

PENERAPAN DATA MINING UNTUK KLASTERISASI DAERAH RAWAN PENYAKIT MENULAR DI KABUPATEN KARAWANG DENGAN MENGGUNAKAN ALGORITMA K-MEANS

Indira Nur Amalia, Yuyun Umaidah, Rini Mayasari

Informatika, Universitas Singaperbangsa Karawang
Jl. HS. Ronggowaluyo, Telukjambe Timur, Karawang, Indonesia
indiraamalia9@gmail.com

ABSTRAK

Penyakit menular adalah varian penyakit infeksi yang disebabkan oleh mikroorganisme seperti virus, bakteri, parasit, atau jamur, yang dapat dengan mudah menular ke individu lain yang masih dalam keadaan sehat. Kabupaten Karawang telah mengalami perkembangan positif, termasuk dalam hal pendapatan daerah, situasi ekonomi, dan tingkat pendidikan penduduknya. Kenaikan populasi dan ukuran wilayah Kabupaten Karawang tidak terlepas dari aspek kesehatan penduduknya. Saat ini, jumlah kasus penyakit menular semakin meningkat setiap tahun, sehingga mungkin diperlukan upaya untuk mengelompokkan daerah penyebaran penyakit tersebut. Tujuan dari penelitian ini adalah untuk menerapkan algoritma *K-means* dalam melakukan klusterisasi daerah rawan penyakit menular di Kabupaten Karawang. Metode yang digunakan dalam penelitian ini adalah dengan menggunakan metode KDD (*Knowledge Discovery in Database*) dengan tools R Studio. Hasil dari penelitian ini adalah klusterisasi data penyakit menular menjadi 3 kluster, yaitu kluster 1 sebanyak 6 kecamatan, kluster 2 sebanyak 15 kecamatan, dan kluster 3 sebanyak 9 kecamatan. Dengan hasil evaluasi menggunakan SSE dan DBI menghasilkan titik kluster optimal di $C=3$ dengan nilai SSE 80.0712 dan nilai DBI sebesar 0.6507.

Kata kunci : Penyakit Menular, Klusterisasi, Kabupaten Karawang, *K-means*.

1. PENDAHULUAN

Penyakit menular adalah varian penyakit infeksi yang disebabkan oleh mikroorganisme seperti virus, bakteri, parasit, atau jamur, yang dapat dengan mudah menular ke individu lain yang masih dalam keadaan sehat. Upaya pencegahan penyakit menular tertentu yang sering terjadi di Indonesia melibatkan vaksinasi serta menerapkan gaya hidup yang bersih dan sehat [1]. Penyakit menular dapat disebarkan melalui dua cara, baik secara langsung maupun tidak langsung. Seseorang yang menjadi sumber penularan mungkin tidak menunjukkan gejala penyakit dan tampak sehat, terutama jika dia hanya menjadi pembawa (*carrier*) penyakit. Penyakit menular juga bisa menyebar secara tidak langsung, seperti ketika seseorang menyentuh permukaan yang terkontaminasi seperti gagang pintu, kran air, atau pegangan di kereta [2].

Dalam konteks penyakit menular, ini diwakili dengan jelas dalam berkomitmen untuk mengakhiri epidemi AIDS, tuberkulosis (TB), malaria, serta penyakit tropis yang terabaikan, sambil berupaya mengatasi hepatitis, penyakit yang bersumber dari air, dan penyakit menular lainnya pada tahun 2030 [3]. Penyakit menular masih menjadi faktor utama dalam angka kematian global, karena adanya kemunculan penyakit menular baru (*new emerging disease*), kembalinya penyakit menular yang sudah ada sebelumnya (*re-emerging disease*), dan juga karena kelanjutan penyakit yang sulit diatasi [4]. Kabupaten Karawang telah mengalami perkembangan positif, termasuk dalam hal pendapatan daerah, situasi ekonomi, dan tingkat pendidikan penduduknya [5]. Menurut data dari Badan Pusat Statistik Kabupaten

Karawang pada tahun 2021, terdapat peningkatan jumlah penduduk Kabupaten Karawang sebanyak 2468,58 jiwa. Kenaikan populasi dan ukuran wilayah Kabupaten Karawang tidak terlepas dari aspek kesehatan penduduknya. Saat ini, jumlah kasus penyakit menular semakin meningkat setiap tahun, sehingga mungkin diperlukan upaya untuk mengelompokkan daerah penyebaran penyakit tersebut [6]. Selain itu, variasi dalam geografi dan keragaman sosial juga memiliki potensi untuk memengaruhi cara penyakit menular menyebar.

Penyakit menular memiliki potensi dampak serius pada kesehatan masyarakat jika tidak ditangani dengan baik. Oleh karena itu, penting untuk memiliki pemahaman yang mendalam tentang pola penyebaran dan pengelompokan (klusterisasi) dalam rangka mengidentifikasi wilayah-wilayah di Kabupaten Karawang yang rentan terhadap penyakit menular. Dengan cara ini, upaya pencegahan dan pengendalian dapat didasarkan dengan lebih tepat dan difokuskan pada daerah-daerah yang memerlukan perhatian khusus. Dengan pengetahuan mengenai wilayah yang berisiko tinggi terhadap penyakit menular, langkah-langkah pencegahan dan pemantauan dapat ditingkatkan untuk mencegah penyebaran penyakit yang lebih lanjut di Kabupaten Karawang.

2. TINJAUAN PUSTAKA

Setelah mengkaji beberapa penelitian, ditemukan bahwa beberapa di antaranya berkaitan dengan penelitian yang sedang penulis jalani.

Berdasarkan penelitian yang dilakukan oleh [7] berjudul *Data Mining Clustering Menggunakan*

Algoritma *K-means* Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen dengan tujuan untuk memprediksi tingkat pencapaian tridarma seorang dosen dengan cara menilai koreksi dokumen yang diunggah oleh dosen tersebut. Hasil ini menunjukkan bahwa terdapat 53.33% perbedaan dalam tingkat akurasi prediksi yang diberikan oleh algoritma *K-means* terhadap 15 data.

Penelitian lainnya yang dilakukan oleh [8] berjudul Pemetaan Siswa Kelas Unggulan Menggunakan Algoritma *K-means Clustering* dengan tujuan untuk mengelompokkan data siswa kelas unggulan dengan tujuan membuat proses pembelajaran di sekolah menjadi lebih efisien. Hasil penelitian ini menunjukkan *Cluster* 1 terdiri dari 10 orang siswa yang memiliki kemahiran dalam pemrograman Android, *Cluster* 2 memiliki 62 orang siswa yang ahli dalam pemrograman web, dan *Cluster* 3 terdiri dari 48 orang siswa yang memiliki kompetensi dalam pemrograman desktop.

