LAPORAN PROYEK DATA MINING

JUDUL

Analisis Distribusi Kasus TBC di
DKI Jakarta dengan Metode Clustering

Oleh:

Kelas: LA08
Kelompok 6
2702217402 - Matthew Nathanael Halim
2702269021 - Yohanes Wenanta
2702223834 - Jonathan Hopi Pranata
27022247310 - Kenneth Andrew Lukita

Kata Pengantar

Penyakit Tuberkulosis atau TBC adalah salah satu masalah kesehatan global yang menjadi salah satu perhatian utama berbagai organisasi kesehatan di dunia termasuk Indonesia. Meskipun sudah terdapat kemajuan dalam pengobatan TBC, penyakit ini masih menjadi tantangan yang cukup serius, terutama di kota-kota besar seperti Jakarta. Dimana tingginya kepadatan penduduk, masalah sanitasi, serta tingkat kesadaran yang masih rendah di kalangan masyarakat turut menjadi faktor penyebab prevalensi TBC yang lumayan tinggi.

Laporan ini bertujuan untuk menganalisis persebaran penyakit TBC di Jakarta dengan menggunakan algoritma Agglomerative Clustering, salah satu metode hierarchical clusterization yang efektif untuk mengidentifikasi pola tersembunyi dalam data kesehatan. Dengan menggunakan data kasus TBC yang tersebar di berbagai wilayah Jakarta, analisis ini berusaha untuk menemukan area-area dengan tingkat prevalensi yang lebih tinggi dan mengidentifikasi faktor-faktor risiko yang mungkin berkontribusi terhadap penyebaran penyakit.

Abstrak

Penelitian ini bertujuan untuk mengelompokan wilayah di DKI Jakarta berdasarkan karakteristik kasus tuberkulosis (TBC) dengan tujuan memberikan dukungan berbasis data kepada pemangku kebijakan dalam menentukan wilayah prioritas penanganan. Data yang digunakan mencakup 267 kelurahan dengan atribut seperti jumlah penduduk, kepadatan, tingkat kemiskinan, serta ketersediaan fasilitas kesehatan. setelah dilakukan *preprocessing* dan *handling missing values*, metode *Agglomerative Hierarchical Clustering* diterapkan untuk membentuk *cluster* wilayah yang memiliki kemiripan dalam indikator terkait TBC. Hasil analisis menunjukan adanya cluster yang membedakan wiilayah dengan jumlah kasus TBC tinggi dan kepadatan penduduk besar dari wilayah dengan beban TBC yang lebih rendah dan infrastruktur kesehatan yang lebih baik.Hal ini memberikan wawasan praktis dalam penyusunan strategi pengendalian TBC yang lebih tepat sasaran, khususnya pada wilayah dengan kepadatan tinggi dan akses terbatas terhadap layanan kesehatan. Penelitian ini menunjukkan potensi penerapan teknik klasterisasi dalam perencanaan kebijakan kesehatan berbasis data.

Daftar Isi

Introduction

Latar Belakang

Tuberkulosis atau TBC adalah salah satu penyakit menular yang disebabkan oleh bakteri *Mycobacterium tuberculosis,* yang umumnya menyerang paru-paru manusia. Meskipun TBC dapat disembuhkan dengan pengobatan yang tepat, penyakit ini masih menjadi masalah kesehatan masyarakat global yang signifikan Menurut data dari Organisasi Kesehatan Dunia (WHO), Indonesia merupakan salah satu negara dengan jumlah kasus TBC tertinggi di dunia. Penyakit ini terutama menyerang sistem pernapasan dan dapat menular melalui udara, menyebabkan penurunan kualitas hidup dan meningkatkan angka kematian jika tidak ditangani dengan tepat meskipun berbagai upaya telah dilakukan untuk menanggulangi penyakit ini.

Jakarta, sebagai ibu kota negara Indonesia, memiliki kepadatan penduduk yang sangat tinggi, serta kondisi sosial dan ekonomi yang bervariasi, yang membuatnya menjadi salah satu daerah dengan tingkat risiko tinggi terhadap penyebaran penyakit, termasuk TBC. Faktor-faktor seperti kemiskinan, kepadatan penduduk, kualitas udara yang buruk, serta keterbatasan akses terhadap layanan kesehatan di beberapa wilayah, berkontribusi pada tingginya angka kejadian TBC di Jakarta.

Meskipun data kasus TBC sudah tersedia secara keseluruhan, tetapi masih belum ada pemetaan yang jelas mengenai daerah-daerah mana saja di Jakarta yang paling rawan terhadap penyebaran TBC. Untuk itu, diperlukan analisis yang lebih mendalam untuk mengidentifikasi dan mengklasterkan daerah-daerah di Jakarta berdasarkan tingkat kerawanan terhadap TBC. Hal ini penting agar upaya pencegahan dan penanganan dapat difokuskan di wilayah-wilayah yang paling membutuhkan.

Rumusan Masalah

Tingginya jumlah kasus Tuberkulosis (TBC) di DKI Jakarta menunjukan adanya ketimpangan dalam sistem deteksi dini serta penanganan yang belum optimal. Kompleksitas kondisi sosial ekonomi dan beragamnya karakteristik wilayah membuat strategi pengendalian secara umum kurang efektif dalam menekan penyebaran TBC. Oleh karena itu, pendekatan berbasis data untuk mengelompokan wilayah berdasarkan variabel seperti jumlah kasus TBC, indikator sosial ekonomi, dan ketersediaan fasilitas kesehatan sangat dibutuhkan. segmentasi ini diharapkan dapat mengidentifikasi wilayah prioritas untuk intervensi yang lebih terarah dan efisien.

Solusi terhadap Masalah

Untuk menyelesaikan permasalahan tersebut, riset ini menerapkan metode *Agglomerative Clustering,* yaitu salah satu teknik *hierarchical clustering*, untuk mengelompokkan wilayah di Jakarta berdasarkan data kasus TBC dan indikator sosial ekonomi lainnya . Melalui eksplorasi data dan segmentasi wilayah, penelitian ini bertujuan untuk menemukan pola penyebaran penyakit serta mengidentifikasi kelompok wilayah yang memerlukan perhatian dan penanganan khusus.

Studi Terkait

1. Pengelompokan dan Pemetaan Penyakit Tuberkulosis Paru menurut Provinsi di Indonesia Tahun 2016 menggunakan Analisis Cluster K-Means
Penulis: Arif Anjang Laksono, Bana Ali Fikri, Muhammad Atma Yadin, Sendhyka Cakra Pradana, Tegar Anugrah Widi, Edy Widodo
Sumber: Prosiding Konferensi Nasional Penelitian Matematika dan Pembelajarannya, 2018
Ringkasan: Penelitian ini menggunakan algoritma K-Means untuk mengelompokkan provinsi di Indonesia berdasarkan jumlah kasus TBC paru pada tahun 2016. DKI Jakarta termasuk dalam cluster dengan jumlah penderita sedang, bersama dengan Sumatera Utara, Jawa Tengah, dan Jawa Timur. Studi ini memberikan wawasan tentang distribusi geografis kasus TBC dan dapat membantu dalam penentuan prioritas penanganan di berbagai wilayah.

