PENERAPAN ALGORITMA *DBSCAN* DAN *K-MEANS* UNTUK *CLUSTERING* PENDERITA *PNEUMONIA* DI KABUPATEN KARAWANG

Muh. Yoga Fauzan, Garno, Yuyun Umaidah
Informatika, Univeristas Singaperbangsa Karawang
Jl. HS.Ronggo Waluyo, Puseurjaya, Kec. Telukjambe Timur, Kab. Karawang, Jawa Barat, Indonesia
myoga.fauzan@gmail.com

ABSTRAK

Pneumonia adalah infeksi pernapasan akut yang menyerang paru-paru dan dapat disebabkan oleh bakteri, virus, atau jamur. Penyakit ini menjadi salah satu penyebab utama *morbiditas* dan *mortalitas* di berbagai wilayah, termasuk Kabupaten Karawang. Penelitian ini bertujuan untuk mengklasifikasikan wilayah rawan penyebaran *pneumonia* di Kabupaten Karawang menggunakan algoritma *K-Means* dan *DBSCAN* pada rentang tahun 2019-2023. Metode *Knowledge Discovery in Database* (KDD) diterapkan dengan pemrograman *Python* dalam *Google Colab* sebagai editor teks. *DBSCAN* menghasilkan 3 *cluster*, *cluster* -1 (5 kecamatan), *cluster* 0 (40 kecamatan), dan *cluster* 1 (6 kecamatan). Sementara itu, *K-Means* menghasilkan 4 klaster: *Cluster* 0 (6 kecamatan), *Cluster* 1 (31 kecamatan), *cluster* 2 (7 kecamatan), dan *cluster* 3 (7 kecamatan). Pemilihan C= 4, menggunakan *Sum of Square Error* (SSE) menunjukkan nilai 14.54426 pada K-Means, dengan Silhouette Coefficient sebesar 0.61082, menunjukkan struktur klasterisasi yang cukup baik. DBSCAN dengan nilai epsilon 3 memiliki Silhouette Coefficient lebih tinggi, yaitu 0.76746, menunjukkan struktur klasterisasi yang lebih kuat. Selain itu, Davies-Bouldin Index (DBI) menunjukkan nilai 0.684 untuk *K-Means* dan 0.273 untuk *DBSCAN*, mengindikasikan keunggulan *DBSCAN* dalam membentuk *cluster* yang lebih optimal. Hasil penelitian ini dapat menjadi referensi dalam pengambilan keputusan terkait intervensi kesehatan untuk mengendalikan penyebaran pneumonia di Kabupaten Karawang.

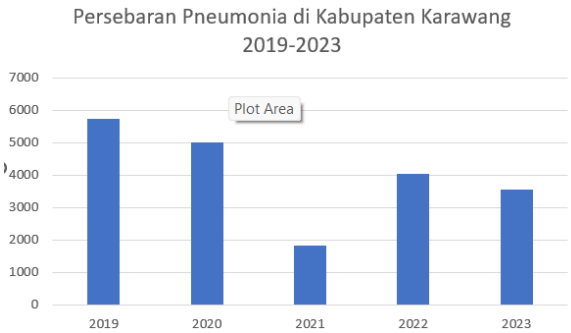
Kata kunci : *Pneumonia, Clustering, K-Means, DBSCAN, Silhouette Coefficient, Davies Bouldin Index*

1. PENDAHULUAN

Pemanfaatan teknologi informasi di bidang kesehatan memegang peranan penting dalam pelayanan medis tidak terkecuali pemetaan penyakit. Kabupaten Karawang, dengan jumlah penduduknya yang padat, menghadapi tantangan pemerataan layanan kesehatan, termasuk pengobatan *pneumonia*, yang merupakan penyebab kematian utama di Indonesia. Dengan *data mining*, analisis pola penyebaran penyakit menjadi lebih efektif dalam mendukung pengambilan keputusan di bidang kesehatan [1].

Pneumonia merupakan salah satu jenis peradangan yang menyerang jaringan paru dengan spesifik menyerang bagian *alveolus* dan *bronkiolus*. Secara umum, penyakit ini dapat disebabkan oleh berbagai *mikroorganisme* meliputi virus, jamur serta bakteri. Para penderita *Pneumonia* biasanya mengidap beberapa gejala fisik seperti demam tinggi, batuk berdahak, gangguan pernafasan hingga merasakan nyeri di ulu hati yang menyebabkan mual dan muntah [2].

Dalam kurun waktu 5 tahun kebelakang, berdasarkan data dari Dinas Kesehatan Kabupaten Karawang, yaitu pada tahun 2019-2023, terdapat 20.177 orang yang terindikasi mengidap *pneumonia*. Jumlah tersebut mengidentifikasikan adanya masalah serius terkait penyebaran dan penanganan penyakit *pneumonia* di Kabupaten Karawang. Disajikan rincian kasus *pneumonia* setiap tahunnya, terdapat pada Gambar 1.



Gambar 1. Jumlah kasus *pneumonia* di kabupaten karawang 2019-2023

Pada penelitian ini, dilakukan implemetasi *clustering* pada penderita penyakit *pneumonia* di Kabupaten Karawang, dengan memanfaatkan algoritma *DBSCAN* dan *K-Means*. Pemilihan kedua algoritma didasari karena kelebihannya didalam pengelompokkan data berdasarkan pola yang sudah ada, serta sangat jeli terhadap keberadaan *outlier* yang mana hal tersebut sangat berepengaruh terhadap proses *clustering* [3].

