Scan
Jurnal Teknologi Informasi dan Komunikasi

ISSN (Print): 1978-0087
ISSN (Online): 2686-6099
Volume 19 Nomor 3, Oktober 2024
Halaman: 142-155

Klasterisasi Wilayah Berdasarkan Penyebaran Penyakit Menular di DKI Jakarta Menggunakan Algoritma *K-Means*

Raisah Nurul Faridah*, Lativa Yulia Taviani, Thalita Syahlani Putri
Fakultas Ilmu Komputer
Universitas Pembangunan Nasional ”Veteran” Jawa Timur, Indonesia

Diterima: Juli, 2024 | Revisi: September, 2024 | Diterbitkan: Oktober 2024

DOI: https://doi.org/10.33005/scan.v19i3.5031

ABTSRAK

DKI Jakarta merupakan wilayah dengan kepadatan penduduk tinggi yang meningkatkan risiko penyebaran penyakit menular seperti HIV/AIDS, malaria, dan tuberkulosis. Penelitian ini bertujuan untuk mengelompokkan wilayah berdasarkan tingkat penyebaran penyakit menggunakan algoritma K-Means. Data yang digunakan berasal dari Dinas Kesehatan DKI Jakarta pada tahun 2015 hingga 2020, mencakup atribut wilayah, tahun, jenis penyakit, dan jumlah kasus. Proses penelitian dilakukan melalui preprocessing, normalisasi data, penerapan algoritma K-Means, dan evaluasi menggunakan SSE dan DBI. Hasil penelitian menunjukkan bahwa klasterisasi dengan tiga klaster memberikan pemetaan distribusi penyakit yang optimal, dengan nilai SSE sebesar 0.779 dan DBI sebesar 0.487. Tiga klaster yang dihasilkan mengelompokkan wilayah menjadi sangat rawan, rawan, dan tidak rawan, dengan wilayah sangat rawan didominasi oleh pusat kota yang memiliki kasus penyakit menular tinggi. Penelitian ini diharapkan dapat memberikan kontribusi dalam penyusunan kebijakan kesehatan berbasis data untuk intervensi yang lebih efektif.

Kata Kunci: Penyakit Menular, DKI Jakarta, *K-Means*, Klasterisasi, *Data Mining*.

Regional Clustering Based on Infectious Disease Spread in DKI Jakarta Using the K-Means Algorithm

ABSTRACT

DKI Jakarta, as a densely populated area, has a heightened risk of spreading infectious diseases such as HIV/AIDS, malaria, and tuberculosis. This study aims to classify regions based on the level of disease spread using the K-Means algorithm. The dataset, sourced from the DKI Jakarta Health Department from 2015 to 2020, includes attributes such as region, year, disease type, and the number of cases. The research process involved data preprocessing, normalization, application of the K-Means algorithm, and evaluation using SSE and DBI metrics. The findings show that clustering with three clusters provides optimal disease distribution mapping, with an SSE value of 0.779 and a DBI value of 0.487. The resulting clusters categorize regions into highly vulnerable, vulnerable, and non-vulnerable areas, with highly vulnerable regions predominantly located in urban centers experiencing high infectious disease cases. This study is expected to contribute to data-driven health policy development for more effective interventions.

Keywords: Infectious Diseases, DKI Jakarta, K-Means, Clustering, Data Mining.

***Corresponding Author:**
Email : 21081010105@student.upnjatim.ac.id
Alamat : Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur 60294

PENDAHULUAN

DKI Jakarta adalah wilayah dengan kepadatan penduduk yang sangat tinggi, sehingga menjadi salah satu daerah yang rentan terhadap penyebaran penyakit menular. Tingginya mobilitas dan kepadatan penduduk menjadikan penyakit seperti *HIV/AIDS*, malaria, tuberkulosis, dan penyakit lainnya sebagai masalah kesehatan yang serius. Berdasarkan data dari Kementerian Kesehatan, meskipun berbagai upaya pencegahan dan pengendalian telah dilakukan, jumlah kasus penyakit menular di Jakarta masih mengalami naik turun yang mengkhawatirkan, terutama di wilayah tertentu yang dipengaruhi oleh kondisi sosial, ekonomi, dan lingkungan yang beragam [1][2][3]. Ketimpangan dalam persebaran penyakit ini menunjukkan perlunya penerapan pendekatan berbasis data untuk memetakan dan menyusun kebijakan yang lebih efektif dalam pengendalian penyakit menular.

Untuk mengatasi masalah ini, penelitian ini menggunakan pendekatan data mining dengan algoritma klasterisasi, yaitu *K-Means*. Metode ini memungkinkan pengelompokan wilayah berdasarkan kesamaan tingkat penyebaran penyakit [4]. Melalui klasterisasi, wilayah dengan pola penyebaran penyakit yang serupa dapat diidentifikasi, sehingga sumber daya kesehatan dapat diarahkan ke area yang paling membutuhkan. Penelitian-penelitian sebelumnya telah membuktikan bahwa metode ini efektif dalam menganalisis distribusi penyakit di berbagai daerah. Evaluasi klaster dilakukan dengan menggunakan metrik *Sum of Squares Error (SSE)* dan *Davies-Bouldin Index (DBI),* yang berguna untuk menentukan jumlah klaster yang paling optimal.