Penelitian dengan menggunakan algoritma yang sama juga dilakukan oleh [9] berjudul Analisis *Clustering* Kasus Covid 19 Di Indonesia Menggunakan Algoritma *K-means* dengan tujuan untuk menganalisis pola penyebaran Covid-19 dengan mengelompokkan kasus Covid-19 di Indonesia. Hasil penelitian ini menunjukkan tiga *cluster* terbaik dengan nilai DBI terendah, yakni sekitar 0,47. *Cluster* pertama terdiri dari 30 provinsi, sementara *Cluster* kedua dan ketiga masing-masing terdiri dari 2 provinsi.

2.1. Penyakit Menular

Penyakit menular atau sering disebut juga sebagai penyakit infeksi, adalah penyakit yang disebabkan oleh agen biologis, seperti virus, bakteri, atau parasit, bukan oleh faktor fisik, seperti luka bakar atau cedera, atau eksposur terhadap bahan kimia beracun. Penyakit-penyakit ini dapat menular atau menyebar ke individu lain melalui berbagai saluran, seperti melalui makanan dan minuman yang sanitasinya kurang memadai, penggunaan jarum suntik, atau melalui transfusi darah [10].

2.2. KDD (Knowledge Discovery in Database)

Knowledge Discovery in Database (KDD) adalah proses lengkap untuk mengeksplorasi dan mengenali pola dalam data, di mana pola yang ditemukan harus sah, inovatif, bermanfaat, dan dapat dimengerti [11]. *Knowledge Discovery in Database* (KDD) memusatkan perhatian pada langkah-langkah penemuan pengetahuan yang dapat diterapkan pada basis data. Proses ini melibatkan pengenalan data-data yang valid, baru, memiliki potensi manfaat, dan akhirnya membentuk pola yang dapat dimengerti.

2.3. Clustering

Clustering atau klasterisasi adalah aktivitas yang dilakukan untuk mengelompokkan data berdasarkan tingkat kemiripan antara satu data dengan data lainnya ke dalam klaster atau kelompok. Tujuan utamanya

adalah agar data dalam satu klaster memiliki tingkat kemiripan maksimum satu sama lain, sementara tingkat kemiripan antara klaster berbeda sekecil mungkin [12]. Dalam proses *clustering*, algoritma berupaya mengelompokkan data ke dalam kelompok-kelompok sedemikian rupa sehingga data dalam satu kelompok memiliki tingkat kesamaan yang tinggi, sementara data di antara kelompok memiliki perbedaan yang besar. Hal ini memungkinkan pengidentifikasian pola, tren, atau kelompok data yang mungkin tidak terlihat secara langsung.

2.4. Algoritma K-means

K-means adalah algoritma yang digunakan untuk pelatihan tanpa pengawasan (*unsupervised learning*) yang pertama kali diperkenalkan oleh Stuart Lloyd pada tahun 1984. Ini adalah algoritma *clustering* yang sangat umum digunakan. Algoritma ini relatif mudah untuk diimplementasikan dan dieksekusi, memiliki kecepatan yang baik, dapat disesuaikan dengan mudah, dan banyak digunakan dalam berbagai konteks [13].

Adapun langkah-langkah *clustering* dengan algoritma *K-means* dapat dijabarkan sebagai berikut:

1. Menentukan jumlah *cluster* (k).
2. Menentukan pusat *cluster* atau yang biasa disebut centroid secara acak.
3. Menghitung jarak pada setiap data terhadap masing-masing pusat. Rumus untuk menghitung jarak menggunakan rumus jarak Euclidean (*Euclidean Distance*) dengan persamaan (1) yaitu:

$$d(a, b) = \sqrt{\sum_{k=1}^n (x_{ak} - x_{bk})^2} \quad (1)$$

Keterangan:

$d(a, b)$ = Distance (jarak) antara objek a dan b

n = Jumlah atribut

x_{ak} = Nilai pusat dari objek a pada dimensi k

x_{bk} = Nilai pusat dari objek b pada dimensi k

4. Mengelompokkan data yang didasarkan pada jarak terdekat ke titik pusat.
5. Menentukan nilai pusat yang baru dengan menghitung rata-rata dari *cluster* yang bersangkutan, dengan persamaan (2) yaitu:

$$C_k = \frac{1}{n_k} \sum d_a \quad (2)$$

Keterangan:

n_k = Jumlah data pada *cluster* k

d_a = Nilai jarak yang masuk dalam masing-masing *cluster*

2.5. SSE (Sum of square error)

SSE (*Sum of square error*) adalah jumlah dari jarak antara setiap data dengan pusat clusternya. Semakin kecil nilai SSE, semakin seragam data dalam setiap *cluster*, dan semakin baik hasil *clustering*nya [14].

$$SSE = \sum_{k=1}^K \sum_{i=1}^n s_k \|X_i - C_k\| \quad (3)$$

Dengan:

K = Total *cluster*

X_i = Data ke- i

C_k = Awal *cluster*

n = Total data

2.6. Davies Bouldin Index

DBI (*Davies bouldin index*) adalah parameter yang digunakan untuk mengukur mutu pengelompokan (*clustering*) dalam analisis data. DBI pertama kali dikemukakan oleh David L. Davies dan Donald W. Bouldin pada tahun 1979. Evaluasi menggunakan DBI merupakan skema evaluasi cluster internal, yang mempertimbangkan kuantitas dan kedekatan antara data dalam hasil pengelompokan untuk menentukan sejauh mana hasil *clustering* yang baik atau tidak [15].

2.7. R Studio

R Studio adalah platform sumber terbuka yang menyajikan lingkungan pengembangan terpadu (IDE) untuk bahasa pemrograman R. Ini juga merupakan tools profesional yang mendukung kolaborasi tim ilmu data dalam mengembangkan dan berbagi pekerjaan mereka [16]. R Studio adalah salah satu tools yang banyak digunakan oleh ilmuwan data, analis data, dan pengembang tools untuk melakukan berbagai tugas seperti analisis data, pemrosesan statistik, visualisasi data, dan pekerjaan terkait dengan bahasa pemrograman R. R Studio dikenal sebagai alat yang sangat berguna dan populer di kalangan komunitas analisis data dan ilmuwan data.