2. Analysis of Tuberculosis Disease Case Growth From Medical Record Data, Viewed Through Clustering Algorithms (Case Study: Islamic Hospital Bogor)
Penulis: La Dodo, Nenden Siti Fatonah, Gerry Firmansyah, Habibullah Akbar
Sumber: Jurnal Indonesia Sosial Sains, Vol. 4 No. 09 (2023)
Ringkasan: Studi ini menganalisis pertumbuhan kasus TBC berdasarkan data rekam medis di Rumah Sakit Islam Bogor menggunakan beberapa algoritma clustering, termasuk K-Means, Fuzzy C-Means, dan Gaussian Mixture. Fokus utama adalah pada distribusi usia dan jenis kelamin pasien, dengan tujuan untuk mengidentifikasi kelompok yang lebih rentan terhadap TBC. Hasil penelitian ini dapat memberikan panduan dalam pengambilan keputusan untuk tindakan pencegahan dan penyediaan fasilitas kesehatan.

Relevansi :

1. Pengelompokan dan Pemetaan Penyakit Tuberkulosis Paru menurut Provinsi di Indonesia Tahun 2016 menggunakan Analisis Cluster K-Means

Penelitian ini menggunakan algoritma K-Means untuk mengelompokkan provinsi di Indonesia berdasarkan jumlah kasus TBC paru pada tahun 2016. DKI Jakarta termasuk dalam cluster dengan jumlah penderita sedang, bersama dengan Sumatera Utara, Jawa Tengah, dan Jawa Timur. Studi ini memberikan wawasan tentang distribusi geografis kasus TBC dan dapat membantu dalam penentuan prioritas penanganan di berbagai wilayah.

2. Analysis of Tuberculosis Disease Case Growth From Medical Record Data, Viewed Through Clustering Algorithms (Case Study: Islamic Hospital Bogor)

Studi ini menganalisis pertumbuhan kasus TBC berdasarkan data rekam medis di Rumah Sakit Islam Bogor menggunakan beberapa algoritma clustering, termasuk K-Means, Fuzzy C-Means, dan Gaussian Mixture. Fokus utama adalah pada distribusi usia dan jenis kelamin pasien, dengan tujuan untuk mengidentifikasi kelompok yang lebih rentan terhadap TBC. Hasil penelitian ini dapat memberikan panduan dalam pengambilan keputusan untuk tindakan pencegahan dan penyediaan fasilitas kesehatan.

Dengan mengacu pada kedua studi tersebut, penelitian Anda dapat mengadopsi pendekatan clustering untuk menganalisis distribusi kasus TBC di DKI Jakarta, baik dari segi geografis maupun demografis. Hal ini akan membantu dalam mengidentifikasi area dan kelompok populasi yang memerlukan perhatian khusus, sehingga strategi penanganan TBC dapat lebih tepat sasaran dan efektif.

Data Mentah

Data yang digunakan

Dalam analisis ini, kami mengumpulkan data mentah dari berbagai sumber terpercaya yang kemudian digabungkan dan dibentuk menjadi satu data utama yang digunakan dalam proses analisis. Sumber data meliputi Badan Pusat Statistik (BPS) DKI Jakarta, Dinas Kesehatan Provinsi DKI Jakarta, serta platform pemantauan lingkungan seperti situs ISPU resmi pemerintah. Seluruh data diperoleh untuk mencerminkan kondisi wilayah DKI Jakarta dari tahun 2020 hingga 2023, agar hasil analisis tetap relevan dengan kondisi terkini. Tahun tersebut dipilih karena data yang tersedia tergolong lengkap dan memiliki tingkat granularitas yang tinggi, baik secara spasial maupun temporal. Data-data ini dipilih karena mewakili faktor-faktor kunci yang berperan dalam penyebaran TBC, seperti kepadatan penduduk, tingkat kemiskinan, dan ketersediaan fasilitas kesehatan.

1. Data Fasilitas Kesehatan
   Data ini mencakup informasi mengenai jumlah dan distribusi fasilitas kesehatan di Jakarta, seperti rumah sakit umum dan rumah sakit khusus. Fasilitas kesehatan yang lebih banyak dan lebih tersebar dapat mengurangi risiko penyebaran penyakit, sehingga data ini sangat penting untuk menilai aksesibilitas terhadap pengobatan dan pencegahan TBC.
2. Data Kasus TBC
   Data ini berisi informasi tentang jumlah kasus TBC yang tercatat di tiap kecamatan selama periode tertentu. Data ini akan menjadi indikator utama dalam menilai tingkat kerawanan TBC di setiap wilayah.
3. Data Kepadatan Penduduk
   Data ini berisikan informasi mengenai jumlah penduduk per kilometer persegi di setiap kecamatan. Kepadatan penduduk yang tinggi dapat meningkatkan potensi penyebaran penyakit menular seperti TBC, karena interaksi sosial yang lebih sering dalam ruang yang terbatas.
4. Data Sosial Ekonomi
   Data ini mencakup informasi terkait faktor-faktor sosial-ekonomi seperti tingkat kemiskinan, pendidikan, dan akses terhadap fasilitas dasar lainnya. Daerah dengan tingkat kemiskinan tinggi atau akses terbatas terhadap layanan kesehatan lebih berisiko terhadap penyebaran penyakit.

Pra-pemrosesan

Sebelum diterapkan metode clustering, seluruh data yang telah dikumpulkan dari berbagai sumber perlu melalui tahapan preprocessing untuk memastikan kualitas data yang bersih, konsisten, dan siap dianalisis. Proses ini mencakup empat tahap utama: data integration, data cleaning, data transformation, dan dimensionality reduction. Masing-masing tahap memiliki peran penting dalam menyiapkan data agar hasil analisis lebih akurat dan interpretatif.

1. Data Integration
Tahap ini bertujuan untuk menggabungkan beberapa sumber data menjadi satu tabel yang utuh dan siap digunakan. Dalam konteks analisis TBC ini, penggabungan dilakukan dari berbagai data yang memuat informasi:

- Jumlah kasus TBC per wilayah (2022–2023),
- Indeks Standar Pencemar Udara (ISPU) per wilayah (2020–2022),
- Jumlah fasilitas kesehatan, rumah sakit umum dan rumah sakit khusus (2020–2021),
- Data kepadatan penduduk dan tingkat kemiskinan (2020–2023).

Penggabungan dilakukan berdasarkan kesamaan wilayah (kecamatan/kabupaten), sehingga setiap baris data akhir mencerminkan satu wilayah dengan berbagai fitur pendukung sebagai variabel input.

2. Data Cleaning
Setelah integrasi, data diperiksa untuk mengidentifikasi nilai yang hilang (missing values) atau inkonsistensi data. Ditemukan bahwa beberapa entri pada jumlah rumah sakit umum dan rumah sakit khusus memiliki nilai null. Hal ini kemungkinan terjadi karena di wilayah tersebut tidak terdapat fasilitas rumah sakit yang dicatat.

Oleh karena itu, strategi cleaning yang digunakan adalah imputasi nilai null dengan angka 0. Alasan pemilihan metode ini:

- Secara semantik: Jika data tidak mencatat rumah sakit di suatu wilayah, sangat mungkin memang tidak ada rumah sakit, sehingga 0 adalah representasi yang logis.
- Menghindari distorsi data: Imputasi dengan nilai rata-rata atau median tidak relevan dalam konteks keberadaan rumah sakit yang bisa biner (ada/tidak ada).
- Menyederhanakan model: Nilai 0 dapat langsung diterjemahkan sebagai tidak tersedia, tanpa perlu pembentukan variabel dummy tambahan.

3. Data Transformation (Normalization)
Langkah berikutnya adalah menyetarakan skala antar fitur dengan melakukan normalisasi menggunakan RobustScaler dari library sklearn.preprocessing.