Tujuan dari penelitian ini adalah untuk memberikan pemetaan yang lebih akurat tentang penyebaran penyakit menular di DKI Jakarta dengan mengklasterkan daerah-daerah yang memiliki tingkat penyebaran penyakit yang sebanding. Tujuan lain dari penelitian ini adalah untuk menyediakan model yang dapat digunakan oleh pemerintah daerah dalam mengembangkan kebijakan kesehatan yang lebih tepat sasaran. Penelitian ini diharapkan dapat membantu kebijakan pencegahan penyakit berbasis data dengan memfokuskan pada area yang memerlukan lebih banyak perhatian.

Pola penyebaran penyakit menular di wilayah perkotaan seperti Jakarta sangat kompleks dan dipengaruhi oleh faktor sosial, ekonomi, serta lingkungan. Kondisi ini membutuhkan pendekatan berbasis data untuk menghasilkan pemetaan yang lebih akurat. *Algoritma K-Means*, sebagai metode klasterisasi, memiliki keunggulan dalam mengelompokkan wilayah berdasarkan kesamaan atribut, seperti tingkat penyebaran penyakit, sehingga memungkinkan identifikasi area prioritas untuk dilakukan intervensi [5]. Selain itu, hasil klasterisasi dievaluasi menggunakan metrik *Sum of Squares Error (SSE)* yang merupakan jumlah jarak antar data dengan pusat *clusternya* [6]. Selain itu, *Davies-Bouldin Index (DBI)* juga digunakan untuk memastikan kualitas klaster yang optimal. Penelitian ini diharapkan dapat menghasilkan pemetaan yang lebih presisi untuk mendukung kebijakan kesehatan berbasis data yang efektif dalam mengendalikan penyebaran penyakit menular di DKI Jakarta.

METODE PENELITIAN

Penelitian ini bertujuan untuk menganalisis penyebaran penyakit menular di DKI Jakarta dengan menggunakan pendekatan *data mining*. Untuk analisis data, *algoritma K-Means* dipilih karena dapat mengelompokkan data ke dalam *cluster* berdasarkan kedekatan fitur dan menangani volume data yang besar, dalam hal ini tingkat penyebaran penyakit, relatif cepat [7].

Proses penelitian terdiri dari beberapa tahapan utama, termasuk pemilihan data, pengolahan data, dan penerapan algoritma.

Rancangan Penelitian

Pola penyebaran penyakit menular dianalisis melalui pendekatan "desain kuantitatif" dan "*data mining*". Data yang digunakan dalam penelitian ini berasal dari Dinas Kesehatan DKI Jakarta, yang mencakup berbagai penyakit menular yang terjadi di wilayah Jakarta. Selanjutnya, data diproses untuk mengidentifikasi daerah dengan tingkat penyebaran penyakit yang sama. *Algoritma K-Means* akan digunakan sebagai metode klasterisasi, di mana setiap wilayah akan dikelompokkan berdasarkan jumlah kasus penyakit yang tercatat di masing-masing wilayah. Jumlah kasus, wilayah, jenis penyakit, dan faktor sosial dan ekonomi yang berkontribusi pada penyebaran penyakit adalah beberapa variabel yang digunakan dalam penelitian ini. Setelah data dikumpulkan, langkah berikutnya adalah *preprocessing data*. Data akan dibersihkan, diubah, dan disiapkan untuk digunakan dalam proses data mining melalui tahapan *preprocessing* ini [8].

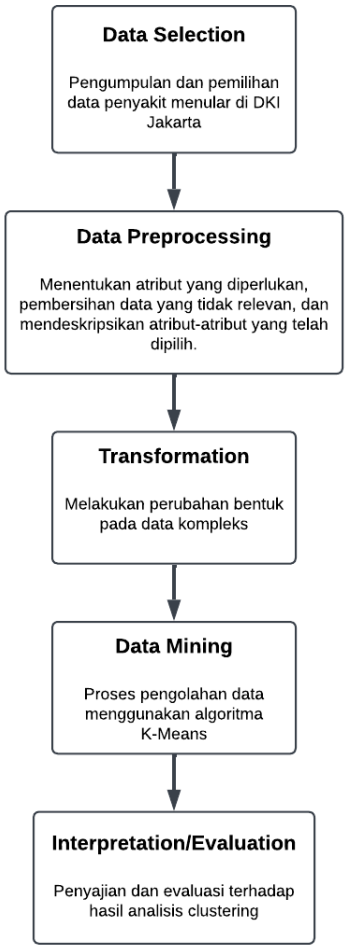
Alur Penelitian

Penelitian ini dirancang secara sistematis melalui tahapan-tahapan agar mencapai hasil yang relevan. Tahapan ini meliputi proses seleksi data, prapemrosesan, transformasi, analisis dengan algoritma *K-Means*, dan evaluasi hasil klasterisasi. Proses ini dilakukan untuk memastikan data yang digunakan bersih, relevan, dan siap untuk menghasilkan klaster yang akurat. Gambaran langkah-langkah utama yang dilakukan selama proses penelitian tertera pada diagram alur yang tertera pada Gambar 1. Alur penelitian ini melibatkan beberapa tahapan penting yang dijelaskan sebagai berikut:

a. *Data Selection* : Data yang digunakan dalam penelitian ini adalah data penyakit menular yang mencakup wilayah DKI Jakarta pada rentang tahun 2015-2020, meliputi data penyakit *HIV/AIDS*, malaria, tuberkulosis, dan penyakit menular lainnya.
b. *Data Preprocessing* : Pada tahap ini, dilakukan pembersihan data yang bertujuan untuk mempersiapkan data agar siap digunakan dalam analisis. Data yang tidak diperlukan akan dihapus, termasuk data yang tidak lengkap, seperti kolom yang kosong. Penghapusan data kosong dilakukan karena bisa mengganggu akurasi analisis dan membuat informasi yang dihasilkan menjadi tidak relevan. Langkah pertama adalah menangani *missing values* (nilai yang hilang), dengan mengisi nilai yang hilang menggunakan nilai yang sesuai, seperti rata-rata atau median. Selain itu, data encoding numerikal juga dilakukan untuk mengubah data kategorikal menjadi format numerik yang dapat diproses oleh *algoritma machine learning*.
c. *Data Transformation* : Setelah data diproses, dilakukan normalisasi untuk memastikan bahwa semua data berada dalam skala yang seragam. Normalisasi

ini penting untuk menghindari perbedaan skala antar atribut yang bisa mempengaruhi akurasi dalam proses klasterisasi. Penskalaan ini memastikan bahwa setiap atribut berkontribusi secara proporsional terhadap hasil klasterisasi.

d. *Data Mining* : Pada tahap *data mining*, data akan diproses untuk mencari pola atau informasi dengan menggunakan teknik *Clustering* melalui algoritma *K-means*. Ini adalah langkah utama dalam pengolahan data, di mana data digali menggunakan teknik klasterisasi dengan *K-means*, serta dilakukan penentuan jumlah *cluster* yang paling optimal.
e. *Evaluation* : Pada tahap akhir, dilakukan evaluasi terhadap hasil klaster untuk menilai sejauh mana data ditempatkan dengan tepat pada cluster yang sesuai. Evaluasi dilakukan menggunakan *SSE (Sum of Squared Errors)* yang diperkuat dengan Davies-Bouldin Index. Setelah itu, hasil akhir divisualisasikan dalam bentuk peta klasterisasi menggunakan pustaka *Python* seperti *Matplotlib, Geopandas*, dan *Contextily* untuk mempermudah interpretasi hasil.



Gambar 1. Alur Penelitian

HASIL DAN PEMBAHASAN

Hasil penelitian ini bertujuan untuk melakukan klasterisasi pada wilayah-wilayah yang rentan terhadap penyakit menular di DKI Jakarta pada periode 2015 hingga 2020 dengan memanfaatkan *algoritma K-means*..

Data Selection

Pada tahap ini, data yang digunakan diperoleh dari *platform Kaggle*, yang mencatat informasi mengenai jumlah kasus penyakit menular di DKI Jakarta selama periode 2015 hingga 2020. Dataset yang diperoleh terdiri dari 216 baris data dan 4 kolom yang mencakup atribut tahun, lokasi, jenis penyakit, dan jumlah kasus penyakit yang bisa dilihat pada Tabel 1. Data ini digunakan untuk menganalisis dan mengelompokkan wilayah yang memiliki angka kejadian penyakit menular yang tinggi, serta melihat pola penyebaran penyakit selama periode waktu tersebut.

Tabel 1
Dataset penyakit menular

No.	Year	Reagion	Disease Name	Number of Cases
1.	2015	Thousand Islands	Malaria	0
2.	2015	South Jakarta	Malaria	6
3.	2015	East Jakarta	Malaria	2
4.	2015	Central Jakarta	Malaria	4
5.	2015	West Jakarta	Malaria	5
6.	2015	North Jakarta	Malaria	3
7.	2015	Thousand Islands	Gastro Entritis	0
8.	2015	South Jakarta	Gastro Entritis	0
9.	2015	East Jakarta	Gastro Entritis	0
10.	2015	Central Jakarta	Gastro Entritis	0
11.	2015	West Jakarta	Gastro Entritis	0
12.	2015	North Jakarta	Gastro Entritis	0
13.	2015	Thousand Islands	Cholera	0
14.	2015	South Jakarta	Cholera	0
15.	2015	East Jakarta	Cholera	0
16.	2015	Central Jakarta	Cholera	0
17.	2015	West Jakarta	Cholera	0
18.	2015	North Jakarta	Cholera	0
19.	2015	Thousand Islands	Leprosy	3
20.	2015	South Jakarta	Leprosy	69
21.	2015	East Jakarta	Leprosy	81
22.	2015	Central Jakarta	Leprosy	12
23.	2015	West Jakarta	Leprosy	89
24.	2015	North Jakarta	Leprosy	63
25.	2015	Thousand Islands	TBC	10

Sumber: Data Diolah

Tabel 2
Nilai yang hilang setelah penanganan

Numeric Column	Missing Value
Year	0
Region	0
Disease Name	0
Number of Cases	0

Sumber: Data Diolah

Data preprocessing

Pada tahap data preprocessing, beberapa langkah penting dilakukan untuk mempersiapkan dataset yang digunakan dalam penelitian ini agar dapat dianalisis dengan lebih efektif. Proses ini mencakup dua bagian utama, yaitu penanganan missing values dan data encoding numerikal, yang bertujuan untuk membersihkan dan mengubah format data menjadi lebih sesuai untuk analisis klasterisasi.

Pada langkah pertama, dilakukan penanganan terhadap *missing values* (nilai yang hilang) yang terdapat dalam dataset. Nilai yang hilang dapat mempengaruhi hasil analisis, terutama dalam algoritma klasterisasi seperti *K-means* yang memerlukan data lengkap untuk memisahkan data ke dalam klaster-klaster yang tepat. Oleh karena itu, nilai yang hilang pada kolom numerik, seperti jumlah kasus penyakit, diisi dengan menggunakan nilai rata-rata dari kolom tersebut. Pengisian nilai rata-rata ini bertujuan untuk menjaga kelengkapan data tanpa mengubah distribusi data secara signifikan. Dengan demikian, data yang hilang dapat dipulihkan tanpa menyebabkan distorsi pada hasil analisis.