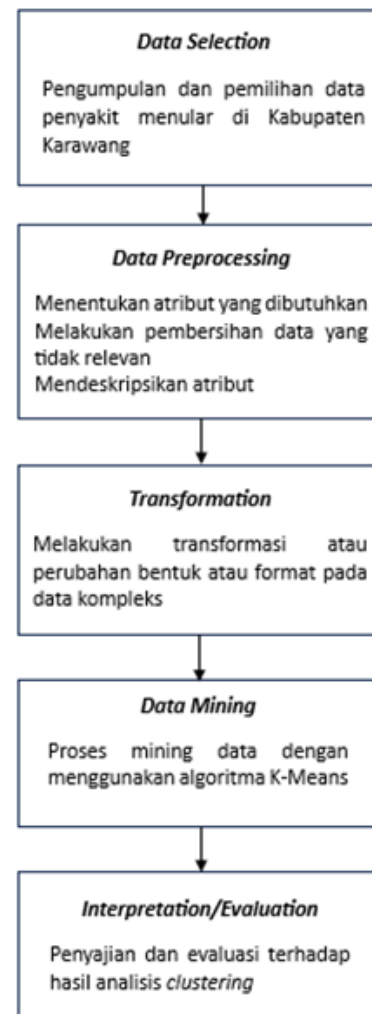
2.8. QGIS

QGIS (*Quantum Geographic Information System*) adalah salah satu tools Sistem Informasi Geografis (SIG) yang berbasis open source. QGIS menyediakan fungsi dan fitur umum yang mudah digunakan oleh penggunanya. QGIS dapat dijalankan pada sistem operasi Linux (Ubuntu), Unix, Mac OS, Windows dan Android, serta mendukung banyak format dan fungsionalitas pengolahan data vektor, raster, dan database [17].

3. METODE PENELITIAN

Metode yang digunakan dalam penelitian ini adalah Knowledge Discovery in Database (KDD). Informasi tersebut terdapat dalam basis data yang besar, di mana manfaat potensial dari data ini sebelumnya belum diketahui.

Dalam penelitian ini pun penulis membuat sebuah rancangan penelitian yang berisi serangkaian kegiatan yang akan dilakukan untuk menyelesaikan penelitian. Proses yang dilakukan yaitu melalui tahapan:



Gambar 1. Alur Penelitian

Berdasarkan gambar diatas alur penelitian yang akan dilakukan adalah sebagai berikut:

a. Data Selection

Data yang diambil merupakan data jumlah kasus penyakit menular berdasarkan jenis-jenis penyakit menular di Kabupaten Karawang pada rentang waktu 2020-2023.

b. Data Preprocessing

Tahap ini melibatkan proses pembersihan data yang menjadi fokus penelitian. Pada proses ini dilakukan penghapusan data yang tidak diperlukan. Penghapusan data juga dilakukan jika terdapat data yang tidak lengkap, seperti atribut kolom yang bernilai *null* (kosong). Penghapusan data kosong dilakukan karena dapat mengakibatkan ketidakakuratan analisis dan membuat informasi menjadi tidak relevan

c. Data Transformation

Pada proses ini akan mengubah nilai nominal kedalam bentuk numerik. Transformasi data akan mencakup penskalaan data untuk menghindari ketidakakuratan analisis yang mungkin terjadi jika rentang jarak data terlalu jauh.

d. *Data Mining*

Pada tahap data *mining*, data akan mengalami proses pengolahan dan pencarian pola atau informasi dengan memanfaatkan teknik *Clustering* menggunakan algoritma *K-means*. Ini merupakan tahap utama dalam pengelolaan data, di mana proses penggalian data dilakukan dengan teknik klasterisasi menggunakan algoritma *K-means*, dan penentuan jumlah *cluster* yang optimal.

e. *Evaluation*

Pada tahap akhir, dilakukan evaluasi pada hasil *cluster* untuk mengetahui seberapa optimal atau seberapa baik suatu data ditempatkan sesuai pada *clusternya*. Evaluasi menggunakan SSE yang diperkuat dengan *Davies bouldin index*. Setelah itu pada hasil *final* akan dilakukan visualisasi peta klasterisasi menggunakan *tools Quantum Geographic Information System (QGIS)* agar lebih mudah dipahami.

4. HASIL DAN PEMBAHASAN

Hasil penelitian ini digunakan untuk melakukan klasterisasi terhadap daerah-daerah yang rawan terhadap penyakit menular di Kabupaten Karawang dari tahun 2020 hingga 2023 dengan menggunakan algoritma *K-means*. Klasterisasi ini kemudian divisualisasikan dengan menggunakan *Quantum Geographic Information System (QGIS)*.

4.1. Pengumpulan Data

Pada proses ini dilakukanlah pengumpulan data-data yang diperlukan menggunakan metode yaitu observasi. Data ini berasal dari Dinas Kesehatan Kabupaten Karawang, yang terdiri dari 720 baris, dan 7 kolom yang mencatat jumlah korban penyakit menular dari tahun 2020 hingga 2023.

Tabel 1. Dataset Awal

Id	Nama Kabupaten	Nama Kecamatan	Bentuk Penyakit	Jumlah Korban	Satuan	Tahun
1	Kabupaten Karawang	Pangkalan	HIV/AIDS	0	Orang	2020
2	Kabupaten Karawang	Pangkalan	IMS	0	Orang	2020
3	Kabupaten Karawang	Pangkalan	DBD	8	Orang	2020
4	Kabupaten Karawang	Pangkalan	Diare	1092	Orang	2020
5	Kabupaten Karawang	Pangkalan	TB	37	Orang	2020
6	Kabupaten Karawang	Pangkalan	Malaria	0	Orang	2020
7	Kabupaten Karawang	Tegalwaru	HIV/AIDS	1	Orang	2020
8	Kabupaten Karawang	Tegalwaru	IMS	0	Orang	2020
9	Kabupaten Karawang	Tegalwaru	DBD	8	Orang	2020
10	Kabupaten Karawang	Tegalwaru	Diare	1092	Orang	2020
11	Kabupaten Karawang	Tegalwaru	TB	37	Orang	2020
12	Kabupaten Karawang	Tegalwaru	Malaria	0	Orang	2020
13	Kabupaten Karawang	Ciampel	HIV/AIDS	12	Orang	2020
14	Kabupaten Karawang	Ciampel	IMS	0	Orang	2020
15	Kabupaten Karawang	Ciampel	DBD	16	Orang	2020
16	Kabupaten Karawang	Ciampel	Diare	1081	Orang	2020
17	Kabupaten Karawang	Ciampel	TB	37	Orang	2020
18	Kabupaten Karawang	Ciampel	Malaria	0	Orang	2020
19	Kabupaten Karawang	Telukjambe Timur	HIV/AIDS	21	Orang	2020
20	Kabupaten Karawang	Telukjambe Timur	IMS	0	Orang	2020
21	Kabupaten Karawang	Telukjambe Timur	DBD	200	Orang	2020
22	Kabupaten Karawang	Telukjambe Timur	Diare	3181	Orang	2020
23	Kabupaten Karawang	Telukjambe Timur	TB	1110	Orang	2020
24	Kabupaten Karawang	Telukjambe Timur	Malaria	0	Orang	2020
25	Kabupaten Karawang	Telukjambe Barat	HIV/AIDS	5	Orang	2020

Setelah proses pengumpulan data awal selesai, data dianalisis atau dideskripsikan untuk menjelaskan format dan atribut yang digunakan.