Diterapkan juga metodologi *Knowledge Discovery in Database* (KDD), dengan tujuan supaya penelitian lebih berjalan dengan sistematis, dimulai dengan tahapan data *selection*, *preprocessing*, *transformation*, *data mining*, hingga *evaluation* [4]. Dengan menerapkan metode ini, penelitian dapat menghasilkan model analisis yang tidak hanya valid

secara akademik, tetapi juga aplikatif dalam dunia kesehatan.

Penerapan *clustering* memikiki tujuan untuk melakukan identifikasi pola penyebaran *pneumonia* untuk optimalisasi pelayanan serta fasilitas kesehatan, baik distribusi tenaga medis maupun obat-obatan. Selain itu *clustering* juga bermanfaat untuk mengidentifikasi wilayah dengan kasus tinggi untuk dilakukan pemantauan yang lebih intensif, yang mana hal tersebut berguna untuk menganalisis lonjakan kasus yang terjadi. Sehingga penelitian ini tidak hanya berperan dalam hal akademik tetapi secara langsung memberikan manfaat bagi pihak yang bersangkutan.

Guna meningkatkan kualitas serta keefektivitasannya dalam proses data *mining*, diperlukan kecermatan, selain pengolahan data yang harus sistematis, ada faktor lain yang harus diperhatikan yaitu keberadaan *outlier*, yang sangat berpengaruh terhadap kualitas clustering [5]. Selain itu, penggunaan data harus memperhatikan kualitasnya. Seringkali data yang bernilai kosong dapat mempengaruhi hasilnya, sering terjadi ketidak akuratan data [6].

Melalui proses ini, peneliti dapat mengidentifikasi berbagai pola penyebaran penyakit yang mungkin tidak terlihat secara langsung namun memiliki dampak besar dalam pengambilan keputusan strategis dalam bidang kesehatan.

Maka berdasarkan latar belakang diatas, tujuan dari penelitian ini ialah untuk menentukan daerah mana yang memiliki penyebaran kasus *pneumonia* yang tertinggi di Kabupaten Karawang. Diharapkan dengan adanya penelitian ini, akan memberikan masukan untuk pemerintah Kabupaten Karawang mengenai perhatian yang lebih serius terhadap daerah-daerah yang menjadi klaster tertinggi.

2. TINJAUAN PUSTAKA

2.1. *Data Mining*

Data *mining* merupakan suatu upaya dilakukan untuk mengeksplorasi data menggunakan *statistika* matematika dan *artificial intelegent* untuk mengelompokan informasi penting yang tentunya mempunyai manfaat yang terkandung dalam suatu *database* yang besar [7].

2.2. *Knowledge Discovery in Database*

Knowledge Discovery in Database merupakan Langkah dalam melakukan pengolahan yang berasal dari data mentah menjadi data yang sudah berbentuk informasi dan dapat dimanfaatkan oleh orang lain .Pendapat lain yang dikemukakan oleh

2.3. *K-Means*

K-Means yaitu salah satu algoritma dapat digunakan untuk melatih data tanpa adanya kontrol (*unsupervised learning*). Algoritma ini diciptakan oleh Stuart Lloyd akhir abad ke-19. K-MEANS menjadi salah satu algoritma *clustering* yang banyak digunakan, dikarenakan algoritma ini tergolong mudah dipakai serta dijalankan. Selain itu algoritma ini juga memiliki kecepatan saat mengolah data dalam bentuk apapun [4].

2.4. *DBSCAN*

Secara bahasa *Density-Based Spatial Clustering of Application with Noise* (DBSCAN) dapat diartikan sebagai salah satu algoritma yang berfokus ke dalam pengelompokkan objek ke dalam satu *cluster* tanpa memperhatikan nilai atau kategorinya belum diketahui sama sekali. Algoritma ini berfokus ke dalam pengelompokan wilayah dengan tingkat kepadatan yang sangat tinggi dalam sebuah *database* yang acak. Dalam hal ini, *DBSCAN* menggunakan dua parameter yang menjadi acuan dalam melakukan *cluster*, yang pertama ada *epsilon* dan (ε) *minimum point* (MinPts) [8].

2.5. *Silhouette Coefficient*

Silhouette Coefficient merupakan salah satu cara dalam mengukur serta menganalisis nilai dari suatu data. *Silhouette Coefficient* sendiri merupakan gabungan dari dua perhitungan, yang pertama yaitu *cohesion* dan juga *separation*. *Cohesion* sendiri merupakan sebuah rumus yang digunakan untuk menghitung korelasi antara suatu objek dengan objek lainnya. Adapun *separation* mempunyai tujuan untuk menghitung besaran jarak antar cluster atau seberapa besar perbedaan cluster nya [4].

Adapun rangkaian perhitungan di dalam silhouette coefficient sebagai berikut didalam persamaan (1) [3].

$$sil(c) = sil(k)\frac{1}{|k|}\sum_{i=1}^{k} sil(c_i) \qquad (1)$$

Keterangan:
sil (k) : *Value* dari *Silhouette* keseluruhan *cluster*
|k| : Banyak jumlah *cluster* k
sil (ci) : *Mean* nilai *Silhouette*

Berikut disajikan tabel 1, yang berisikan rentang nilai dari *silhouette coefficient*.

