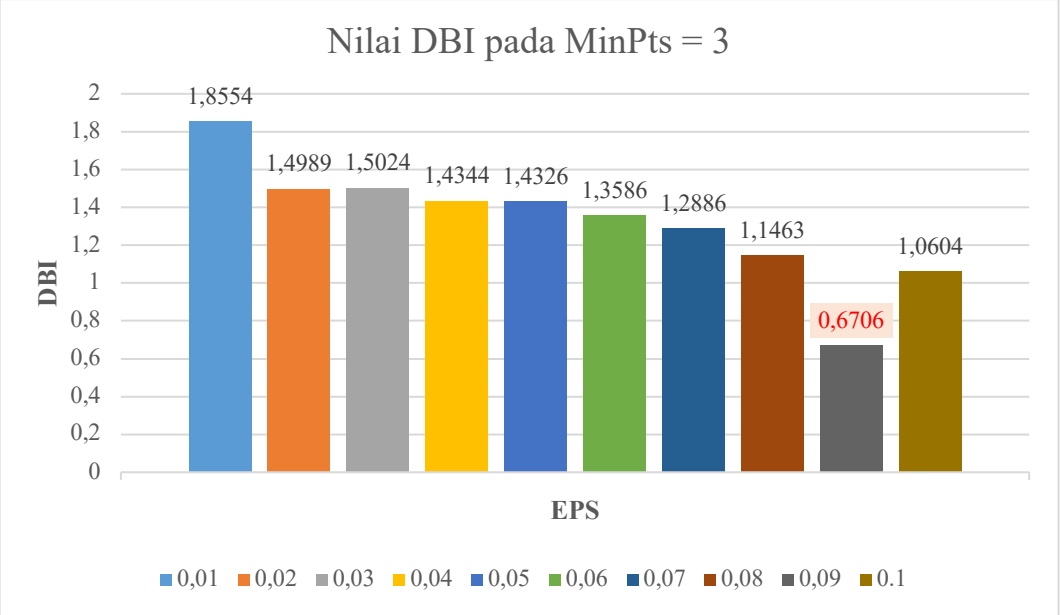
C. Perhitungan Algoritma DBSCAN

Proses perhitungan Algoritma DBSCAN dilakukan sebanyak 20 kali dengan nilai Epsilon (Eps) dan MinPoints (MinPts) pada setiap percobaan antara nilai 0,01 sampai 0,1 dengan nilai MinPts yang digunakan 2 dan 3. Selanjutnya melakukan validitas terhadap *cluster* menggunakan DBI menghasilkan nilai sebagai berikut.