4.2. Data Preprocessing

Langkah pertama dalam proses ini adalah menghapus atribut data yang dianggap tidak penting. Atribut data dianggap tidak relevan dengan penelitian karena tidak memengaruhi proses analisis data. Misalnya, atribut yang hanya berisi 'id', 'kabupaten karawang', 'orang' dan 'satuan' tidak diperlukan. Selain itu, karena penelitian menggunakan rentang waktu 2020 hingga 2023, maka jumlah korban di

gabungkan pada kecamatan yang sama dalam rentang waktu tersebut.

Tabel 2. Dataset Setelah Penghapusan Data

Nama Kecamatan	Bentuk Penyakit	Jumlah Korban
Pangkalan	HIV/AIDS	0
Pangkalan	IMS	0
Pangkalan	DBD	8
Pangkalan	Diare	1092
Pangkalan	TB	37
Pangkalan	Malaria	0
Tegalwaru	HIV/AIDS	1
Tegalwaru	IMS	0

Nama Kecamatan	Bentuk Penyakit	Jumlah Korban
Tegalwaru	DBD	8
Tegalwaru	Diare	1092
Tegalwaru	TB	37
Tegalwaru	Malaria	0
Ciampel	HIV/AIDS	12
Ciampel	IMS	0
Ciampel	DBD	16
Ciampel	Diare	1081
Ciampel	TB	37
Ciampel	Malaria	0
Telukjambe Timur	HIV/AIDS	21
Telukjambe Timur	IMS	0
Telukjambe Timur	DBD	200
Telukjambe Timur	Diare	3181
Telukjambe Timur	TB	1110

Nama Kecamatan	Bentuk Penyakit	Jumlah Korban
Telukjambe Timur	Malaria	0
Telukjambe Barat	HIV/AIDS	5

Proses ini melibatkan perubahan struktur tabel dengan mengubah kolom yang memiliki atribut bentuk penyakit menjadi 6 kolom baru dengan label yang menggambarkan jenis penyakit yang sebelumnya tersedia dalam kolom bentuk penyakit. Selain itu, dilakukan penambahan 1 kolom baru dengan label "total korban" yang mencakup total jumlah korban dari semua jenis penyakit. Atribut yang sebelumnya bernama "jumlah korban" kemudian digabungkan dalam rentang waktu 2020-2023 dan diintegrasikan menjadi isi dari atribut sesuai dengan kategori penyakit yang bersangkutan.

Tabel 3. Dataset Setelah Diintegrasikan

Nama Kecamatan	HIV/AIDS	IMS	DBD	Diare	TB	Malaria	Total Korban
Pangkalan	677	0	42	1689	182	0	2590
Tegalwaru	886	0	32	1336	138	0	2392
Ciampel	965	0	112	1657	162	0	2896
Telukjambe Timur	1657	3	846	4647	2132	15	9300
Telukjambe Barat	1371	0	92	4086	275	0	5824
Klari	6594	4	483	7850	1147	1	16079
Cikampek	2010	12	352	7767	842	0	10983
Purwasari	1213	0	235	3185	156	1	4790
Tirtamulya	695	0	75	3369	80	0	4219
Jatisari	2035	0	54	2822	752	0	5663
Banyusari	904	0	41	2358	357	0	3660
Kotabaru	1809	4	471	5158	1416	0	8858
Cilamaya Wetan	1124	0	20	3291	664	1	5100
Cilamaya Kulon	840	0	51	3436	615	0	4942
Lemahabang	1013	1	64	3088	248	0	4414
Telagasari	954	0	106	3920	219	1	5200
Majalaya	624	0	143	1745	157	0	2669
Karawang Timur	2020	3	440	4861	884	1	8209
Karawang Barat	3333	2	630	9951	2001	1	15918
Rawamerta	1173	0	71	3153	462	0	4859
Tempuran	1115	0	42	2890	467	0	4514
Kutawaluya	1076	0	15	3355	347	0	4793
Rengasdengklok	2471	0	55	4458	969	3	7956
Jayakarta	1204	0	17	4117	648	0	5986
Pedes	1105	0	18	5345	257	0	6725
Cilebar	474	0	19	1760	395	0	2648
Cibuaya	373	0	13	2623	319	0	3328
Tirtajaya	1021	0	13	4439	148	0	5621
Batujaya	1115	0	14	3311	376	0	4816
Pakisjaya	409	0	5	2396	136	0	2946

Setelah data diintegrasikan, langkah selanjutnya adalah melakukan pengecekan terhadap nilai yang kosong atau *missing value* menggunakan tools R Studio

```
> is.na(tabelfix)
```

	HIV_AIDS	IMS	DBD	Diare	TB	Malaria	Total.Korban
Pangkalan	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Tegalwaru	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Ciampel	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Telukjambe Timur	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Telukjambe Barat	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Klari	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Cikampek	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Purwasari	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Tirtamulya	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Jatisari	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Banyusari	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Kotabaru	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Cilamaya wetan	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Cilamaya kulon	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Lemahabang	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Telagasari	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Majalaya	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Karawang Timur	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Karawang Barat	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Rawamerta	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Tempuran	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Kutawaluya	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Rengasdengklok	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Jayakarta	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Pedes	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Cilebar	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Cibuaya	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Tirtajaya	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Batujaya	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Pakisjaya	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Gambar 2. Pengecekan Missing Value

Apabila terdapat data yang tidak lengkap, seperti atribut kolom yang bernilai *null*, maka data akan dibuang. Namun, pada Gambar 2, tidak ada nilai TRUE yang menunjukkan bahwa data tidak memiliki nilai kosong. Setelah tahap tersebut, dilakukan pemeriksaan lanjutan, yaitu pemeriksaan terhadap redundansi data, apabila terdapat indikasi bahwa ada duplikat data. Karena data pada kolom "nama kecamatan" adalah data kunci, pengecekan dilakukan hanya pada kolom tersebut. Kolom lainnya tidak dievaluasi karena bukan merupakan data kunci, sehingga tidak menjadi masalah jika terdapat nilai atau angka yang sama di dalamnya.