Setelah melakukan pengecekan terhadap *missing values* dalam *dataset*, dapat dipastikan bahwa seluruh nilai yang hilang telah berhasil diatasi. Hasil pengecekan menunjukkan bahwa tidak ada lagi missing values yang tersisa seperti pada Tabel 2, yang berarti dataset sekarang sudah bersih dan tidak akan mempengaruhi hasil klasterisasi.

Setelah langkah pembersihan data selesai, langkah selanjutnya adalah data *encoding numerikal* untuk mengubah data kategorikal menjadi format numerik yang dapat diproses lebih lanjut oleh *algoritma K-means*. Beberapa kolom dalam *dataset*, seperti lokasi dan jenis penyakit, berisi informasi kategorikal yang perlu diubah menjadi format numerik. Pada penelitian ini, dilakukan *Label Encoding*, di mana setiap kategori pada kolom tersebut diberi label numerik yang sesuai. Sebagai contoh, nama lokasi yang semula berupa teks, 12seperti "*South Jakarta", "Central Jakarta",*89 dan seterusnya, diubah menjadi angka, yaitu 063, 1, 2, dan seterusnya yang bisa dilihat pada Tabel 3. Dengan cara ini, data kategorikal dapat diubah menjadi representasi numerik yang siap digunakan dalam proses klasterisasi.

Data Transformation

Transformasi data merupakan langkah penting dalam mempersiapkan dataset agar siap untuk diproses lebih lanjut pada tahap data mining. Pada tahap ini, data diubah menjadi format yang memungkinkan untuk analisis lebih lanjut, sehingga hasil yang diperoleh menjadi lebih akurat dan relevan. Salah satu proses utama dalam transformasi data adalah penskalaan (*scaling*), yang bertujuan untuk menyamakan skala antar atribut

dalam dataset. Pada penelitian ini, salah satu teknik transformasi yang diterapkan adalah penskalaan pada kolom jumlah kasus penyakit menular yang tercatat di setiap kecamatan di DKI Jakarta. Nilai-nilai pada kolom ini memiliki rentang yang cukup besar, dengan beberapa kecamatan memiliki angka kasus yang sangat tinggi, sementara yang lainnya memiliki angka yang rendah atau bahkan nol. Untuk mengatasi masalah perbedaan skala ini, dilakukan normalisasi pada kolom jumlah penyakit. Dapat dilihat pada Tabel 4 yang menunjukkan sampel data hasil setelah dilakukan normalisasi. Proses normalisasi ini bertujuan untuk menyelaraskan rentang nilai agar seluruh data berada pada skala yang seragam, sehingga setiap atribut memiliki bobot yang sama dalam proses klasterisasi.

Tabel 3
Data Setelah *Encoding*

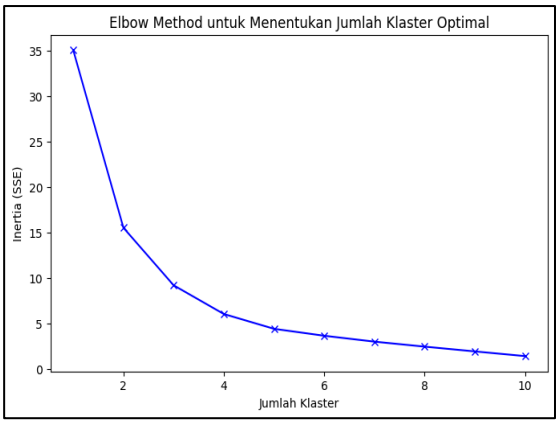
	Region	Reagion Encoded	Disease Name	Disease Encoded
0	Thousand Islands	4	Malaria	4
1	South Jakarta	3	Malaria	4
2	East Jakarta	1	Malaria	4
3	Central Jakarta	0	Malaria	4
4	West Jakarta	5	Malaria	4

Sumber: Data Diolah

Tabel 4
Data Setelah Normalisasi

	Number of Cases	Cases Normalized	Year	Year Normalized
0	0	0.000000	2015	0.0
1	6	0.000456	2015	0.0
2	2	0.000152	2015	0.0
3	4	0.000304	2015	0.0
4	5	0.000380	2015	0.0

Sumber: Data Diolah



Gambar 2. Elbow Method untuk Menentukan Jumlah Klaster Optimal

Data Mining

Dalam proses data mining, algoritma yang diterapkan adalah *K-means*. Algoritma ini memungkinkan pengelompokan data ke dalam beberapa klaster berdasarkan kedekatan data dengan *centroid* klaster. Penentuan jumlah klaster yang optimal dilakukan menggunakan *Elbow Method*, yang didasarkan pada perhitungan *Sum of Squared Errors* (SSE) atau *Within-Cluster Sum of Squares* (WCSS). Metode ini mengukur tingkat kompaknya data dalam klaster; semakin kecil nilai SSE, semakin baik data dikelompokkan di dalam klasternya.