Gambar 3 Grafik nilai DBI Algoritma DBSCAN MinPts = 2

ISSN: 2715-2766
Vol. IV No: 1, Januari 2023

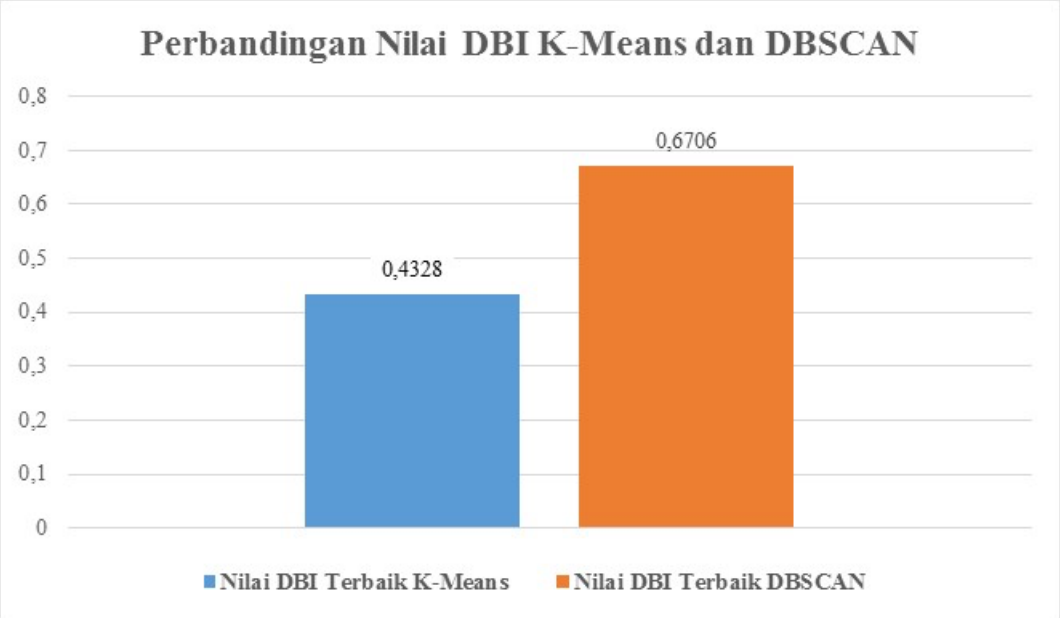


Gambar 4 Grafik nilai DBI Algoritma DBSCAN MinPts = 3

Pada gambar 3 dan gambar 4 diatas diperoleh nilai DBI terbaik pada percobaan dengan nilai Eps antara 0,01 sampai 0,1 dan nilai MinPts 3 dan 4 berada pada nilai Eps 0,09 dan MinPts 3 yaitu sebesar 0,6706 .

D. Hasil Perbandingan Algoritma

Perbandingan Algoritma *K-Means* dan DBSCAN dilakukan berdasarkan kualitas *cluster* yang terbentuk dari perhitungan nilai *Davies-Bouldin Index* (DBI). Berikut hasil grafik perbandingan nilai DBI untuk menentukan kualitas *cluster* terbaik sebagai berikut.



Gambar 5 Grafik perbandingan nilai DBI

Hasil validitas *cluster* terbaik antara Algoritma *K-Means* dan DBSCAN menggunakan metode *Davies-Bouldin Index* (DBI) untuk perhitungan Algoritma *K-Means* diperoleh nilai terbaik dalam percobaan k=5 dengan nilai sebesar 0,4328, sedangkan perhitungan Algoritma DBSCAN diperoleh nilai terbaik dalam percobaan Eps 0,09 dan MinPts 3 yaitu sebesar 0,6706 . Maka dalam penelitian ini diketahui Algoritma *K-Means* memiliki hasil validitas *cluster* yang lebih optimal daripada Algoritma DBSCAN.

IV. KESIMPULAN

Kesimpulan yang diperoleh berdasarkan perhitungan Algoritma *K-Means* dan DBSCAN dalam pengelompokan data penyebaran *Covid-19* seluruh kecamatan di provinsi Jawa Barat ditemukan nilai DBI terbaik Algoritma *K-Means* pada *cluster* k=5 dengan nilai 0,4328, sementara nilai DBI terbaik Algoritma DBSCAN ditemukan pada nilai Eps 0,09 dan MinPts 3 dengan nilai 0,6706. Dengan demikian, berdasarkan hasil dari perhitungan *Davies-Bouldin Index* (DBI) dalam menguji validitas *cluster* pengelompokan data penyebaran *Covid-19* menghasilkan Algoritma *K-Means* lebih optimal dari Algoritma DBSCAN.

Pada penelitian ini tentunya terdapat banyak kekurangan maupun kelemahan. Oleh karena itu, berikut saran yang peneliti berikan mungkin bermanfaat untuk peneliti lainnya agar lebih baik lagi yaitu :

1. Perlu dilakukan penentuan variabel data yang lebih variatif agar hasil *clustering* yang diperoleh lebih maksimal.
2. Dalam melakukan *clustering* dapat menggunakan *tools* yang lainnya untuk memperkuat hasil *clustering*.
3. Percobaan dengan Algoritma *clustering* lain untuk melakukan perbandingan Algoritma dalam pengelompokan data *Covid-19* seluruh Kecamatan di Jawa Barat untuk menemukan Algoritma yang benar-benar optimal.