Tabel 4. Hasil Pengecekan Redudansi Data

Nama Kecamatan	HIV/AIDS	IMS
Pangkalan	677	0
Tegalwaru	886	0
Ciampel	965	0
Telukjambe Timur	1657	3
Telukjambe Barat	1371	0
Klari	6594	4
Cikampek	2010	12
Purwasari	1213	0
Tirtamulya	695	0
Jatisari	2035	0
Banyusari	904	0
Kotabaru	1809	4
Cilamaya Wetan	1124	0
Cilamaya Kulon	840	0
Lemahabang	1013	1
Telagasari	954	0
Majalaya	624	0
Karawang Timur	2020	3
Karawang Barat	3333	2
Rawamerta	1173	0
Tempuran	1115	0
Kutawaluya	1076	0
Rengasdengklok	2471	0
Jayakarta	1204	0
Pedes	1105	0
Cilebar	474	0
Cibuaya	373	0

Nama Kecamatan	HIV/AIDS	IMS
Tirtajaya	1021	0
Batujaya	1115	0
Pakisjaya	409	0

Kolom IMS dijadikan contoh untuk menunjukkan bahwa jika terdapat data yang duplikat, tabel akan menunjukkan warna kuning. Hasil pemeriksaan menunjukkan bahwa pada kolom nama kecamatan tidak ditemukan data yang duplikat.

4.3. Data Transformation

Transformasi data merupakan langkah penting dalam mempersiapkan data untuk proses data mining, di mana data diubah ke dalam format yang memungkinkan untuk dilanjutkan ke tahap selanjutnya. Salah satu proses transformasi data yang akan diterapkan adalah penggunaan fungsi penskalaan (*scale*), di mana nilai dari setiap jumlah penyakit menular pada setiap kecamatan akan disesuaikan agar rentang nilainya tidak terlalu jauh.

	HIV/AIDS	IMS	DBD	Diare	TB	Malaria	Total.Korban
1	-0.62432037	-0.39812552	-0.51567172	-1.08434191	-0.72790978	-0.28992753	-0.96435315
2	-0.44598375	-0.39812552	-0.56239522	-1.26548464	-0.81152660	-0.28992753	-1.02152254
3	-0.37857422	-0.39812552	-0.18860720	-1.10076278	-0.76591743	-0.28992753	-0.87600047
4	0.21189917	0.83743644	3.24089786	0.43356229	2.97783577	5.14621361	0.97305368
5	-0.03214041	-0.39812552	-0.28205421	0.14568391	-0.55117422	-0.28992753	-0.03058658
6	4.2456841	1.24929043	1.54483473	2.07718878	1.10595917	0.07248188	2.93038318
7	0.51310886	4.54412235	0.93275685	2.03459715	0.52634256	-0.28992753	1.45899343
8	-0.16695948	-0.39812552	0.38609187	-0.31666622	-0.77731972	0.07248188	-0.32913780
9	-0.60896123	-0.39812552	-0.36148416	-0.22224622	-0.92174878	-0.28992753	-0.49400506
10	0.53444099	-0.39812552	-0.45960352	-0.50294047	0.35530815	-0.28992753	-0.07707280
11	-0.43062462	-0.39812552	-0.52034407	-0.74104309	-0.39534287	-0.28992753	-0.65540750
12	0.34159853	1.24929043	1.48876653	0.69578307	1.61716202	-0.28992753	0.84543314
13	-0.24290187	-0.39812552	-0.61846342	-0.26227209	0.18807450	0.07248188	-0.23963018
14	-0.48523487	-0.39812552	-0.47362057	-0.18786502	0.09495577	-0.28992753	-0.28525019
15	-0.33761653	0.01372847	-0.41288001	-0.36644198	-0.60248454	-0.28992753	-0.43770188
16	-0.38796036	-0.39812552	-0.21664131	0.06050064	-0.65759563	0.07248188	-0.21075676
17	-0.66954449	-0.39812552	-0.04376435	-1.05560539	-0.77541934	-0.28992753	-0.94154315
18	0.52164171	0.83743644	1.34392367	0.54337686	0.60615862	0.07248188	0.65804461
19	1.64200524	0.42558246	2.23167021	3.15532155	2.72888568	0.07248188	2.88389696
20	-0.20109089	-0.39812552	-0.38017356	-0.33308709	-0.19580273	-0.28992753	-0.30921514
21	-0.25058143	-0.39812552	-0.51567172	-0.46804612	-0.18630081	-0.28992753	-0.40882845
22	-0.28385956	-0.39812552	-0.64182518	-0.22943035	-0.41434669	-0.28992753	-0.32827160
23	0.90647336	-0.39812552	-0.45493117	0.33657653	0.76769112	0.79730070	0.58499484
24	-0.17463905	-0.39812552	-0.63248047	0.16159163	0.15766839	-0.28992753	0.01618837

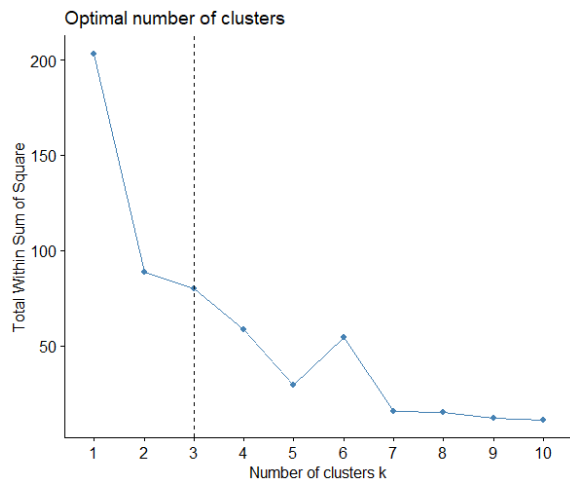
Gambar 3. Hasil Transformasi Data

Pada gambar 3 menampilkan hasil dari proses penskalaan. Hal ini dilakukan untuk mencegah adanya rentang nilai yang terlalu besar yang dapat memengaruhi kualitas dari klaster yang dihasilkan.

4.4. Data Mining

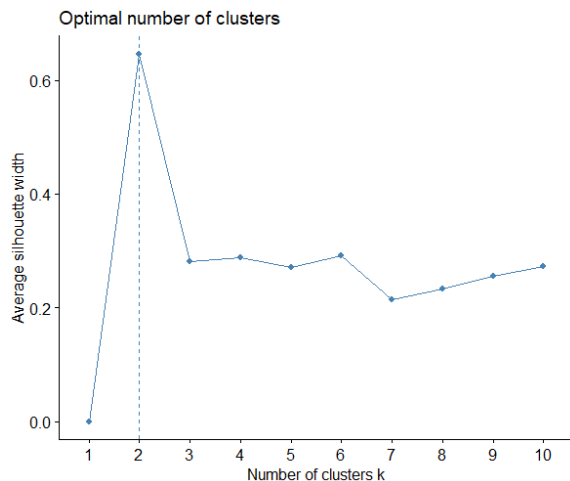
Dalam proses data mining, algoritma yang diterapkan adalah K-means. Algoritma ini memungkinkan penentuan jumlah *cluster* yang optimal menggunakan metode *Elbow*, yang didasarkan pada perhitungan *Sum of Squares*.