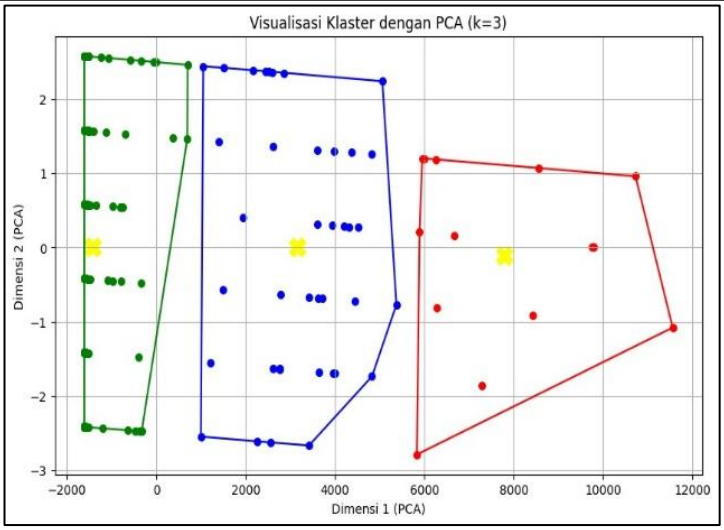
Pada metode *Elbow*, grafik yang menunjukkan hubungan antara jumlah klaster (k) dengan nilai SSE diplot untuk menemukan titik optimal. Titik optimal ini disebut titik siku (*elbow point*), yang merupakan jumlah klaster optimal. Grafik *Elbow Method* di bawah ini menunjukkan penurunan nilai SSE yang signifikan ketika jumlah klaster bertambah, tetapi setelah k = 3, penurunan nilai SSE menjadi lebih kecil. Oleh karena itu, jumlah klaster yang optimal pada dataset ini adalah 3 klaster.

Grafik yang bisa dilihat pada Gambar 2 menunjukkan bahwa dengan 3 klaster, dataset telah dikelompokkan dengan baik. Penambahan jumlah klaster setelah angka ini hanya menghasilkan sedikit penurunan pada nilai SSE, yang menunjukkan bahwa klaster tambahan tidak memberikan manfaat signifikan untuk meningkatkan kualitas pengelompokan. Dengan demikian, 3 klaster dipilih sebagai konfigurasi yang optimal untuk analisis klasterisasi dalam penelitian ini.

Dalam implementasi K-Means, setiap data di dalam dataset dikelompokkan ke dalam salah satu dari tiga klaster berdasarkan kedekatan nilai terhadap centroid masing-masing klaster. Centroid ini dihitung sebagai rata-rata dari data yang termasuk dalam klaster tersebut dan terus diperbarui hingga proses iterasi mencapai konvergensi, yaitu ketika tidak ada lagi perubahan signifikan pada posisi centroid. Proses klasterisasi ini menghasilkan tiga klaster yang memiliki pola distribusi berbeda. Grafik hasil klasterisasi dapat ditempatkan setelah penjelasan ini untuk memberikan gambaran visual mengenai distribusi data berdasarkan klaster yang terbentuk.

Visualisasi hasil klasterisasi dilakukan menggunakan dua pendekatan utama. Pendekatan pertama adalah visualisasi data berdasarkan atribut asli, yaitu Jumlah Kasus Penyakit yang Dinormalisasi dan Tahun yang Dinormalisasi. Grafik ini menunjukkan distribusi data dalam ruang dua dimensi dengan klaster yang diberi warna berbeda untuk mempermudah identifikasi pola penyakit di setiap klaster.

Hasil visualisasi klasterisasi menggunakan algoritma K-Means dengan reduksi dimensi menggunakan Principal Component Analysis (PCA) menunjukkan pola distribusi data yang terbagi menjadi tiga klaster. Pada visualisasi yang bisa dilihat pada Gambar 3, sumbu X merepresentasikan Dimensi 1 (PCA), sedangkan sumbu Y merepresentasikan Dimensi 2 (PCA). Kedua dimensi ini merupakan komponen utama yang dihasilkan dari reduksi dimensi awal, yang mencakup fitur -fitur number_of_cases, year, region_encoded, dan disease_encoded. PCA digunakan untuk menyederhanakan kompleksitas data, namun tetap mempertahankan variabilitas utama yang relevan untuk analisis klasterisasi.



Gambar 3. Hasil Klasterisasi Penyakit Menular di Jakarta

Klaster pertama, yang ditandai dengan warna merah, terletak di sisi kanan visualisasi dengan nilai Dimensi 1 yang relatif tinggi, di atas 5000. Klaster ini mencakup data dengan jumlah kasus yang besar, sehingga dapat diidentifikasi sebagai wilayah dengan risiko atau tingkat kasus yang tinggi. Klaster kedua, yang ditandai dengan warna biru, berada di area tengah, dengan nilai Dimensi 1 berkisar antara 1000 hingga 5000. Klaster ini merepresentasikan wilayah dengan jumlah kasus sedang, yang berada di antara klaster risiko tinggi dan rendah. Klaster ketiga, yang berwarna hijau, terletak di sisi kiri visualisasi, dengan nilai Dimensi 1 yang rendah, kurang dari 1000. Data pada klaster ini dapat diinterpretasikan sebagai wilayah dengan jumlah kasus rendah atau risiko yang lebih kecil.

Setiap klaster memiliki centroid yang ditandai dengan simbol silang kuning. Centroid ini merupakan titik pusat dari setiap klaster yang dihitung berdasarkan rata-rata nilai dari seluruh data pada masing-masing klaster. Posisi centroid membantu dalam memahami distribusi data serta karakteristik utama dari setiap klaster.