Tabel 1. Rentang Nilai *Silhouette Coefficient*

Rentang Nilai	Keterangan
0.71 – 100	Kuat
0.51 – 0.70	Baik
0.26 – 0.50	Lemah
≤ 0.25	Buruk

Range nilai dari *Silhouette Coefficient* berkisar dari nilai -1 sampai 1. *Mean* yang dihasilkan dari *Silhouette Coefficient* dari suatu objek yang merupakan anggota dari *cluster*. Dalam hal ini akan dijadikan sebagai indikator kecocokan antara semua anggota data yang berada dalam *cluster*. Apabila suatu *cluster* mendekati bahkan hampir menyentuh nilai 1 maka akan semakin baik juga pengelompokan data tersebut. Akan tetapi, ketika nilai yang dihasilkan mendekati dan menyentuh nilai -1 maka akan dikatakan semakin buruk.

2.6. *Davies Bouldin Index*

Davies Bouldin Index atau yang bisa disebut juga sebagai validitas pengukuran klasifikasi, pertama kali ditemukan dan dipopulerkan oleh *David L. Davies* dan temannya yaitu *Donald W. Bouldin* pada tahun 1979. Teori ini didefinisikan sebagai mean jarak dari setiap cluster [9].

Semakin rendah index yang dihasilkan, maka semakin baik juga kualitas dari clustering yang dihasilkan [10].

Dalam perhitungannya, menggunakan rumus seperti yang tercantum dibawah ini dalam persamaan (2).

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j} \, (R_{i,j}) \qquad (2)$$

Keterangan:

K : Jumlah kelompok

$R_{i,\,j}$: Rasio (i) dan (j)

Max : Mencari rasio terbesar

3. METODE PENELITIAN

3.1. Sumber Data

Objek yang digunakan dalam penelitian kali ini kalsterisasi daerah penyebaran penyakit *pneumonia* dengan studi kasus wilayah Kabupaten Karawang dengan mengaplikasikan algoirtma *DBSCAN* dan *K-Means*. Data yang dipakai merupakan data penyebaran penyakit *Pneumonia* di Kabupaten Karawang dengan rentang waktu 2019-2023 yang diperoleh dari data primer yaitu dari Dinas Kesehatan Kabupaten Karawang. Berikut disajikan contoh dataset yang digunakan pada gambar 2.

No	Nama Kecamatan	Tahun	Pneumonia (L < 1 Tahun)	Pneumonia (P < 1 Tahun)	Pneumonia (L 1 -< 5 Tahun)
1	ADIARSA	2019-2023	78	63	240
2	ANGGADITA	2019-2023	26	25	110
3	BALONGSARI	2019-2023	17	19	127
4	BATUJAYA	2019-2023	119	96	162
5	BAYUR LOR	2019-2023	52	47	61
6	CIAMPEL	2019-2023	22	18	92
7	CIBUAYA	2019-2023	152	101	174
8	CICINDE	2019-2023	13	17	155
9	CIKAMPEK	2019-2023	97	73	185
10	CIKAMPEK UTARA	2019-2023	131	150	170
11	CILAMAYA	2019-2023	55	33	90
12	CURUG	2019-2023	72	61	51
13	GEMPOL	2019-2023	3	10	46
14	JATISARI	2019-2023	64	55	130
15	JAYAKERTA	2019-2023	11	13	45

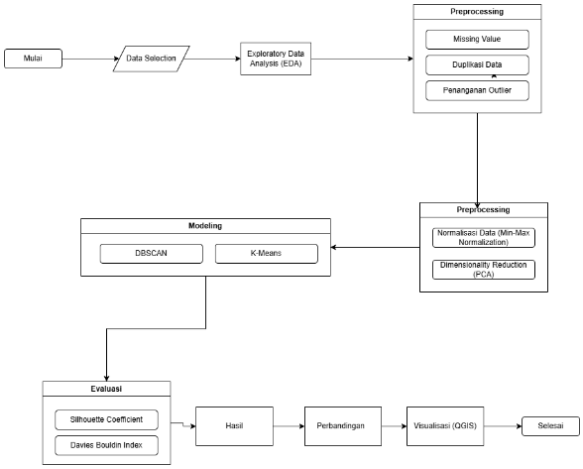
Gambar 2. Contoh dataset penderita *pneumonia* 2019-2023

3.2. Teknik Pengumpulan Data

Pengumpulan data dilakukan secara langsung melalui observasi lapangan untuk memperoleh informasi yang akurat dan relevan. Selain itu, data juga dikumpulkan dengan mengajukan permohonan resmi serta berkoordinasi langsung dengan Dinas Kesehatan Kabupaten Karawang guna mendapatkan data yang valid dan sesuai dengan kebutuhan penelitian. Metode ini memastikan bahwa data yang diperoleh berasal dari sumber terpercaya dan mencerminkan kondisi sebenarnya di lapangan.

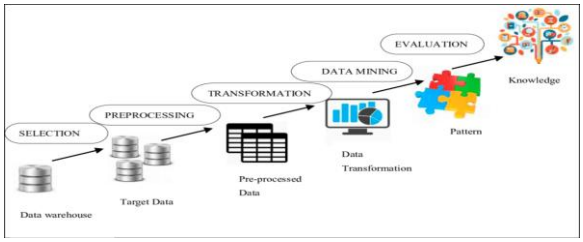
3.3. Tahapan Perancangan

Penelitian ini membandingkan algoritma K-Means dan DBSCAN untuk clustering kasus pneumonia di Kabupaten Karawang. Komparasi dilakukan untuk menentukan metode yang paling optimal dalam mengelompokkan data berdasarkan pola penyebaran penyakit. K-Means dipilih karena kemampuannya membentuk klaster terstruktur, sementara DBSCAN unggul dalam mendeteksi outlier. Alur metode clustering ditunjukkan pada gambar 3.