Metode *Elbow* mengacu pada nilai *k* yang rendah dan memiliki nilai *witness* yang kecil. Pada metode ini, titik yang menunjukkan *cluster* optimal akan membentuk siku dalam sebuah grafik



Gambar 4. Grafik Metode *Elbow*

Analisis grafik *Elbow* mengindikasikan bahwa titik di mana siku mulai membentuk garis lurus dan stabil berada pada $C=3$. Oleh karena itu, menurut metode *Elbow* jumlah *cluster* optimal adalah 3 *cluster*. Selain menggunakan metode *Elbow*, juga dilakukan pengecekan menggunakan *Silhouette Coefficient* sebagai pertimbangan penentuan jumlah *cluster*.



Gambar 5. *Silhouette Coefficient*

Berdasarkan grafik *Silhouette Coefficient* dalam Gambar 5, tampak bahwa titik optimal terletak pada $C=2$. Meskipun metode *Elbow* menunjukkan jumlah *cluster* optimal pada $C=3$, grafik *Silhouette Coefficient* menunjukkan bahwa *cluster* optimal berada pada titik $C=2$. Oleh karena itu, dilakukan proses *clustering* dengan jumlah *cluster* $C=2$ dan $C=3$ pada data.

K-means clustering with 2 clusters of sizes 6, 24

Cluster means:							
	HIV.AIDS	IMS	DBD	Diare	TB	Malaria	Total.Korban
1	1.2758037	1.5238598	1.7971416	1.4899716	1.593724	0.7973007	1.6249675
2	-0.3189509	-0.3809649	-0.4492854	-0.3724929	-0.398431	-0.1993252	-0.4062419

clustering vector:

[1] 2 2 2 1 2 1 1 2 2 2 2 1 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2

within cluster sum of squares by cluster:

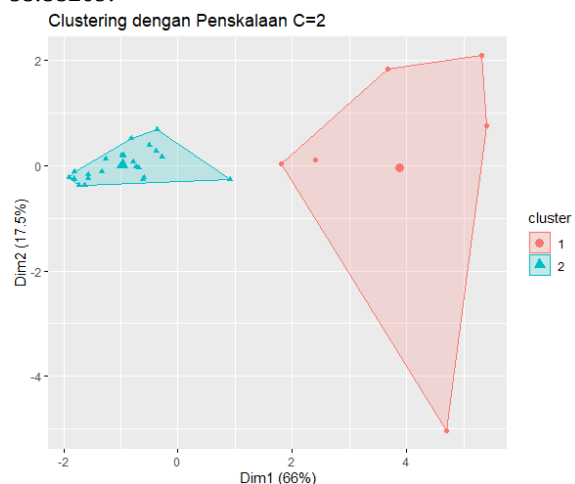
```
[1] 67.76226 21.11983
(between_SS / total_SS = 56.2 %)
```

Available components:

```
[1] "cluster"      "centers"      "tototss"      "withinss"     "tot.withinss" "betweenss"
[9] "ifault"
> #elliptical nlat SSE
> finaldata$scale$tot.withinss
[1] 88.88209
> finaldata$scale$betweenss
[1] 114.1179
```

Gambar 6. Hasil *Clustering* C=2

Pada Gambar 6, menampilkan hasil *clustering* pada data yang telah di-skala dengan jumlah nilai $C=2$, di mana jumlah anggota atau ukuran dari *cluster* 1 adalah 6, sedangkan pada *cluster* 2 terdapat 24 anggota. Dalam hasil *clustering* terlihat bahwa *within* nilai *between*ss yang merupakan perhitungan jarak antara *cluster*, adalah 114.1179. Adapun total jumlah kuadrat dalam model *clustering* ini, yang juga merupakan total dari *within*ss *cluster* atau dikenal sebagai nilai *Sum of square error* (SSE), adalah 88.88209.



Gambar 7. Hasil Visualisasi *Clustering* C=2

Setelah melakukan *clustering* pada jumlah cluster=2, akan dilanjutkan proses *clustering* dengan menggunakan jumlah cluster=3.

```
> print(finaldatascale)
```

K-means clustering with 3 clusters of sizes 6, 15, 9

Cluster means:								
	HIV_AIDS	IMS	DBD	Diare	TB	Malaria	Total	Korban
1	1.2758037	1.5238598	1.7971416	1.48997162	1.5937240	0.7973007		1.6249675
2	-0.1308371	-0.3706686	-0.4418486	-0.07305577	-0.2372311	-0.1449638		-0.1491024
3	-0.6324740	-0.3981255	-0.4616801	-0.87155480	-0.667095	-0.2899275		-0.8348077

clustering vector:

[1] 3 3 3 1 2 1 1 2 3 2 3 1 2 2 2 2 3 1 1 2 2 2 2 2 3 3 2 2 3

within cluster sum of squares by cluster:

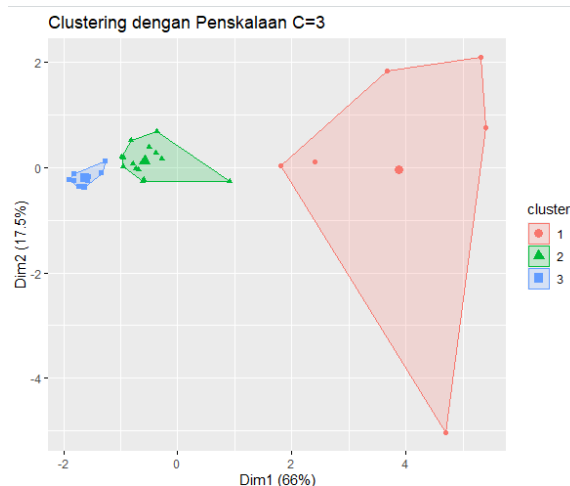
```
[1] 67.762263 10.190290 2.118658
(between_SS / total_SS = 60.6 %)
```

Available components:

```
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
> #melihat nilai SSE
> finaldatascals$tot.withinss
[1] 80.07121
> finaldatascals$betweenss
[1] 122.9288
```

Gambar 8. Hasil *Clustering* C=3

Pada Gambar 8 menampilkan hasil *clustering* pada data yang telah di-skala dengan jumlah nilai $C=3$. Terlihat bahwa jumlah anggota atau ukuran dari *cluster* 1 adalah 6, pada *cluster* 2 terdapat 15 anggota, dan pada *cluster* 3 terdapat 9 anggota. Dalam hasil *clustering* terlihat bahwa nilai *betweeness*, atau perhitungan jarak antara *cluster* adalah 122.9288. Adapun total jumlah kuadrat dalam model *clustering* ini, yang juga merupakan total dari *withinss cluster* atau dikenal sebagai nilai *Sum of square error* (SSE), adalah 80.07121.