PENGAKUAN

Naskah ilmiah ini merupakan bagian dari penelitian Tugas Akhir milik Bayu Biantara yang berjudul Perbandingan Algoritma *K-Means* dan DBSCAN untuk Pengelompokan Data *Covid-19* Seluruh Kecamatan di Jawa Barat, yang dibimbing oleh Tatang Rohana, ST., M.Kom dan Ayu Ratna Juwita, M.Kom.

DAFTAR PUSTAKA

[1] Kemenkes, "Pedoman Pencegahan dan Pengendalian Coronavirus Disease (COVID-19)," *Germas*, pp. 0–115, 2020.
[2] N. Dwitri, J. A. Tampubolon, S. Prayoga, F. Ilmi Zer, and D. Hartama, "Penerapan Algoritma *K-Means* Dalam Menentukan Tingkat Kepuasan Pembelajaran Online Pada Masa Pandemi Covid-19 di Indonesia," *Jti (Jurnal Teknol. Informasi)*, vol. 4, no. 1, pp. 101–105, 2020.
[3] E. Ramadanti and M. Muslih, "Analisis Persebaran Kasus Covid-19 Di Jawa Barat Menggunakan Metode *K-Means* Clustering," 2021.
[4] D. Noviyanti, A. Emma Pravitasari, and S. Sahara, "Analisis Perkembangan Wilayah Provinsi Jawa Barat Untuk Arahan Pembangunan Berbasis Wilayah Pengembangan," *J. Geogr.*, vol. 12, no. 01, p. 280, 2020.
[5] N. Mirantika, A. Tsamratul'ain, and F. D. Agnia, "Penerapan Algoritma *K-Means* Clustering Untuk Pengelompokan Penyebaran Covid-19," vol. 15, pp. 92–98, 2021.
[6] D. D. Darmansah, "Analisis Penyebaran Penularan Virus Covid-19 di Provinsi Jawa Barat Menggunakan Algoritma *K-Means* Clustering," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 3, pp. 1188–1199, 2021.
[7] S. Sindi, W. R. O. Ningse, I. A. Sihombing, F. Ilmi R.H.Zer, and D. Hartama, "Analisis algoritma K-Medoids clustering dalam pengelompokan penyebaran Covid-19 di Indonesia," *Jti (Jurnal Teknol. Informasi)*, vol. 4, no. 1, pp. 166–173, 2020.
[8] R. Adha, N. Nurhaliza, U. Soleha, P. Studi, S. Informasi, and F. Sains, "Perbandingan Algoritma DBSCAN dan *K-Means* Clustering untuk Pengelompokan Kasus Covid-19 di Dunia," vol. 18, no. 2, pp. 206–211, 2021.
[9] Z. I. Alfianti, U. Bina, S. Informatika, K. Kabupaten, J. Barat, and D. Mining, "Algoritma *K-Means*," pp. 111–122, 2020.
[10] M. Nana Nurhaliza, "Clustering of Data Covid-19 Cases in the World Using DBSCAN Algorithms Pengelompokan Data Kasus Covid-19 di Dunia Menggunakan Algoritma," vol. 1, no. 1, pp. 1–8, 2021.
[11] Y. P. Sari, A. Primajaya, and A. S. Y. Irawan, "Implementasi Algoritma *K-Means* untuk Clustering Penyebaran Tuberkulosis di Kabupaten Karawang," *INOVTEK Polbeng - Seri Inform.*, vol. 5, no. 2, p. 229, 2020.
[12] D. P. Sari, "Implementasi Algoritma *K-Means* Dalam Menentukan Tingkat Penyebaran Pandemi Covid-19 Di Sumatera Barat," *CBIS J.*, vol. 01, pp. 50–56, 2021.
[13] T. Juninda, Mustasim, and E. Andri, "Penerapan Algoritma K-Medoids untuk Pengelompokan Penyakit di Pekanbaru Riau," *Semin. Nas. Teknol. Informasi, Komun. dan Ind.*, vol. 11, no. 1, pp. 42–49, 2019.