Gambar 3. Rancangan Penelitian

Gambar diatas menunjukan rancangan penelitian dengan membandingkan kedua algoritma. Untuk analisis dan pemrosesan data yang lebih terperinci, penelitian ini menerapkan metode *Knowledge Discovery in Databases* (KDD) pada gambar 4 dibawah.



Gambar 4. Proses KDD

Disajikan gambar 3, sebagai tahapan yang ada pada *knowledge discovery in database*, yang terdiri dari:

a. *Data Selection*
Pada tahapan ini dilakukan pengambilan serta pemilihan data yang akan dilakukan proses pengolahan. Dalam hal ini data bersumber dari Dinas Kesehatan Kabupaten Karawang, dengan data penderita penyakit *pneumonia* rentang tahun 2019-2023

b. *Data Pre-processing*
Untuk proses ini, data akan mengalami pembersihan. Diantaranya yaitu pembersihan dari *missing value*, *outlier*, dan duplikasi data.

c. *Data Tranformation*
Setelah melewati tahapan pembersihan, selanjutnya data akan dilakukan tranformasi kedalam format yang lebih mudah untuk dilakukan proses *data mining*.

d. *Data Mining*
Untuk langkah selanjutnya yaitu *data mining*. Memiliki tujuan untuk menerapkan *model* atau algoritma yang dipilih untuk menemukan pola atau informasi yang terkandung didalamnya.

e. *Evaluation*
Tahapan terakhir yaitu dilakukannya proses evaluasi, yang bertujuan untuk mencari serta mengukur kualitas dari suatu *cluster*. Dalam kasus ini menggunakan *silhouette coefficient* dan *davies bouldin index.*

4. HASIL DAN PEMBAHASAN

4.1. *Data Selection*

Pada tahap *data selection*, akan dilakukan proses pemilihan variabel dari data yang diperoleh dari Dinas Kesehatan Kabupaten Karawang. Dataset yang diperoleh merupakan data penderita *Pneumonia* di Kabupaten Karawang pada periode 2019-2023, dengan 17 atribut yaitu 'No', 'Nama Kecamatan', 'Tahun', 'Pneumonia (L < 1 Tahun)', 'Pneumonia (P < 1 Tahun)', 'Pneumonia (L 1 -< 5 Tahun)', 'Pneumonia (P 1 -< 5 Tahun)', 'PB (L < 1 Tahun)', 'PB (P < 1 Tahun)', 'PB (L 1 -< 5 Tahun)', 'PB (P 1-<5 Tahun)', 'Laki-Laki', 'Perempuan', 'Total Kasus', 'Meninggal (L)', 'Meninggal (P)', 'Total Meninggal'. Dan dari semua atribut yang ada, 4 atribut dihilangkan karena tidak sesuai dengan kriteria clustering, yaitu atribut ‘No’, ‘Tahun’, ‘Meninggal (L)’, dan ‘Meninggal (P)’. Proses data selection terdapat pada Gambar 5 sedangkan untuk hasilnya terdapat pada Gambar 6.

```
x = data[['Pneumonia (L < 1 Tahun)', 'Pneumonia (P < 1 Tahun)',
x
```

Gambar 5. Contoh proses *data selection*

	Pneumonia (L < 1 Tahun)	Pneumonia (P < 1 Tahun)	Pneumonia (L 1 -< 5 Tahun)
0	78	63	240
1	26	25	110
2	17	19	127
3	119	96	162
4	52	47	61
5	22	18	92
6	152	101	174
7	13	17	155

Gambar 6. Contoh hasil *data selection*

4.2. *Data Pre-Processing*

Selanjutnya pada tahapan data preprocessing, dilakukan proses pembersihan data dengan tujuan menghilangkan *missing value* dan *data duplicate*. Gambar 7, menunjukkan semua atribut dilakukan pengecekan missing value. Dan didapatkan hasil bahwa data tidak mempunyai missing value.

```
print(x.isnull().sum())
```

```
Pneumonia (L < 1 Tahun)      0
Pneumonia (P < 1 Tahun)      0
Pneumonia (L 1 -< 5 Tahun)   0
Pneumonia (P 1 -< 5 Tahun)   0
PB (L < 1 Tahun)             0
PB (P < 1 Tahun)             0
PB (L 1 -< 5 Tahun)          0
PB (P 1 -< 5 Tahun)          0
Laki-Laki                    0
Perempuan                    0
Total Kasus                  0
```

Gambar 7. Hasil *missing value*

Selain dilakukan pengecekan *missing value*, juga dilakukan pengecekan duplikasi data pada setiap atribut, guna memastikan semua data bernilai unik dan tidak ada yang sama, disajikan gambar 8 sebagai hasilnya.

```
x.duplicated().sum()
```

```
0
```

Gambar 8. Hasil Cek Duplikasi Data

4.3. *Data Transformation*

Tahap berikutnya adalah transformasi data menggunakan metode min-max normalization untuk algoritma K-Means. Teknik ini diterapkan untuk memastikan bahwa setiap atribut memiliki skala yang sama, dengan mengonversi seluruh nilai data ke dalam rentang 0-1. Transformasi ini bertujuan untuk mempermudah proses pengolahan data. Sekaligus meningkatkan kualitas evaluasi data. Proses transformasi data ditunjukkan pada Gambar 9 dan hasilnya pada Gambar 10.

```
minmax=preprocessing.MinMaxScaler().fit_transform(x_clipped)
data_transform = pd.DataFrame(minmax,index=x_clipped.index, columns=x_clipped.columns)
data_transform
```

Gambar 9. Contoh proses *min-max normalization*

Pneumonia (L < 1 Tahun)	Pneumonia (P < 1 Tahun)	Pneumonia (L 1 -< 5 Tahun)
0.439306	0.472795	0.801358
0.138728	0.187617	0.359932
0.086705	0.142589	0.417657
0.676301	0.720450	0.536503
0.289017	0.352720	0.193548
0.115607	0.135084	0.298812
0.867052	0.757974	0.577250

Gambar 10. Contoh hasil *min-max normalization*

Pada algoritma *DBSCAN*, *dimensionality reduction* dengan PCA diterapkan untuk mengurangi fitur, meningkatkan efisiensi, serta mengatasi *curse of dimensionality* dan *overfitting*. Prosesnya ditampilkan pada Gambar 11.