Gambar 9. Hasil Visualisasi Clustering C=3

4.5. Evaluation

Setelah melewati tahapan data *mining* dan menyelesaikan proses klasterisasi, langkah berikutnya adalah melakukan evaluasi untuk mengukur hasil atau kualitas dari *cluster* dan model yang telah dibentuk. Terdapat perbedaan dalam penentuan jumlah *cluster* antara hasil grafik *Elbow*, yang menunjukkan jumlah *cluster* optimal pada titik $C=3$, dan grafik *Silhouette Coefficient*, yang menunjukkan jumlah *cluster* optimal pada titik $C=2$. Oleh karena itu, evaluasi dilakukan menggunakan nilai *Sum of Square Error* (SSE) yang didukung oleh indeks *Davies Bouldin Index*.

Tabel 5. Perbandingan Indeks Cluster

Metode	Cluster	SSE/ Tot. Withinss	Betweeness
<i>Silhouette</i>	2	88.88209	114.1179
<i>Elbow</i>	3	80.07121	122.9288

Dalam R Studio, indeks SSE dinyatakan menggunakan fungsi *Tot.Witthniss*. *Cluster* dianggap optimal atau baik jika nilai SSE-nya rendah. Sebaliknya, untuk indeks *betweeness*, *cluster* dianggap optimal jika rasio nilainya lebih tinggi. Berdasarkan tabel 1, jumlah *cluster* dengan nilai SSE yang lebih rendah dan *betweeness* yang lebih tinggi terletak pada titik $C=3$. Untuk memperkuat hasil analisis, evaluasi juga dilakukan menggunakan indeks *Davies Bouldin Index* (DBI).

Tabel 6. Perbandingan DBI

Metode	Jumlah Cluster	DBI
<i>Silhouette</i>	2	0.8816
<i>Elbow</i>	3	0.6507

Dalam *Davies Bouldin Index* (DBI), pengelompokan data dianggap optimal atau semakin baik jika indeksnya lebih rendah atau mendekati angka 0. Berdasarkan informasi dalam Tabel 2, titik optimal untuk *cluster*, baik berdasarkan nilai SSE maupun DBI, berada pada $C=3$.

Langkah berikutnya adalah menampilkan data agar lebih mudah dipahami dan dibaca. Pada tahap data *mining*, hasil *clustering* hanya berupa informasi vektor dan plot visualisasi. Oleh karena itu, dilakukan ekstraksi hasil *clustering* ke dalam bentuk tabel.

Nama.Kecamatan	HIV/AIDS	IMS	DBD	Diare	TB	Malaria	Total.Korban	finaldatascale.cluster
Pangkalan	677	0	42	1689	182	0	2590	3
Tegalwaru	886	0	32	1336	138	0	2392	3
Ciampe	965	0	112	1657	162	0	2896	3
Telukjambe Timur	1657	3	846	4647	2132	15	9300	1
Telukjambe Barat	1371	0	92	4086	275	0	5824	2
Klari	6594	4	483	7850	1147	1	16079	1
Cikampek	2010	12	352	7767	842	0	10983	1
Purwasari	1213	0	235	3185	156	1	4790	2
Tirtamulya	695	0	75	3369	80	0	4219	3
Jatisari	2035	0	54	2822	752	0	5663	2
Banyusari	904	0	41	2358	357	0	3660	3
Kotabaru	1809	4	471	5158	1416	0	8858	1
Cilamaya Wetan	1124	0	20	3291	664	1	5100	2
Cilamaya Kulon	840	0	51	3436	615	0	4942	2
Lemahabang	1013	1	64	3088	248	0	4414	2
Telagasari	954	0	106	3920	219	1	5200	2
Majalaya	624	0	143	1745	157	0	2669	3
Karawang Timur	2020	3	440	4861	884	1	8209	1
Karawang Barat	3333	2	630	9951	2001	1	15918	1
Rawamerta	1173	0	71	3153	462	0	4859	2
Tempuran	1115	0	42	2890	467	0	4514	2
Kutawaluya	1076	0	15	3355	347	0	4793	2
Rengasdengklok	2471	0	55	4458	969	3	7956	2
Jayakarta	1204	0	17	4317	648	0	5986	2

Gambar 10. Hasil Ekstraksi Clustering

Setelah data disajikan dalam format tabel, dilakukan juga pelabelan *cluster* berdasarkan karakteristik dari masing-masing *cluster*, menghasilkan tiga label: tidak rawan, rawan, dan sangat rawan. Karakteristik dari setiap *cluster*, yang merupakan rata-rata dari masing-masing *cluster* yang terbentuk.

cluster	HIV.AIDS	IMS	DBD	Diare	TB	Malaria	Total.Korban
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2904.	4.67	537	6706.	1404.	3	11558.
2	1255.	0.0662	57.8	3660.	440.	0.4	5414.
3	667.	0	53.6	2104.	214	0	3039.

Gambar 11. Karakteristik Cluster

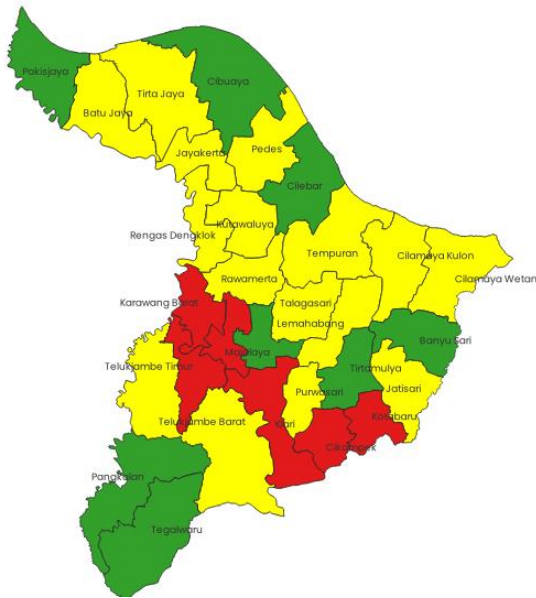
Untuk memudahkan dalam membaca hasil karakteristik *cluster*, maka dibuatlah informasi label untuk setiap *cluster* tercantum dalam Tabel 3.

Tabel 7. Label Cluster

Cluster	Label
C1	Daerah Sangat Rawan
C2	Daerah Rawan
C3	Daerah Tidak Rawan

Agar penelitian ini dapat lebih jelas pada proses akhirnya maka dibentuk sebuah penyajian data dengan melakukan visualisasi peta clusterisasi melalui *tools*

Quantum Geographic Information System (QGIS). Hasil visualisasi tercantum pada Gambar 12.