Alasan pemilihan RobustScaler:

- RobustScaler mengurangi pengaruh outlier, karena proses scaling-nya berbasis interquartile range (IQR), bukan mean dan standard deviation seperti StandardScaler.
- Data ini mengandung beberapa variabel dengan distribusi ekstrem, seperti jumlah penduduk atau jumlah kasus TBC yang bisa sangat besar di kota besar dan sangat kecil di daerah rural. Scaling dengan teknik lain seperti MinMax atau StandardScaler akan membuat nilai-nilai ekstrem terlalu dominan.
- RobustScaler mempertahankan distribusi asli secara lebih adil, karena tidak memaksa semua nilai masuk ke dalam range tertentu seperti MinMax.

Dengan normalisasi, algoritma clustering (yang sensitif terhadap perbedaan skala) dapat bekerja secara optimal.

4. Dimensionality Reduction (PCA)
Untuk menyederhanakan kompleksitas data dan mempercepat proses komputasi, dilakukan reduksi dimensi menggunakan PCA (Principal Component Analysis), dengan mengambil dua komponen utama.

Alasan penggunaan PCA dan pemilihan dua komponen:

- PCA membantu mengurangi noise dan mengeliminasi multikolinearitas antar variabel.
- Mengurangi jumlah dimensi dari banyak fitur menjadi dua komponen membuat proses visualisasi hasil clustering lebih mudah dilakukan dalam bentuk scatter plot 2D.
- mampu menangkap proporsi variansi terbesar dalam data, dan dinilai cukup representatif untuk tujuan eksploratif serta pemodelan awal.
- Selain efisiensi, pendekatan ini juga membantu dalam analisis interpretatif terhadap pola-pola utama yang tersembunyi dalam data.

Data

Setelah data digabungkan dan dibersihkan, selanjutnya dilakukan *Exploratory Data Analysis* (EDA) untuk memahami pola distribusi, hubungan antar fitur, dan karakteristik umum dari data yang digunakan dalam proses *clustering*. Tahapan ini penting untuk mengidentifikasi hubungan antar variabel, distribusi nilai, serta potensi keberadaan nilai *outliers* yang dapat mempengaruhi kualitas hasil segmentasi.

1. Struktur dan Statistik Data
   terdiri dari 14 fitur numerik yang mencakup berbagai aspek seperti jumlah kasus TBC, jumlah penduduk, prevalensi, indikator sosial ekonomi, serta ketersediaan fasilitas kesehatan. Pemeriksaan struktur data melalui fungsi describe() menunjukan bahwa tidak terdapat nilai kosong (missing values) dan setiap kolom memiliki nilai numerik yang sesuai.

```
Statistical Summary:
       jumlah target kasus tb  jumlah penderita tb  \
count              267.000000          267.000000
mean              3472.138577          808.588015
std              10394.223643         2504.227311
min                 30.000000            3.000000
25%                861.000000          198.000000
50%               1686.000000          345.000000
75%               2601.500000          569.000000
max             123030.000000        29588.000000

       Garis Kemiskinan (rupiah/kapita/bulan)  Jumlah Penduduk Miskin (ribu)  \
count                              267.000000                     267.000000
mean                             26418.084921                       3.765658
std                              26283.108985                       3.973534
min                               1121.220004                       0.124217
25%                              12589.491185                       1.079793
50%                              19054.342400                       2.951599
75%                              33478.174865                       4.330590
max                             202545.617700                      23.471535

       Persentase Penduduk Miskin  Jumlah Rumah Sakit Umum  \
count                  267.000000                267.000000
mean                     0.210751                  1.044944
std                      0.314166                  3.900771
min                      0.007892                  0.000000
25%                      0.075869                  0.000000
50%                      0.121210                  0.000000
75%                      0.170432                  1.000000
max                      2.414976                 37.000000

       Jumlah Rumah Sakit Khusus
count                 267.000000
mean                    0.295880
std                     0.937172
min                     0.000000
25%                     0.000000
50%                     0.000000
75%                     0.000000
max                     9.000000
```

1. Jumlah target kasus TBC
    - Mean: 3.472 kasus per wilayah
    - Median: 1.686 kasus (50% wilayah memiliki target kasus ≤ 1.686)
    - Maksimum: 123.030 kasus, menunjukkan adanya wilayah ekstrem (*outlier*)
    - Standar deviasi: 10.394, sangat tinggi, menunjukkan penyebaran yang lebar
    - Insight: Distribusi target kasus sangat tidak merata; terdapat wilayah dengan beban target yang jauh lebih besar dibandingkan lainnya.

2. Jumlah Penderita TBC
    - Mean: 808 penderita per wilayah
    - Median: 345 penderita, banyak wilayah berada di bawah rata-rata
    - Maksimum: 29.588 penderita (wilayah ekstrem, outlier)
    - Standar deviasi: 2.504

- insight Sebagian besar wilayah berada di bawah rata-rata, namun terdapat beberapa wilayah dengan beban kasus yang sangat tinggi. kemungkinan karena faktor lingkungan, kepadatan, atau akses ke layanan kesehatan.

3. Garis Kemiskinan (rupiah/kapita/bulan)
    - Mean: Rp 26.418, rata-rata ambang kemiskinan cukup tinggi
    - Min: Rp 1.121, sangat rendah dan kemungkinan outlier/error
    - Max: Rp 202.545, ekstrem, bisa jadi salah input atau wilayah dengan standar hidup sangat tinggi
    - Insight: Ketimpangan ekonomi antar wilayah sangat tinggi, beberapa wilayah memiliki standar hidup yang sangat rendah atau sangat tinggi.

4. Jumlah Penduduk Miskin (ribu)
    - Mean: 3.76 ribu orang
    - Min: 0.12 ribu (sekitar 120 orang)
    - Max: 23.47 ribu
    - Insight: Beberapa wilayah memiliki konsentrasi kemiskinan yang besar, berpotensi berkorelasi dengan tingginya kasus TBC.

5. Persentase Penduduk Miskin
    - Mean: 0.21%, kecil tapi ada yang mencapai 2.41%
    - 50% wilayah memiliki tingkat kemiskinan $\leq 0.12\%$
    - Insight: Meskipun secara umum angka kemiskinan rendah, beberapa wilayah menunjukkan nilai ekstrem yang perlu diperhatikan dalam analisis risiko kesehatan.

6. Jumlah Rumah Sakit Umum
    - Mean: 0.94 rumah sakit per wilayah
    - Median: 0, lebih dari 50% wilayah tidak memiliki RS umum
    - Max: 37 rumah sakit, sangat timpang
    - Insight: Terjadi ketimpangan yang signifikan dalam distribusi fasilitas kesehatan, banyak wilayah yang sama sekali tidak memiliki RS umum.

7. Jumlah Rumah Sakit Khusus
    - Mean: 0.29, sangat rendah
    - Median: 0, mayoritas wilayah tidak memiliki RS khusus
    - Max: 9 RS khusus
    - Insight: RS khusus yang dapat menangani kasus TBC kemungkinan hanya terpusat di beberapa wilayah besar.