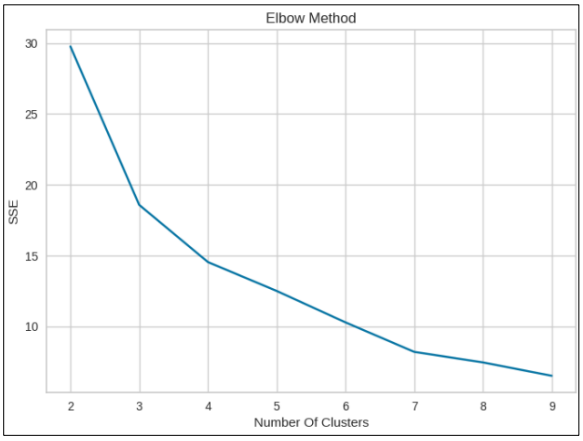
```
from sklearn.decomposition import PCA
pca=PCA(n_components=2)
data=pca.fit_transform(data)

data_new=pd.DataFrame(data,columns=['X','Y'])
```

Gambar 11. Proses *dimensionality reduction* dengan pca

4.4. *Data Mining*

4.4.1. *K-Means Clustering*

Metode elbow digunakan untuk menentukan *cluster* optimal dengan menghitung *Sum of Square Error* (SSE), seperti ditunjukkan pada Gambar 12.



Gambar 12. Grafik metode *elbow*

Gambar 10 menunjukkan penentuan *cluster* menggunakan metode *elbow*. Titik siku (elbow) terlihat pada cluster 4, di mana grafik mulai mendatar dengan perubahan nilai yang tidak signifikan. Oleh karena itu, jumlah cluster terbaik adalah 4. Setelah jumlah cluster optimal ditentukan, yaitu 4, dilakukan pemodelan menggunakan metode K-Means. Data didistribusikan ke dalam 4 *cluster*: *cluster* 0 (C0), *cluster* 1 (C1), *cluster* 2 (C2), dan *cluster* 3 (C3). Seperti pada Gambar 13.

```
array([[0.73121387, 0.79643527, 0.70769666, 0.77278776, 0.08484848,
        0.08333333, 0.02857143, 0.06666667, 0.64169675, 0.68405339,
        0.65576923],
       [0.13071042, 0.17091327, 0.19617723, 0.23192325, 0.07741935,
        0.06129032, 0.04423963, 0.07741935, 0.14842203, 0.17417315,
        0.15852978],
       [0.76630884, 0.83918521, 0.69463983, 0.71285999, 0.82857143,
        0.8       , 0.86530612, 0.82857143, 0.7062919 , 0.72542526,
        0.70879121],
       [0.25681255, 0.24122219, 0.20858598, 0.22469347, 0.77662338,
        0.97142857, 0.7755102 , 0.7047619 , 0.2599278 , 0.25139244,
        0.25315934]])
```

Gambar 13. Hasil perhitungan *k-means*

Jumlah anggota setiap klaster dijelaskan melalui deskripsi pada Gambar 14.

```
Cluster 0 : 6
Cluster 1 : 31
Cluster 2 : 7
Cluster 3 : 7
```

Gambar 14. Anggota *cluster k-means*

Gambar 14, menunjukan anggota dari setiap *cluster* yang ada pada algoritma *k-means*, dengan rincian *cluster* 0 berjumlah 6, *cluster* 1 berjumlah 31, *cluster* 2 berjumlah 7 dan *cluster* 3 berjumlah 7.

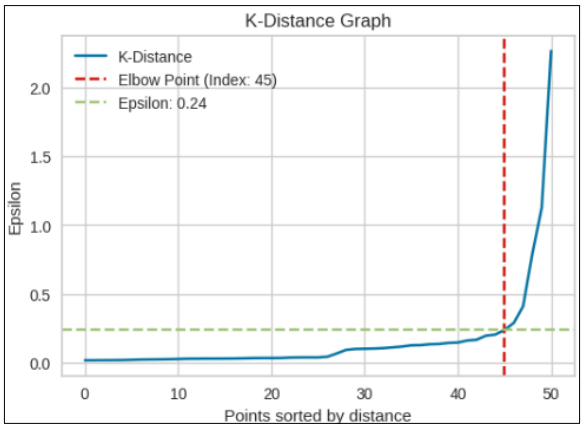
4.4.2. *DBSCAN*

Selanjutnya pada tahapan ini, dilakukan pencarian jumlah *epsilon* yang optimal menggunakan *Kneelocator*, seperti yang tertera pada Gambar 15.

```
kneedle = KneeLocator(
    x=range(len(dis)),
    y=dis,
    curve="convex",
    direction="increasing",
    S=1.0
)
epsilon_index = kneedle.elbow
epsilon_value = dis[epsilon_index]
```

Gambar 15. Proses pencarian epsilon

Setelah dilakukan proses pencarian, maka langkah selanjutnya merupakan visualisasi hasil kedalam bentuk grafik, disajikan pada Gambar 16.