Gambar 12. Hasil Visualisasi Peta

Setelah dilakukan visualisasi terhadap peta kabupaten karawang. Berikut merupakan keterangan mengenai pewarnaan dalam peta tersebut:

Tabel 8. Keterangan Warna

Keterangan	
	Sangat Rawan
	Rawan
	Tidak Rawan

5. KESIMPULAN DAN SARAN

Berdasarkan penelitian yang telah dilakukan yaitu klasterisasi daerah rawan penyakit menular menggunakan data penyakit menular yang didapat dari Dinas Kesehatan Kabupaten Karawang maka terdapat beberapa hal yang dapat disimpulkan yaitu Algoritma *K-means* dapat diimplementasikan dengan baik untuk melakukan klasterisasi data penyakit menular dengan metode KDD dan software excel, R Studio serta menggunakan bahasa pemrograman R. Hasil dari klasterisasi dengan menggunakan *K-means* mengelompokkan data penyakit menular menjadi 3 klaster, klaster 1 dengan anggota 6 kecamatan yang memiliki kategori daerah sangat rawan, klaster 2 terdiri dari 15 kecamatan yang memiliki kategori daerah rawan, dan klaster 3 yang terdiri dari 9 kecamatan yang memiliki kategori daerah tidak rawan.

Hasil klasterisasi dengan menggunakan algoritma *K-means* dievaluasi dengan menggunakan SSE yang diperkuat dengan *Davies Bouldin Index* dan menghasilkan titik klaster optimal di $C=3$ dengan nilai SSE 80.0712 dan nilai DBI sebesar 0.6507. Pada penelitian selanjutnya disarankan dapat menggunakan metode dan tools yang berbeda atau bahkan dilakukan perbandingan, dan dapat memperluas cakupan data dengan melibatkan rentang waktu yang lebih luas dan menggabungkan variabel yang lebih beragam.

DAFTAR PUSTAKA

- [1] D. N. Yuzar, "Penyakit Menular Dan Wabah Penyakit Covid-19," May 2020.
- [2] I. M. Indra *Et Al.*, *Epidemiologi Penyakit Menular*. Tahta Media Group, 2022. [Online]. Available: <https://www.researchgate.net/publication/361864542>
- [3] T. Y. Aditama, "Penyakit Menular Masalah Dunia, Masalah Kita," Media Indonesia.
- [4] I. Y. Dewi, "Pengembangan Sistem Kewaspadaan Dini Penyakit New-Emerging Dan Re-Emerging, Studi Kasus Pada Penyakit Mers-Cov Dan Ebola Di Kantor Kesehatan Pelabuhan Soekarno-Hatta," *Journal Of Information Systems For Public Health*, Vol. 4, No. 1, Pp. 1–8, 2019.
- [5] I. Krisnawati, I. A. Ubaidi, H. Rais, And R. L. Batu, "Strategi Digital Marketing Dalam Perdagangan Hasil Tani Untuk Meningkatkan Pendapatan Kabupaten Karawang," *Jurnal Ilmiah Ekonomi Global Masa Kini*, Vol. 10, No. 02, Pp. 70–75, 2019, [Online]. Available: www.karawangkab.go.id
- [6] Y. P. Sari, A. Primajaya, And A. S. Y. Irawan, "Implementasi Algoritma K-Means Untuk Clustering Penyebaran Tuberkulosis Di Kabupaten Karawang," *Jurnal Inovtek Polbeng*, Vol. 5, No. 2, Pp. 229–239, 2020.
- [7] R. Muliono And Z. Sembiring, "Data Mining Clustering Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen," *Cess (Journal Of Computer Engineering System And Science)*, Vol. 4, No. 2, Pp. 272–279, 2019.
- [8] J. Hutagalung, Y. H. Syahputra, And Z. P. Tanjung, "Pemetaan Siswa Kelas Unggulan Menggunakan Algoritma K-Means Clustering," *Jurnal Teknik Informatika Dan Sistem Informasi*, Vol. 9, No. 1, Pp. 606–620, 2022.
- [9] T. Hardiani, "Analisis Clustering Kasus Covid 19 Di Indonesia Menggunakan Algoritma K-Means," *Jurnal Nasional Pendidikan Teknik Informatika : Janapat*, Vol. 11, No. 2, Pp. 156–165, 2022.
- [10] S. Hardianti, A. Tenriawaru, And N. Ransi, "Sistem Pakar Diagnosa Penyakit Menular Pada Anak Menggunakan Metode Forward Chaining Dan Backward Chaining," *Jurnal Sains Terapan Teknologi Informasi*, Vol. 13, No. 2, Pp. 111–120, 2021, Doi: 10.46964/Justti.V13i1.607.
- [11] M. Rizki, D. Devrika, F. S. Lubis, And I. H. Umam, "Aplikasi Data Mining Dalam Penentuan Layout Swalayan Dengan Menggunakan Metode Mba," *Jurnal Teknik Industri*, Vol. 5, No. 2, Pp. 130–138, 2019.
- [12] A. Aditya, I. Jovian, And B. N. Sari, "Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama Di Indonesia Tahun 2018/2019," *Jurnal Media*

- Informatika Budidarma*, Vol. 4, No. 1, Pp. 51–58, Jan. 2020, Doi: 10.30865/Mib.V4i1.1784.
- [13] D. Marlina, N. F. Putri, A. Fernando, And A. Ramadhan, “Implementasi Algoritma K-Medoids Dan K-Means Untuk Pengelompokkan Wilayah Sebaran Cacat Pada Anak,” *Jurnal Coreit*, Vol. 4, No. 2, Pp. 64–71, 2018.
- [14] L. P. Refialy, H. Maitimu, And M. S. Pesulima, “Perbaikan Kinerja Clustering K-Means Pada Data Ekonomi Nelayan Dengan Perhitungan Sum Of Square Error (Sse) Dan Optimasi Nilai K Cluster,” *Techno.Com*, Vol. 20, No. 2, Pp. 321–329, 2021.
- [15] D. Jollyta, S. Efendi, M. Zarlis, And H. Mawengkang, “Optimasi Cluster Pada Data Stunting: Teknik Evaluasi Cluster Sum Of Square Error Dan Davies Bouldin Index,” In *Prosiding Seminar Nasional Riset Information Science (Senaris)*, 2019, Pp. 918–926.
- [16] A. Iskandar, *Pengantar Analisis Data Dengan R Studio*. Makassar: Cendekiawan Inovasi Digital Indonesia, 2023.
- [17] N. Kurniati, B. Tampubolon, And Lm. H. Christanto, “Pengaruh Penggunaan Media Sig Dengan Aplikasi Qgis Pada Pembelajaran Geografi Terhadap Hasil Belajar Siswa,” 2020.