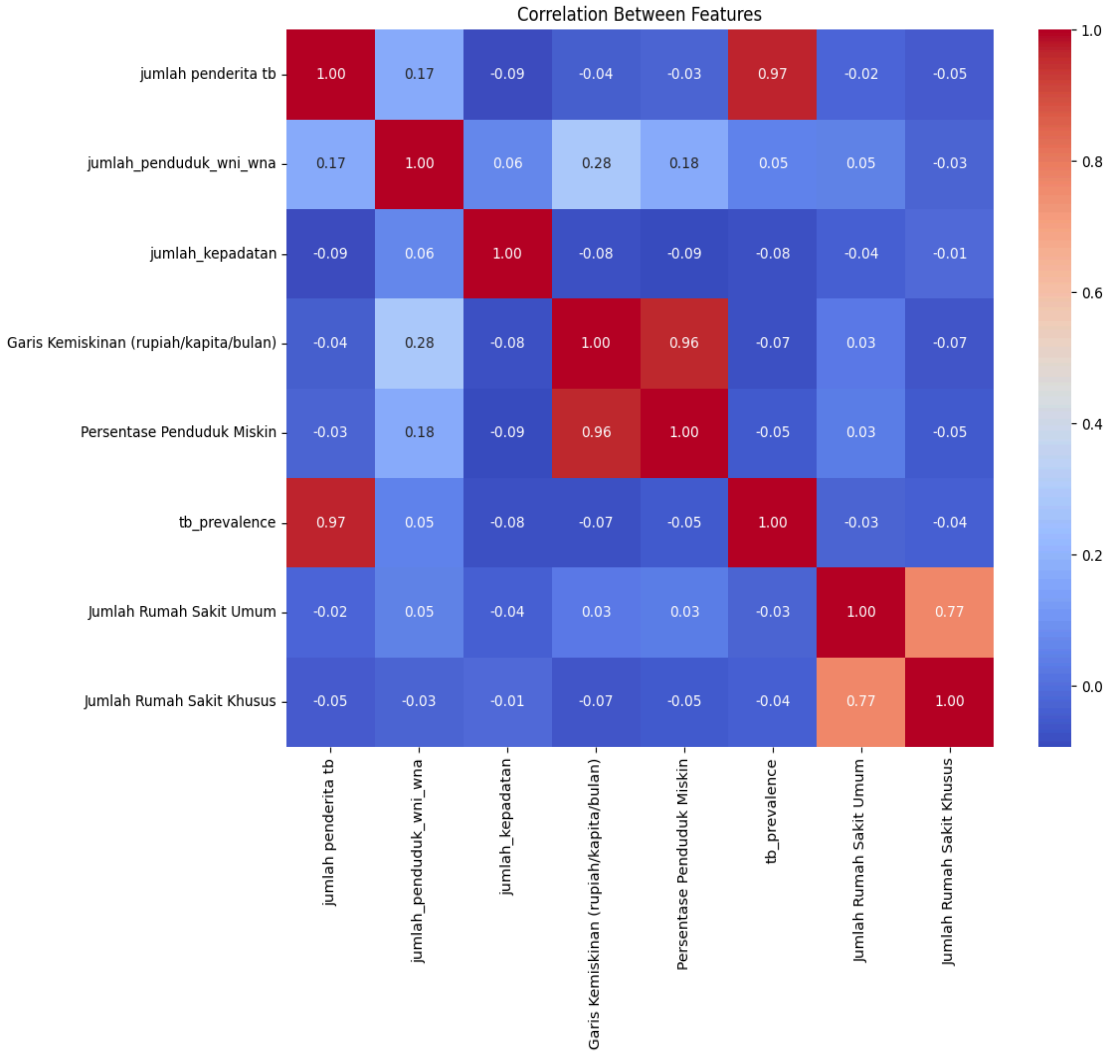
Kesimpulan

Analisis statistik menunjukan ketimpangan ekstrem dalam distribusi jumlah penderita, target kasus, dan fasilitas kesehatan antar wilayah di DKI Jakarta. Mayoritas wilayah tidak memiliki rumah sakit umum maupun khusus. Ketimpangan juga terjadi pada

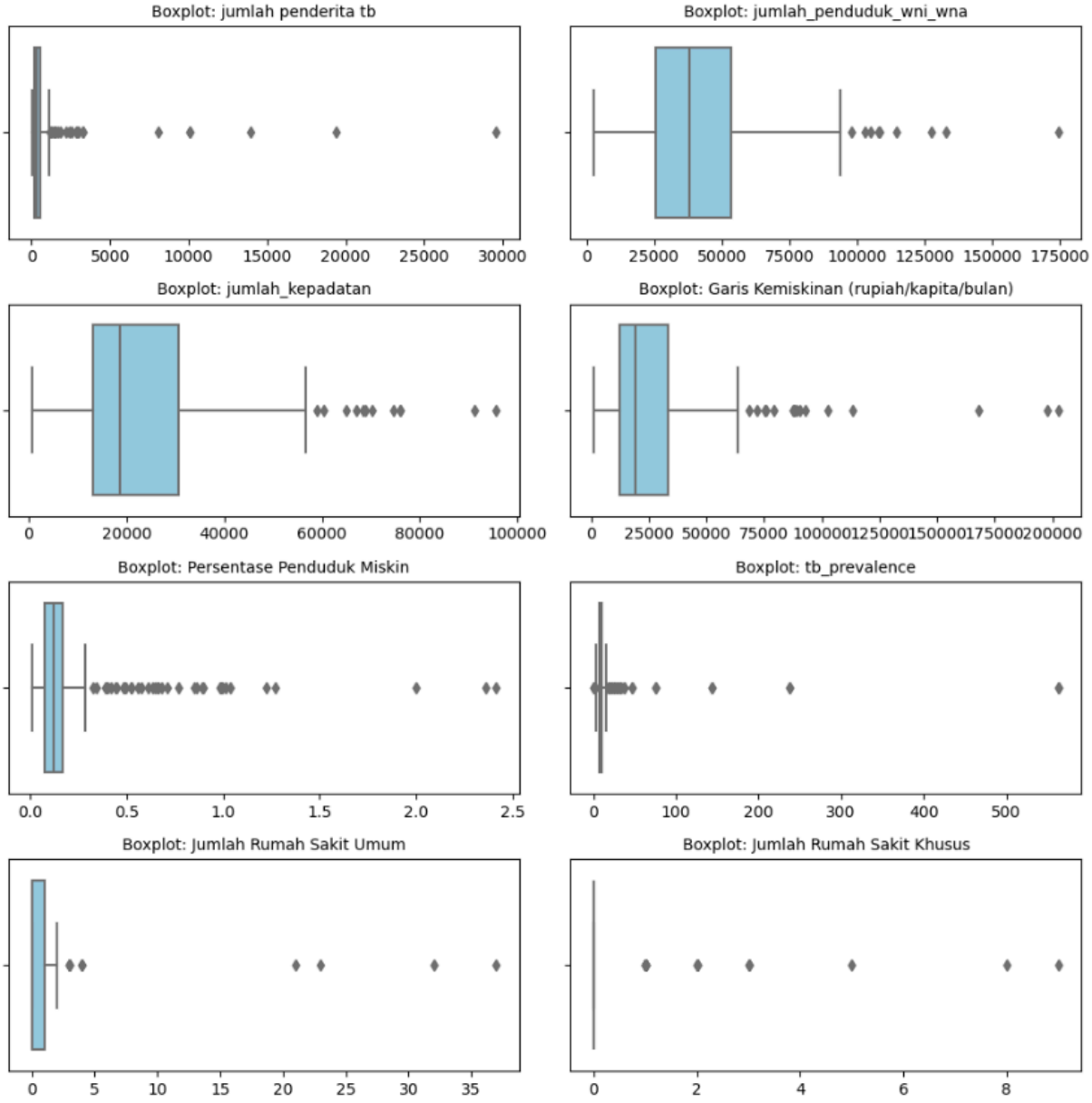
kondisi sosial ekonomi, yang ditunjukkan oleh variasi besar pada garis kemiskinan dan jumlah penduduk miskin. Temuan ini menegaskan pentingnya pendekatan segmentasi wilayah berbasis data, seperti *clustering*, untuk intervensi yang lebih terarah dan efektif dalam pengendalian TBC.