Visualisasi ini menunjukkan bahwa klaster-klaster yang terbentuk cukup terpisah satu sama lain, mengindikasikan bahwa hasil klasterisasi cukup baik dalam mengidentifikasi pola distribusi data berdasarkan karakteristik utama. Penggunaan PCA sangat membantu dalam menampilkan hasil klasterisasi dalam ruang dua dimensi, terutama karena data awal memiliki lebih dari dua fitur. Dengan proyeksi ini, pola distribusi klaster menjadi lebih jelas dan mudah untuk diinterpretasikan.

Hasil klasterisasi ini memberikan informasi yang penting untuk analisis lebih lanjut, seperti mengidentifikasi wilayah dengan tingkat kasus yang tinggi, sedang, dan rendah, yang dapat menjadi dasar dalam pengambilan kebijakan berbasis data. Visualisasi yang jelas dan terstruktur ini mempermudah komunikasi hasil kepada pihak yang berkepentingan, termasuk pembuat kebijakan atau tim analisis lanjutan.

Evaluation

Tahap evaluasi bertujuan untuk memastikan kualitas hasil klasterisasi yang telah dilakukan menggunakan algoritma K-Means. Evaluasi ini mencakup dua metrik utama, yaitu *Sum of Squared Errors* (SSE) dan Davies-Bouldin Index (DBI). *Sum of Squared Errors* (SSE) adalah metrik yang mengukur seberapa kompak klaster yang dihasilkan

oleh algoritma. Nilai SSE dihitung sebagai jumlah kuadrat jarak antara setiap titik data dengan centroid klaster yang bersesuaian. Semakin kecil nilai SSE, semakin baik kekompakan data dalam klaster tersebut. *Davies-Bouldin Index* (DBI) adalah metrik yang mengukur sejauh mana klaster yang dihasilkan terpisah satu sama lain. Nilai DBI dihitung berdasarkan rata-rata rasio antara jarak intra-klaster (kekompakan) dan jarak antar-klaster (keterpisahan). Seperti pada Tabel 5, Nilai DBI yang lebih kecil menunjukkan klaster yang lebih kompak dan terpisah dengan baik.

Sum of Squared Errors (SSE) merupakan metrik yang mengukur kekompakan klaster yang dihasilkan. SSE dihitung dengan menjumlahkan kuadrat jarak antara setiap titik data dalam klaster dan centroid klaster tersebut. Nilai SSE yang lebih kecil menunjukkan bahwa data dalam klaster lebih kompak dan lebih dekat dengan centroidnya. Dalam penelitian ini, nilai SSE yang diperoleh adalah 0.779. Hasil ini menunjukkan bahwa data dalam masing-masing klaster memiliki tingkat kekompakan yang baik, dengan jarak antar data dan centroid yang tidak terlalu besar. Meskipun nilai SSE ini terbilang cukup rendah, yang menunjukkan bahwa klaster yang terbentuk cukup efisien, namun masih mungkin untuk penyempurnaan lebih lanjut.

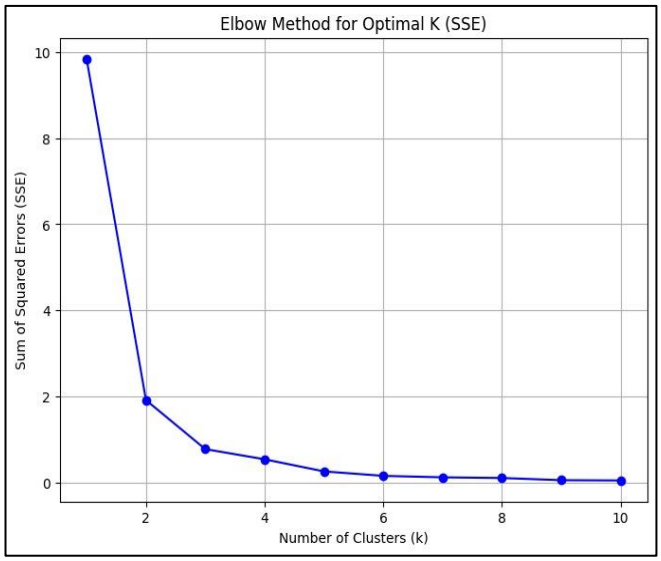
Berdasarkan grafik *Sum of Squared Errors (SSE)* yang bisa dilihat pada Gambar 4, nilai SSE untuk k=3 adalah 0,779. Grafik menunjukkan bahwa nilai SSE mengalami penurunan tajam dari k=1 hingga k=3, dengan pengurangan error yang signifikan di rentang tersebut. Setelah k=3, penurunan SSE menjadi semakin kecil, yang menunjukkan bahwa menambahkan lebih banyak klaster tidak lagi memberikan perbaikan yang signifikan dalam kompaknya data di dalam klaster. Nilai 0,779 pada k=3 menunjukkan bahwa distribusi data dalam tiga klaster sudah cukup baik, di mana jarak data terhadap *centroid* klasternya relatif kecil. Selain itu, titik k=3 dapat dianggap sebagai *elbow point* (titik siku), di mana grafik mulai mendatar.

Sementara itu, *Davies-Bouldin Index (DBI)* digunakan untuk menilai sejauh mana klaster-klaster yang terbentuk terpisah satu sama lain. DBI dihitung berdasarkan perbandingan antara jarak intra-klaster (kekompakan) dan jarak antar-klaster (keterpisahan). Nilai DBI yang lebih rendah menunjukkan bahwa klaster-klaster tersebut lebih terpisah dengan jelas dan memiliki kekompakan yang baik. Dalam penelitian ini, nilai DBI yang diperoleh adalah 0.487, yang menunjukkan bahwa klaster yang terbentuk memiliki tingkat keterpisahan yang baik antar klaster, dan tidak ada tumpang tindih yang signifikan antara klaster-klaster tersebut. Nilai DBI yang lebih rendah dari 1 menandakan bahwa klaster yang dihasilkan efektif dalam memisahkan data ke dalam kelompok yang berbeda berdasarkan pola distribusinya.