Gambar 16. Hasil pencarian *epsilon*

Grafik K-Distance pada DBSCAN digunakan untuk menentukan nilai epsilon (ε) sebagai batas jarak maksimum antar titik. Garis biru menunjukkan jarak ke tetangga terdekat, sedangkan garis merah menandai titik siku pada indeks 45 dengan kenaikan yang signifikan. Nilai epsilon optimal yang dipilih adalah 0,24. Disajikan gambar 17 berisi proses *clustering* nya.

```
cluster=DBSCAN(eps=0.24,min_samples=5)
data_new['Label']=cluster.fit_predict(data_new)
```

Gambar 17. Proses *clustering dbscan*

Pada Gambar 4.15, *eps*=0.24 menentukan jarak maksimum antar titik, sementara *min_samples*=5 menetapkan jumlah minimum data dalam radius *epsilon* untuk membentuk *cluster*. Data diklasifikasikan dengan *fit_predict()*, menghasilkan

label, di mana noise diberi label -1, seperti yang disajikan pada gambar 18.

	X	Y	Label
0	0.155258	-0.039367	0
1	-0.189566	-0.087371	0
2	-0.208070	-0.126304	0
3	0.579160	0.378206	1
4	0.454050	0.515617	1

Gambar 18. Hasil *clustering dbscan*

Jumlah anggota setiap klaster dijelaskan melalui deskripsi pada Gambar 19.

```
Jumlah anggota setiap cluster (DBSCAN):
Label
-1     5
 0    40
 1     6
```

Gambar 19. Jumlah anggota *cluster dbscan*

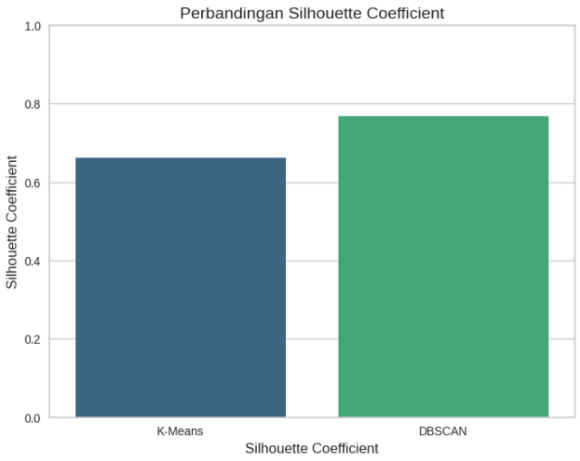
Gambar 19, menunjukan anggota dari setiap *cluster* yang ada pada algoritma *dbscan*, dengan rician *cluster -1* berjumlah 5, *cluster* 0 berjumlah 40 dan *cluster* 1 berjumlah 6.

4.5. *Evaluation*

Tahap terakhir adalah Evaluasi, di mana hasil analisis dinilai untuk melihat kesesuaian dengan tujuan penelitian. Evaluasi akan menggunakan metode *Silhouette Coefficient* dan *Davies-Bouldin Index.* Pada gambar 20 disajikan hasil nilai *silhouette coefficient*, dan gambar 21 grafik perbandingannya.

```
Silhouette Coefficient untuk K-Means: 0.6607292033393881
Silhouette Coefficient untuk DBSCAN: 0.767461522133103
```

Gambar 20. Hasil *silhouette coefficient*



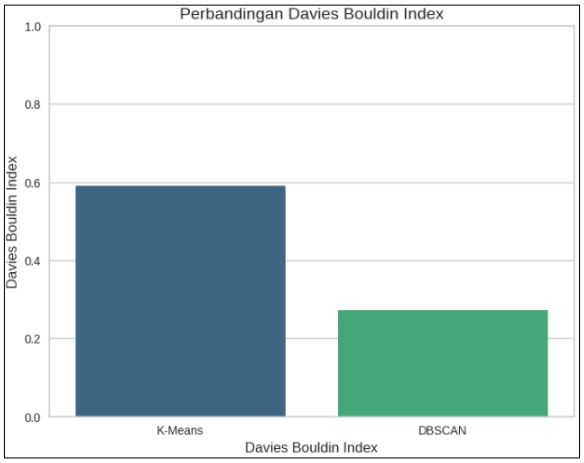
Gambar 21. Grafik perbandingan *Silhouette coefficient*

Hasil evaluasi menggunakan *Silhouette Coefficient* menunjukkan bahwa *DBSCAN* memiliki kinerja lebih unggul dibandingkan *K-Means* dalam mengelompokkan data *pneumonia* di Karawang 2019-2023, dengan nilai 0.7674, lebih tinggi dari K-Means yang hanya 0.6108. *DBSCAN* lebih optimal dalam membentuk klaster yang jelas dan terstruktur.