2. Korelasi antar variabel

   Analisis korelasi dilakukan menggunakan *heatmap* untuk melihat hubungan antar fitur. hasil visualisasi menunjukan bahwa terdapat korelasi yang cukup kuat antara target_tb dan jumlah_penderita, serta antara jumlah_penderita dan prevalensi_tb. hal ini menunjukan bahwa wilayah dengan jumlah kasus tinggi juga cenderung memiliki prevalensi tinggi. Sebaliknya, beberapa variabel seperti fasilitas kesehatan memiliki korelasi yang lemah terhadap prevalensi, mengindikasikan adanya ketimpangan akses layanan.



Eksplorasi Fitur Numerik



- Sebagian besar fitur menunjukkan distribusi yang skewed (miring ke kanan) dengan beberapa outlier yang jauh lebih besar dari nilai mayoritas data. Ini terlihat jelas pada fitur seperti jumlah penderita tb, jumlah_penduduk_wni_wna, Garis Kemiskinan, dan tb_prevalence.
- Outlier ini menandakan ada beberapa wilayah dengan nilai sangat tinggi, misalnya jumlah penderita TB yang jauh lebih besar dibandingkan wilayah lain, atau kemiskinan yang sangat tinggi di beberapa lokasi
- Fitur seperti ‘Jumlah Rumah Sakit Umum’ dan ‘Jumlah Rumah Sakit Khusus’ juga memiliki outlier, menunjukkan beberapa wilayah memiliki fasilitas kesehatan jauh lebih banyak.
- Median dan kuartil menunjukkan sebaran nilai yang sebagian besar terkonsentrasi pada nilai rendah sampai menengah untuk hampir semua fitur, sedangkan outlier mewakili kondisi ekstrim yang perlu diperhatikan.

Kesimpulan:
Data menunjukkan ketimpangan besar antar wilayah, terutama pada kasus TB, kepadatan penduduk, dan kemiskinan. Hal ini penting untuk strategi pengendalian TB yang lebih terfokus pada wilayah dengan nilai ekstrim (outlier) ini.

Metode

Teknik Data Mining yang Digunakan

Pemilihan Metode Clustering: Agglomerative Clustering

Dalam proyek ini, kami menggunakan Agglomerative Clustering sebagai teknik utama untuk melakukan segmentasi wilayah berdasarkan faktor-faktor yang mempengaruhi kasus TBC. Pemilihan metode ini didasarkan pada beberapa pertimbangan teknis dan hasil evaluasi performa model dibandingkan dengan metode clustering lainnya seperti KMeans dan Gaussian Mixture Model (GMM).

Alasan Memilih Agglomerative Clustering

Meskipun KMeans memiliki Silhouette Score tertinggi (~0.96), distribusi data di lapangan tidak mencerminkan asumsi KMeans yang mengharuskan cluster berbentuk bulat dan seragam. DBSCAN mendeteksi outlier dengan baik, tetapi mengabaikan banyak data penting. Oleh karena itu, Agglomerative Clustering dipilih karena memberikan keseimbangan antara akurasi, stabilitas, dan kemudahan interpretasi hierarkis wilayah.

Sementara itu, DBSCAN mampu mendeteksi **outlier** dan bekerja lebih fleksibel terhadap bentuk cluster. Namun, model ini justru **mengabaikan banyak data penting** karena terlalu sensitif terhadap parameter seperti eps dan min_samples, sehingga hasilnya kurang stabil dan sulit untuk diinterpretasikan secara hierarkis. Oleh karena itu, Agglomerative Clustering dipilih karena memberikan keseimbangan terbaik antara:

1. Hasil Evaluasi Lebih Baik
   Berdasarkan Silhouette Score, yang digunakan sebagai metric evaluasi untuk mengukur kualitas cluster (yakni seberapa terpisah dan kompak cluster yang terbentuk), metode Agglomerative Clustering menghasilkan skor yang lebih tinggi dibandingkan KMeans dan GMM.

   - Silhouette Score Agglomerative: Tinggi, menunjukkan pemisahan cluster yang baik
   - Silhouette Score KMeans/GMM: Lebih rendah, menunjukkan ada tumpang tindih atau ketidakjelasan antara cluster

2. Struktur Data Hierarkis
   Agglomerative Clustering membentuk struktur hierarki berdasarkan kedekatan antar data, dimulai dari titik-titik individual lalu digabungkan secara bertahap. Hal ini sangat sesuai untuk data spasial dan demografis yang memiliki struktur alami bertingkat, seperti wilayah administratif atau distribusi fasilitas kesehatan.

3. Fleksibilitas dalam Penentuan Jumlah Cluster
   Kami mengeksplorasi dua pendekatan dalam Agglomerative Clustering:
   - Menggunakan parameter distance_threshold = 10
     Pendekatan ini memungkinkan model untuk menentukan sendiri jumlah cluster berdasarkan ambang jarak antar titik, sehingga lebih eksploratif. Dengan threshold ini, model menghasilkan 9 cluster, yang memberikan pemetaan wilayah yang lebih kaya dan variatif, sesuai untuk eksplorasi dan analisis lebih dalam terhadap kelompok wilayah dengan karakteristik TBC berbeda.

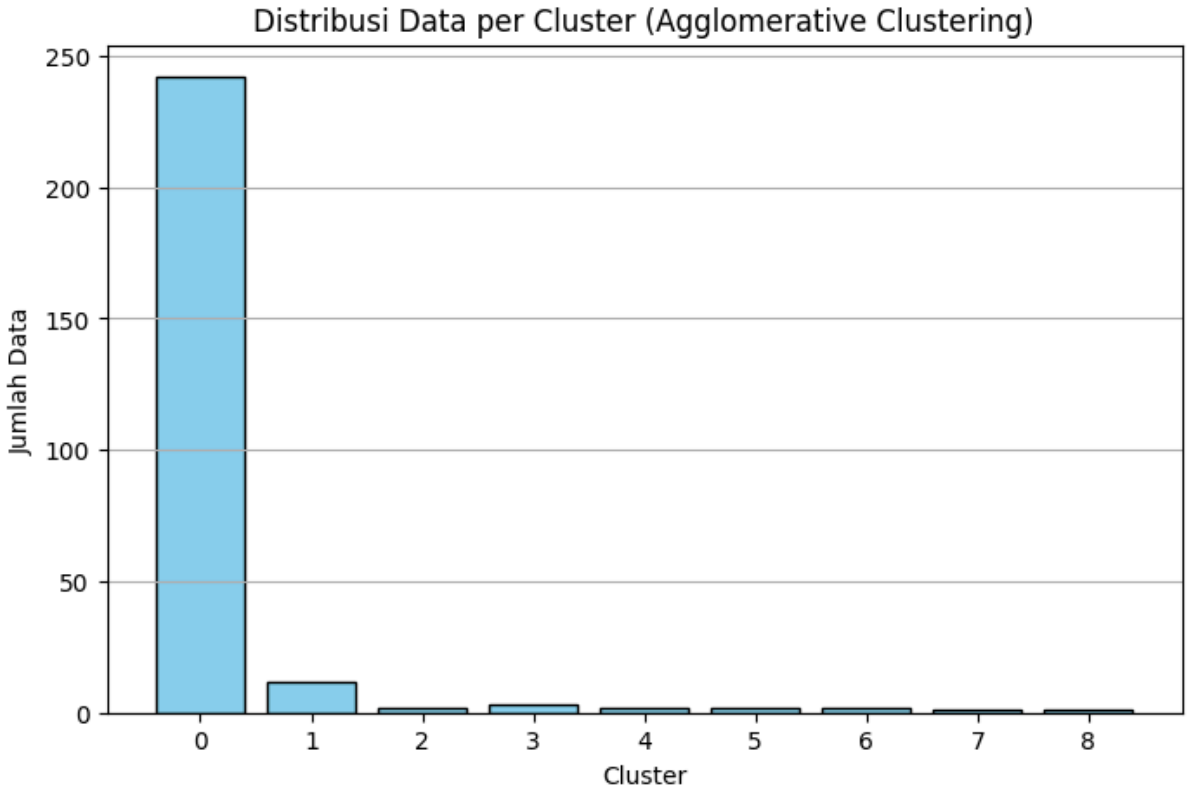
   - Menggunakan parameter n_clusters (misalnya 2 atau 3)
     Meskipun menghasilkan silhouette score yang lebih tinggi, namun jumlah cluster yang terbentuk terlalu sedikit dan kurang eksplanatif untuk menyajikan variasi kondisi antar wilayah secara rinci. Hanya 2–3 kelompok kurang mampu menangkap nuansa perbedaan dalam faktor-faktor sosial, lingkungan, dan fasilitas kesehatan yang kompleks.