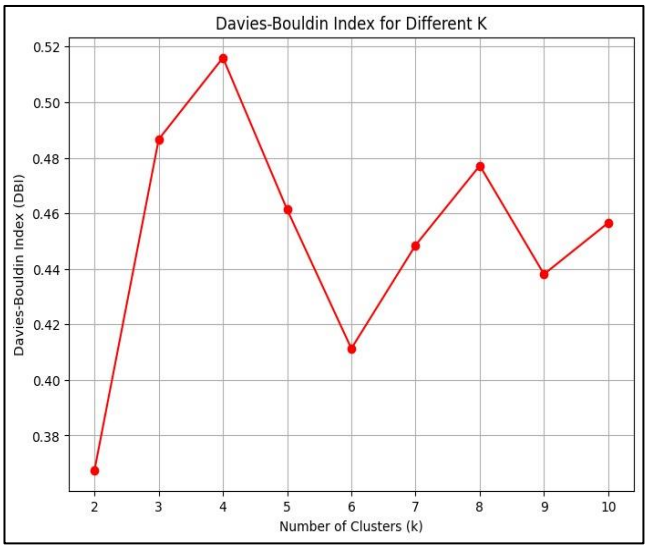
Tabel 5
Hasil Evaluasi Menggunakan DBI dan SSE

Metrik Evaluasi	*Value*
Sum of Squared Errors (SSE)	0.779
Davies-Bouldin Index (DBI)	0.487

Sumber: Data Diolah



Gambar 4. Evaluasi Menggunakan SSE



Gambar 5. Evaluasi Menggunakan DBI

Hasil berdasarkan grafik *Davies-Bouldin Index (DBI)* yang bisa dilihat pada Gambar 5, pemilihan jumlah klaster k=3 dengan nilai DBI sebesar 0,487 menunjukkan bahwa klasterisasi pada jumlah ini memiliki keseimbangan antara kompaknya klaster (intra-cluster) dan terpisahnya antar-klaster (inter-cluster). Meskipun nilai DBI terendah dicapai pada k=2 dengan nilai 0,38, yang menunjukkan klasterisasi terbaik secara matematis, pemilihan k=3 didasarkan pada kebutuhan analisis yang memerlukan pembagian data menjadi lebih banyak kelompok untuk mencerminkan pola yang lebih kompleks dan informatif. Pada k=3, klasterisasi tetap menunjukkan kualitas yang cukup baik, di mana nilai DBI masih berada dalam rentang yang wajar untuk menggambarkan hubungan antar-klaster tanpa terlalu banyak overlap. Dengan demikian, keputusan untuk menggunakan k=3 mempertimbangkan keseimbangan antara interpretasi data yang lebih bermakna dengan kualitas klasterisasi yang tetap terjaga.

Langkah berikutnya adalah menampilkan data agar 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 yang dapat dilihat sampelnya pada Tabel 6.

Tabel 6
Hasil Ekstrasi *Clustering*

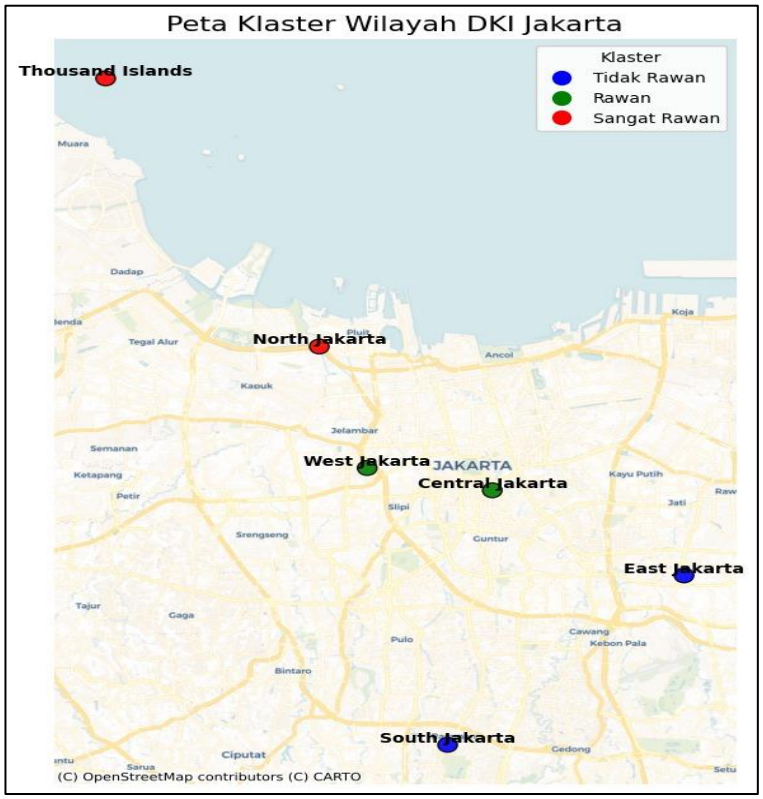
	Year	Region	Disease Name	Number of Cases	Cluster
0	2015	Thousand Island	Malaria	0	2
1	2015	South Jakarta	Malaria	6	1
2	2015	East Jakarta	Malaria	2	1
3	2016	Thousand Island	Malaria	0	2
4	2016	South Jakarta	Malaria	5	1
5	2016	East Jakarta	Malaria	17	0
6	2017	Thousand Island	Malaria	0	2
7	2017	South Jakarta	Malaria	12	1
8	2017	East Jakarta	Malaria	25	0