Selanjutnya akan dilakukan evaluasi dengan menggunakan *davies bouldin index*, seperti yang terlihatr pada gambar 22 dan gambar 23.

```
Davies-Bouldin Index K-Means: 0.590
Davies-Bouldin Index DBSCAN: 0.273
```

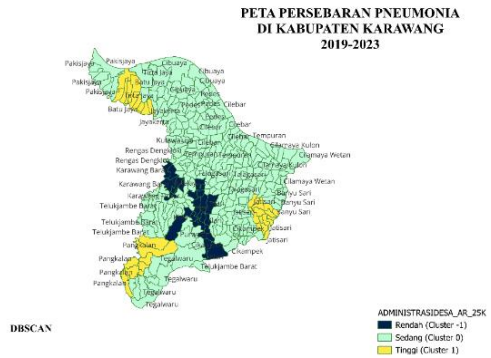
Gambar 22. Nilai *davies bouldin index*



Gambar 23. Grafik perbandingan *davies bouldin index*

DBSCAN menunjukkan performa lebih baik daripada *K-Means*, dengan DBI lebih rendah dan *Silhouette Coefficient* lebih tinggi, menandakan klaster lebih jelas dan terpisah. Setelah *data mining* selesai, laporan berisi pemetaan QGIS dan hasil *clustering* akan diserahkan ke P2PTM Dinas Kesehatan Karawang untuk analisis persebaran *pneumonia* 2019–2023 dan menentukan tindakan di daerah dengan kasus tinggi.

Berikut disajikan gambar 24 sebagai visualisasi *cluster DBSCAn*, dengan menggunakan *software QGIS.*

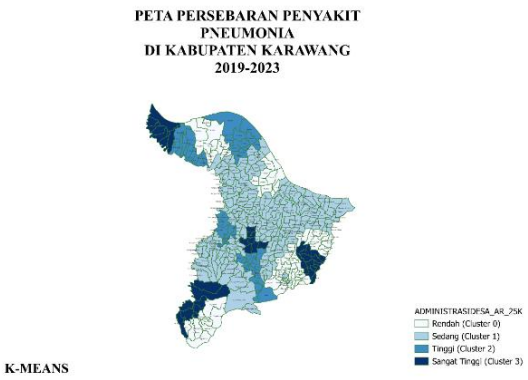


Gambar 24. Interface persebaran penderita *pneumonia* 2019-2023 (*dbscan*)

Disajikan tabel 2, yang berisikan anggota dari setiap *cluster* yang ada pada *clustering* menggunakan *DBSCAN*.

Tabel 2. Anggota *cluster dbscan*

Cluster	Nama Anggota	Jumlah	Label
-1	Cikampek Utara, Karawang Kulon, Kertamukti, Klari, Majalaya	5	Rendah
0	Adiarsa, Anggadita, Balongsari, Ciampel, Cibuaya, Cicinde, Cikampek, Cilamaya, Curug, Gempol, Jayakerta, Jomin, Kalangsari, Karawang, Kota Baru, Kutamukti, Kutawaluya, Lemah Duhur, Lemah Abang, Loji, Medang Asem, Nagasari, Pacing, Pakisjaya, Pasirukem, Pedes, Plawad, Purwasari, Rawamerta, Rngsdengklok, Sungai Buntu, Tanjungpura, Telagasari, Teluk Jambe, Tempuran, Tirtajaya, Tirtamulya, Tunggak Jati, Wanakerta Rs/Sumber Lain	40	Sedang
1	Batujaya Bayur Lor, Jatisari, Pangkalan, Sukatani, Wadas	6	Tinggi



Gambar 25. *Interface* persebaran penderita *pneumonia* 2019-2023 (*k-means*)

Tabel 2 menampilkan hasil *clustering* dengan menggunakan algoritma *dbscan* dengan daerah berdasarkan jumlah anggota dan label kategori. Tabel terdiri dari empat kolom utama: *Cluster*, Nama Anggota, Jumlah, dan Label. *Cluster* -1 memiliki 5 anggota dan diberi label Rendah, *cluster* berikutnya memiliki 40 anggota dengan label Sedang, sementara *cluster* lainnya memiliki 6 anggota dengan label Tinggi. Nama anggota dalam setiap cluster mencantumkan berbagai daerah yang termasuk dalam kategori tersebut.

Selanjutnya disajikan gambar 25, sebagai visualisasi cluster dengan menggunakan algoritma k-means.

Disajikan tabe1 3, yang berisikan anggota dari setiap *cluster* yang ada pada *clustering* menggunakan *K-Means.*

Tabel 3. Anggota *cluster k-means*

Cluster	Anggota	Jumlah	Karakteristik	Kriteria
0	Cikampek, Kertamukti, Kota Baru, Tanjungpura, Tirtajaya, Tirtamulya	6	0 < 250 Kasus	Rendah
1	Anggadita, Balongsari, Ciampel, Cicinde, Cilamaya, Curug, Gempol, Jayakerta, Jomin, Kalangsari, Karawang, Kutamukti, Kutawaluya, Lemah Duhur, Lemah Abang, Loji, Medang Asem, Nagasari, Pacing, Pasirukem, Pedes, Plawad, Purwasari, Rawamerta, Rngsdengklok, Sungai Buntu, Telagasari, Teluk Jambe, Tempuran, Tunggak Jati, Wanakerta	31	250–500 Kasus	Sedang
2	Adiarsa, Batujaya, Cibuaya, Cikampek Utara, Krawangkulon, Klari, Wadas	7	500–1000 Kasus	Tinggi
3	Bayur Lor, Jatisari, Majalaya, Pakisjaya, Pangkalan, Sukatani, Rs/Sumber Lain	7	> 1000 kasus	Sangat Tinggi

Tabel tersebut menyajikan hasil *clustering* daerah dengan menggunakan algoritma *k-means*, berdasarkan jumlah kasus yang tercatat dalam setiap kelompok. Tabel terdiri dari lima kolom utama: *Cluste*r, Anggota, Jumlah, Karakteristik, dan Kriteria. *Cluster* 0 memiliki 6 daerah dengan jumlah kasus di bawah 250 dan dikategorikan sebagai Rendah. *Cluster* 1 mencakup 31 daerah dengan 250–500 kasus dan diberi label Sedang. *Cluster* 2 terdiri dari 7 daerah dengan 500–1000 kasus dan masuk dalam kategori Tinggi. Sementara itu, *cluster* 3 juga mencakup 7 daerah tetapi memiliki lebih dari 1000 kasus, sehingga dikategorikan sebagai Sangat Tinggi.