     Agglomerative Clustering dipilih karena memberikan kombinasi terbaik antara **stabilitas model, interpretabilitas hasil, dan akurasi evaluasi (silhouette score)**. Pemilihan threshold berbasis jarak (bukan jumlah cluster tetap) memungkinkan pemetaan yang lebih granular dan informatif terhadap kondisi wilayah-wilayah yang memiliki kerentanan terhadap TBC. Pendekatan ini sejalan dengan tujuan analisis, yakni menggali pola-pola tersembunyi secara mendalam, bukan sekadar klasifikasi sederhana.

Hasil dan Analisis

Hasil Penerapan Model

1. Visualisasi Distribusi Data

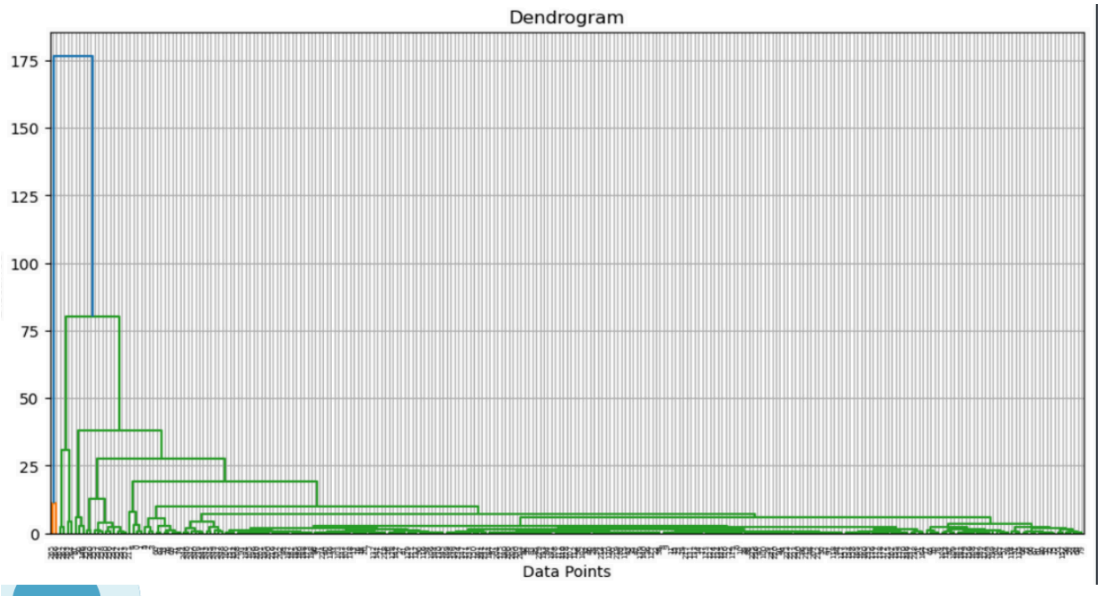


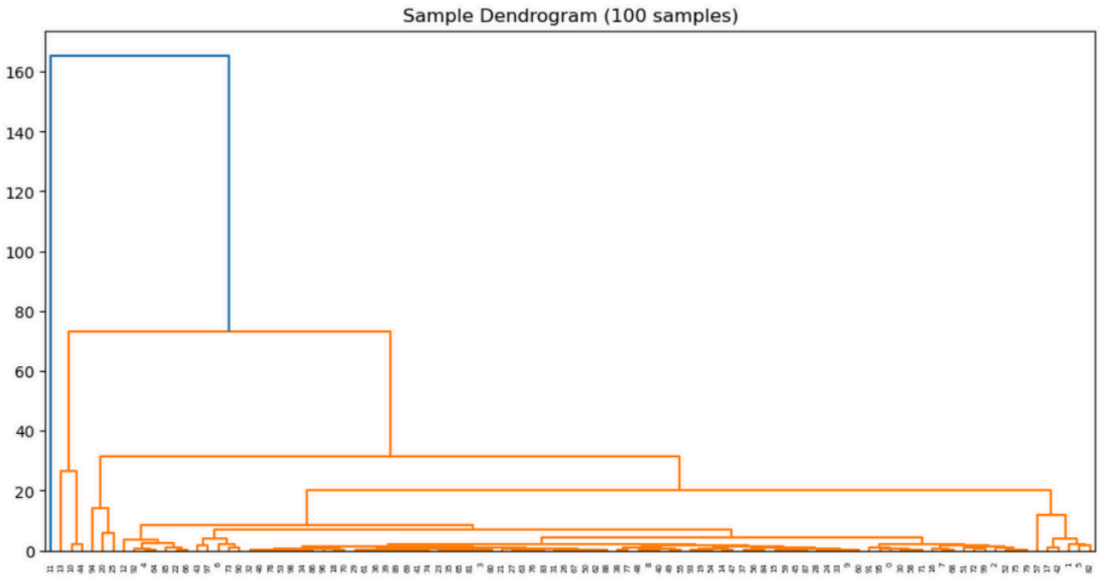
Gambar berikut menunjukkan **distribusi jumlah data pada masing-masing cluster** yang dihasilkan oleh algoritma **Agglomerative Clustering** dengan distance_threshold = 10

Dari grafik tersebut terlihat bahwa:

- **Cluster 0 mendominasi**, mencakup sekitar 243 dari total data (sekitar 90% dari seluruh observasi).
- Sementara **cluster lainnya (1–8)** hanya berisi sangat sedikit data, antara **1 hingga 9 observasi** saja.

2. Visualisasi Dendogram





Visualisasi diatas menunjukkan struktur hierarkis hasil dari metode Agglomerative Clustering. Dendogram digunakan untuk menggambarkan proses penggabungan antar wilayah berdasarkan kemiripan karakteristik, seperti jumlah kasus TBC, kepadatan penduduk, tingkat kemiskinan, dan jumlah fasilitas kesehatan.

Dalam konteks ini, pemotongan dendogram dilakukan pada ambang jarak (distances_treshold = 10) yang menghasilkan 9 cluster utama. Tinggi garis horizontal pada titik potong menunjukkan jarak antar cluster sebelum bergabung, dan menjadi dasar yang digunakan dalam segmentasi wilayah rawan TBC.