Sumber: Data Diolah

Tabel 7
Informasi label untuk setiap *cluster*

Cluster	Label
0	Sangat Rawan
1	Rawan
2	Tidak Rawan

Sumber: Data Diolah



Gambar 6. Peta Klaster Wilayah DKI Jakarta

Untuk memudahkan dalam membaca hasil karakteristik cluster, maka dibuatlah informasi label untuk setiap cluster tercantum dalam Tabel 7. Agar penelitian ini lebih mudah dipahami, hasil akhir dari klasterisasi data ditampilkan melalui visualisasi peta yang dihasilkan menggunakan pustaka Python seperti Geopandas dan Matplotlib. Hasil visualisasi ini dapat dilihat pada Gambar 5.

Hasil visualisasi peta menunjukkan distribusi tingkat kerawanan penyakit menular di wilayah DKI Jakarta. Wilayah yang berwarna merah menunjukkan daerah dengan tingkat kerawanan yang sangat tinggi, sementara wilayah yang berwarna hijau menunjukkan daerah dengan tingkat kerawanan yang sedang. Wilayah yang berwarna biru menunjukkan daerah dengan tingkat kerawanan yang rendah. Visualisasi ini memberikan gambaran yang jelas mengenai sebaran tingkat kerawanan penyakit menular di Jakarta, memudahkan pemahaman mengenai daerah-daerah yang membutuhkan perhatian lebih dalam penanganan penyakit.

SIMPULAN

Berdasarkan penelitian yang telah dilakukan mengenai klasterisasi daerah rawan penyakit menular di DKI Jakarta menggunakan algoritma *K-means*, dapat disimpulkan bahwa algoritma ini efektif dalam mengelompokkan data penyakit menular yang meliputi tahun 2015 hingga 2020. Data yang digunakan dalam penelitian ini mencakup jumlah kasus berbagai penyakit menular, seperti Malaria, TBC, DBD, dan lainnya, di beberapa wilayah di Jakarta. Hasil klasterisasi dengan menggunakan *K-means* menghasilkan tiga klaster, yaitu klaster 0 yang mewakili daerah dengan jumlah kasus tinggi, klaster 1 yang mencakup daerah dengan jumlah kasus sedang, dan klaster 2 yang menunjukkan daerah dengan jumlah kasus rendah. Klasterisasi ini memungkinkan identifikasi daerah dengan tingkat kerawanan yang berbeda terhadap penyakit menular. Evaluasi hasil klasterisasi dilakukan dengan menggunakan *Sum of Squared Errors (SSE)* dan *Davies-Bouldin Index (DBI),* yang menunjukkan bahwa klaster optimal terdiri dari tiga klaster, dengan nilai SSE sebesar 0.779 dan DBI sebesar 0.487. Dengan demikian, *algoritma K-means* berhasil mengelompokkan daerah rawan penyakit menular secara efektif. Untuk penelitian selanjutnya, disarankan untuk menggunakan metode klasterisasi lain, seperti *DBSCAN* atau *hierarchical clustering*, serta memperluas cakupan data dengan rentang waktu yang lebih luas dan variabel tambahan seperti kepadatan penduduk dan kondisi lingkungan, guna meningkatkan ketepatan analisis dan kebijakan kesehatan yang lebih tepat.

DAFTAR PUSTAKA

[1]. Kementerian Kesehatan RI. (2023). Profil Kesehatan Indonesia 2023. Retrieved from https://satusehat.kemkes.go.id/data

[2]. Kementerian Kesehatan RI. (2023). Tahun ini, 5 Provinsi dan 9 Kabupaten/Kota Berhasil Eliminasi Malaria. Retrieved from https://sehatnegeriku.kemkes.go.id

[3]. Kementerian Kesehatan RI. (2023). Kasus HIV dan Sifilis Meningkat, Penularan Didominasi Ibu Rumah Tangga. Retrieved from https://sehatnegeriku.kemkes.go.id

[4]. Alfianti, Z. I. (2021). Pengelompokan Wilayah Penyebaran Covid-19 Di Kabupaten Karawang Menggunakan Algoritma K-Means. Jurnal Ilmiah Informatika Komputer, 26(2), 111-122.

[5]. Fitri, E. M., Suryono, R. R., & Wantoro, A. (2023). Klasterisasi Data Penjualan Berdasarkan Wilayah Menggunakan Metode K-Means Pada Pt Xyz. Jurnal Komputasi, 11(2), 157-168.

[6]. Amalia, I. N., Umaidah, Y., & Mayasari, R. (2024). Penerapan Data Mining Untuk Klasterisasi Daerah Rawan Penyakit Menular Di Kabupaten Karawang Dengan Menggunakan Algoritma K-Means. JATI (Jurnal Mahasiswa Teknik Informatika), 8(4), 5582-5591.

[7]. Kholil, K. A., Rahaningsih, N., & Dana, R. D. (2024). Penerapan Data Mining Untuk Clustering Penyakit Diare Menggunakan Algoritma K-Means. JATI (Jurnal Mahasiswa Teknik Informatika), 8(3), 3124-3131.

[8]. Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.