5. KESIMPULAN DAN SARAN

Langkah awal untuk mengatasi ketidakmerataan pelayanan kesehatan dilakukan dengan menerapkan algoritma *K-Means* dan *DBSCAN* untuk melakukan *clustering* data penyakit *pneumonia* berdasarkan karakteristik wilayah. Hasil klasterisasi menunjukkan pengelompokan wilayah ke dalam kategori rendah, sedang, tinggi, dan sangat tinggi, sehingga intervensi dapat diprioritaskan pada wilayah dengan tingkat kasus yang lebih tinggi guna mengurangi ketidakmerataan dalam pelayanan kesehatan. Untuk mengetahui metode yang paling tepat dalam menangani ketidakmerataan ini, dilakukan evaluasi dengan membandingkan hasil *clustering*

menggunakan metrik *Silhouette Coefficient* dan *Davies-Bouldin Index* (DBI). Hasil evaluasi menunjukkan bahwa algoritma *DBSCAN* lebih unggul dibandingkan *K-Means*, dengan *Silhouette Coefficient* sebesar 0.76746 yang menandakan struktur *cluster* yang kuat serta DBI sebesar 0.273 yang menunjukkan kualitas *clustering* yang baik, sehingga *DBSCAN* dapat dianggap sebagai metode yang lebih tepat dalam studi ini.

Penelitian ini menggunakan data dari tahun 2019 hingga 2023, dan diharapkan pada penelitian berikutnya cakupan data dapat diperluas dengan rentang waktu yang lebih panjang serta menambahkan variabel seperti kepadatan penduduk, sanitasi, dan tingkat pendidikan masyarakat untuk hasil yang lebih mendalam. Selain itu, disarankan untuk menggunakan algoritma lain seperti *Agglomerative Hierarchical Clustering* atau *Gaussian Mixture Model* guna membandingkan hasil klasterisasi dan menentukan algoritma yang paling optimal. Penelitian di masa depan juga dapat mengklasterisasi penyakit lain yang relevan di wilayah Kabupaten Karawang agar memberikan gambaran yang lebih komprehensif mengenai persebaran penyakit menular.

DAFTAR PUSTAKA

[1] M. Andriani, A. Manaor, H. Pardede, and M. Simanjuntak, "Pengelompokan Penyakit pada Pasien Berdasarkan Usia dengan Metode K-Means Clustering mining yang cara kerjanya mencari dan mengelompokan data yang mempunyai kemiripan Data Mining tersembunyi pada suatu database yang sangat besar sehingga ditemukan suatu p," no. 4, 2024.

[2] Y. Isnaini, "Pemodelan Spatial Autoregressive pada Faktor-Faktor yang Mempengaruhi Penyakit Pneumonia Balita di Provinsi Riau [Skripsi]," 2022.

[3] D. P. S. Dewi, "Analisis Clustering Penyebaran Hiv Di Karawang Berdasarkan Kecamatan Dengan Algoritma K-Means," *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.4878.

[4] I. Nur Amalia, Y. Umaidah, and R. Mayasari, "Penerapan Data Mining Untuk Klasterisasi Daerah Rawan Penyakit Menular Di Kabupaten Karawang Dengan Menggunakan Algoritma K-Means," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 4, pp. 5582–5591, 2024, doi: 10.36040/jati.v8i4.9953.

[5] S. Widaningsih, "Penerapan Data Mining untuk Memprediksi Siswa Berprestasi dengan Menggunakan Algoritma K Nearest Neighbor," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 3, pp. 2598–2611, 2022, doi: 10.35957/jatisi.v9i3.859.

[6] Rachmadhany Iman, Basuki Rahmat, and Achmad Junaidi, "Implementasi Algoritma K-Means dan Knearest Neighbors (KNN) Untuk Identifikasi Penyakit Tuberkulosis Pada Paru-Paru," *Repeater Publ. Tek. Inform. dan Jar.*, vol. 2, no. 3, pp. 12–25, 2024, doi: 10.62951/repeater.v2i3.77.

[7] K. Faradila Ilena Putri , Retno Damayanti, "Penerapan Algoritma K-Means," vol. 2, no. mei 2022, pp. 408–418, 2022.

[8] W. Di and M. Raya, "K-MEANS UNTUK ANALISIS POLA PENYEBARAN TEMPAT K-MEANS UNTUK ANALISIS POLA PENYEBARAN TEMPAT," 2023.

[9] I. T. Umagapi, B. Umaternate, S. Komputer, P. Pasca Sarjana Universitas Handayani, B. Kepegawaian Daerah Kabupaten Pulau Morotai, and B. Riset dan Inovasi, "Uji Kinerja K-Means Clustering Menggunakan Davies-Bouldin Index Pada Pengelompokan Data Prestasi Siswa," *Semin. Nas. Sist. Inf. dan Teknol.*, vol. 7, no. 1, pp. 303–308, 2023.

[10] S. Perdana Putra and D. Aryo Anggoro, "Analisis Clustering Global Living Cost Berdasarkan Socioeconomic Status Menggunakan Algoritma DBSCAN," *J. MEDIA Inform. BUDIDARMA*, vol. 8, no. 2, p. 820, Apr. 2024, doi: 10.30865/mib.v8i2.7567.