Visualisasi dendogram ini memperkuat keunggulan metode Agglomerative Clustering karena:

- Mampu memvisualisasikan hubungan hirarkis antar wilayah,
- Memberikan dasar logis untuk menentukan jumlah cluster,
- Mendukung analisis kebijakan berbasis wilayah dengan struktur spasial yang alami.

3. Deskriptif Tiap Cluster

Tabel berikut merupakan ringkasan statistik rata-rata dari masing-masing cluster untuk variabel-variabel utama yang terkait dengan TBC:

Cluster	Jumlah Kasus TBC	Penderita TBC	Penduduk	Kepadatan	Penduduk Miskin (%)	RS Umum	RS Khusus	TB Prevalence	Detection Rate	Count
0	1.766	394.8	41,293	24,013	0.19%	0.53	0.22	**9.19**	0.225	**243**
1	2.560	705.3	58,550	29,889	0.48%	30.7	6.3	**10.67**	0.238	3
2	0.624	141.3	26,180	10,658	1.72%	5.5	1.3	**2.39**	0.128	4
3	9.752	2,345	70,200	25,706	0.03%	0.78	0.11	**35.68**	0.240	9
4	37.605	9,044	62,961	14,766	0.10%	0.5	0.0	**143.64**	0.240	2
5	49.887	11,998	50,366	12,940	0.08%	0.0	0.0	**238.20**	0.240	2

6	80.504	19,360	34,389	7,510	0.21%	1.0	0.0	**562.97**	0.240	1
7	10.676	2,567	33,935	33,426	0.09%	0.0	0.5	**75.64**	0.240	2
8	123.030	29,588	52,555	11,471	0.03%	0.0	0.0	**562.99**	0.240	1

Hasil Analisis

Analisis Hasil Clustering

- **Dominasi Cluster 0**
  Cluster 0 mencakup sebagian besar wilayah yang dianalisis, menggambarkan area dengan jumlah kasus TBC yang relatif rendah namun dengan kepadatan penduduk yang tinggi. Di dalam cluster ini, prevalensi TBC tercatat berada pada angka yang moderat sekitar 9 per 100.000 penduduk, sementara tingkat deteksi kasus TBC juga berada pada angka yang cukup yaitu 0.225. Meskipun demikian, wilayah yang termasuk dalam cluster ini umumnya menghadapi keterbatasan fasilitas kesehatan, dengan rata-rata jumlah rumah sakit umum atau rumah sakit khusus yang sangat minim, kurang dari 1. Hal ini menunjukkan bahwa meskipun jumlah kasus TBC terbilang rendah, pengendalian dan penanggulangan penyakit ini tetap berjalan stabil, kemungkinan berkat intervensi yang cukup efektif di wilayah urban atau semi-urban tersebut. Meskipun begitu, masih ada potensi tantangan untuk memperbaiki layanan kesehatan di daerah-daerah ini.
- **Cluster Minor: Anomali & Outlier**
  Cluster 6 dan Cluster 8 adalah kategori minor yang menunjukkan karakteristik yang sangat mencolok, masing-masing hanya mencakup satu wilayah. Di kedua cluster ini, prevalensi TBC sangat tinggi, mencapai angka ekstrem sebesar 562 per 100.000 penduduk. Meskipun tingkat deteksi kasus TBC di wilayah ini cukup baik, dengan angka 0.24, tingginya prevalensi ini menunjukkan bahwa area ini mungkin merupakan daerah endemik dengan jumlah kasus yang luar biasa tinggi, atau bisa juga menggambarkan populasi yang terbatas namun dengan lonjakan kasus yang sangat signifikan. Menariknya, kedua wilayah ini tidak memiliki rumah sakit khusus, yang mengindikasikan adanya krisis dalam penyediaan layanan kesehatan yang sangat mendesak untuk segera ditangani, mengingat tingginya angka kasus TBC.

- **Cluster 4 dan 5**
  Cluster 4 dan Cluster 5 mewakili wilayah dengan jumlah penderita TBC yang sangat tinggi serta prevalensi yang cukup besar, berkisar antara 143 hingga 238 per 100.000 penduduk. Salah satu permasalahan utama di kedua cluster ini adalah ketiadaan fasilitas kesehatan yang memadai. Wilayah-wilayah ini hampir tidak memiliki rumah sakit atau fasilitas kesehatan yang dapat menangani kasus-kasus TBC secara optimal. Kondisi ini menciptakan potensi besar untuk terjadinya krisis kesehatan masyarakat, di mana penanganan dan pengendalian TBC akan sangat terbatas. Oleh karena itu, wilayah ini memerlukan perhatian serius dari pihak pemerintah, baik dalam hal penambahan fasilitas kesehatan maupun program intervensi yang lebih intensif untuk mengendalikan penyebaran penyakit ini.
- **Cluster 2 dan 3**
  Cluster 2 dan Cluster 3 mencerminkan wilayah dengan jumlah kasus TBC yang relatif sedang, meskipun ada variasi dalam tingkat prevalensi dan kepadatan penduduk di masing-masing area. Contohnya, Cluster 3 memiliki prevalensi TBC yang berada pada angka sedang hingga tinggi, sekitar 35 per 100.000 penduduk, dengan tingkat deteksi kasus yang cukup tinggi yakni 0.24. Namun, meskipun tingkat deteksinya cukup memadai, wilayah ini masih mengalami keterbatasan dalam fasilitas kesehatan, khususnya rumah sakit yang sangat terbatas. Meskipun tingkat prevalensi tidak setinggi cluster lain yang telah disebutkan sebelumnya, masalah kurangnya fasilitas kesehatan di cluster ini masih tetap memerlukan perhatian dan perbaikan agar upaya pengendalian TBC dapat berjalan lebih efektif dan efisien.

Analisis Banding

Cluster	Prevalensi TBC	Jumlah Penderita	Rumah Sakit Umum	Rumah Sakit Khusus	Penduduk Miskin (%)	Catatan
4	143.64	9.044	0.5	0	0.10%	Sangat tinggi prevalensi, minim fasilitas
5	238.20	11.997	0	0	0.08%	Prevalensi sangat ekstrem, nihil RS
6	562.97	19.360	1.0	0.0	0.21%	Prevalensi tertinggi, penderita terbanyak
8	562.99	29.588	0.0	0.0	0.03%	Prevalensi ekstrem + penderita sangat banyak + nihil fasilitas

Interpretasi : Prioritas

Cluster 8 adalah prioritas utama untuk intervensi berdasarkan:

- Prevalensi TBC tertinggi (562.99) → Wabah berat
- Jumlah penderita paling banyak (29.588 orang)
- Tidak ada Rumah Sakit sama sekali
- Populasi tidak terlalu padat, artinya kasus terjadi dalam komunitas kecil = transmisi komunitas sangat mungkin
- Persentase penduduk miskin rendah, artinya ini wilayah cukup mampu tetapi abai terhadap kesehatan

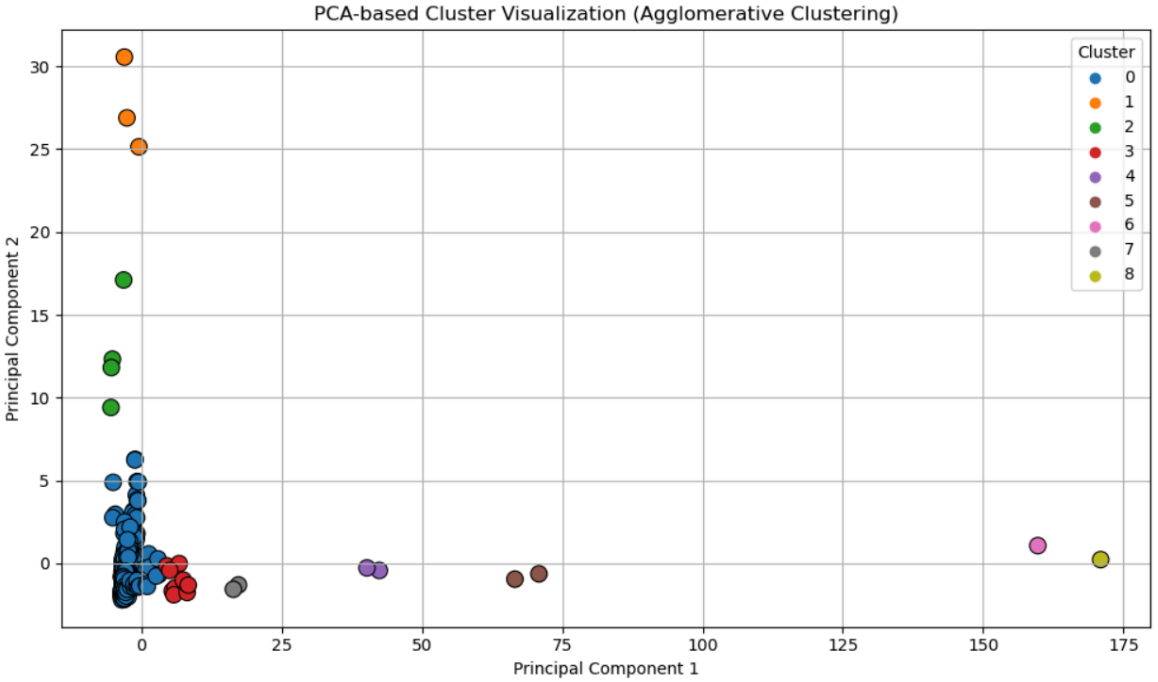
Lokasi :
Kabupaten/Kota : Jakarta Barat
Kecamatan : Kembangan
Kelurahan : Joglo

Cluster 5 dan 6 juga perlu perhatian, karena:

- Memiliki prevalensi sangat tinggi
- Jumlah penderita besar
- Hampir tidak punya fasilitas kesehatan
- Tapi tidak sebanyak dan separah Cluster 8

Evaluasi

1. Visualisasi Clustering



2. Silhouette Score

   Silhouette Score mengukur seberapa baik suatu data cocok dengan cluster-nya sendiri dibandingkan dengan cluster lain. Nilainya berada antara:

   - +1.0: Clustering sempurna; tiap data sangat dekat dengan cluster-nya dan jauh dari cluster lain.
   - 0: Data berada di batas dua cluster.
   - Negatif: Data mungkin salah tempat (mis-clustered).

   Interpretasi Skor 0.7817:
   Nilai 0.7817 merupakan skor yang sangat baik, menunjukkan bahwa:
   Data point pada umumnya cukup rapat di dalam cluster masing-masing (cohesion tinggi),
   Dan jauh dari cluster lainnya (separation tinggi).
   Cluster yang terbentuk jelas terpisah (well-separated) dan konsisten secara struktur.

Kesimpulan:

- Agglomerative Clustering berhasil membentuk cluster yang jelas dan stabil terhadap data TBC.
- Visualisasi PCA membantu memverifikasi pemecahan cluster yang baik secara spasial.
- Silhouette Score sebesar 0.7817 memperkuat bahwa model clustering berkinerja tinggi dan layak untuk interpretasi lebih lanjut.

Kesimpulan

Proyek ini bertujuan untuk mengelompokkan wilayah di DKI Jakarta berdasarkan karakteristik jumlah kasus TBC, faktor sosial ekonomi, dan ketersediaan fasilitas kesehatan menggunakan pendekatan data mining. Melalui proses preprocessing yang komprehensif—meliputi integrasi data, imputasi nilai kosong, normalisasi dengan RobustScaler, serta reduksi dimensi menggunakan PCA—data dianalisis lebih efisien dan siap untuk proses clustering.

Dengan menggunakan metode Agglomerative Clustering dan parameter distance_threshold = 10, diperoleh 9 cluster berbeda dengan silhouette score sebesar 0.7817, yang menunjukkan kualitas pemisahan cluster yang baik. Hasil clustering ini berhasil mengidentifikasi kelompok wilayah dengan tingkat kerentanan TBC yang serupa, termasuk wilayah-wilayah ekstrem dengan kasus sangat tinggi namun minim fasilitas kesehatan.

Dari hasil tersebut, dapat disimpulkan bahwa:

- Pendekatan segmentasi wilayah berbasis data efektif untuk mengungkap ketimpangan dalam beban kasus dan sumber daya kesehatan.
- Cluster yang terbentuk dapat dijadikan dasar untuk prioritisasi intervensi kebijakan, seperti penambahan rumah sakit, program deteksi dini, atau peningkatan literasi kesehatan di wilayah-wilayah rawan.

Dengan demikian, proyek ini berhasil mencapai tujuannya, yakni menghasilkan segmentasi wilayah yang dapat membantu pengambilan keputusan yang lebih tepat sasaran dalam pengendalian TBC di DKI Jakarta.

Implikasi

Konsekuensi

Penggunaan model Agglomerative Clustering dengan jarak ambang distance_threshold = 10 menghasilkan 9 cluster berbeda yang secara struktural lebih informatif dibanding hanya menggunakan jumlah cluster tetap (seperti pada KMeans). Hal ini memberikan konsekuensi bahwa wilayah-wilayah dengan karakteristik ekstrem dalam hal jumlah kasus TBC, prevalensi, dan ketersediaan fasilitas kesehatan dapat terdeteksi secara lebih granular, walaupun dengan risiko adanya distribusi cluster yang tidak merata (mayoritas data berada di 1 cluster besar).

Hasil Langsung

Salah satu hasil langsung yang diperoleh dari pemodelan ini adalah terbentuknya beberapa kelompok wilayah yang memiliki karakteristik kondisi TBC yang seragam. Setiap cluster menggambarkan tingkat keparahan dan distribusi kasus TBC yang berbeda-beda di berbagai daerah.

Berikut adalah deskripsi dari kelompok-kelompok wilayah yang terbentuk berdasarkan hasil analisis tersebut:

1. Cluster 0:
   Wilayah dalam cluster ini memiliki jumlah kasus TBC yang berkisar antara sedang hingga tinggi. Namun, meskipun memiliki tingkat kasus yang cukup signifikan, fasilitas kesehatan di daerah ini masih terbatas. Selain itu, wilayah ini juga menunjukkan persebaran kasus yang paling umum ditemukan di banyak daerah, menjadikannya sebagai cluster yang cukup representatif terhadap kondisi kebanyakan daerah dengan masalah TBC.

2. Cluster 5 dan 8:
   Wilayah pada cluster ini menunjukkan jumlah penderita TBC yang sangat tinggi. Namun, hal yang menarik adalah jumlah penduduk di wilayah-wilayah ini relatif kecil, yang berarti bahwa tingkat prevalensi TBC di daerah ini sangat tinggi bila dibandingkan dengan jumlah penduduknya. Kondisi ini mengindikasikan bahwa daerah ini mengalami masalah kesehatan yang serius terkait dengan TBC, yang perlu perhatian khusus karena prevalensinya yang luar biasa tinggi meskipun populasinya lebih sedikit.

3. Cluster 6 dan 4:
   Wilayah dalam cluster ini menunjukkan kondisi yang sangat ekstrem. Di satu sisi, daerah-daerah ini memiliki beban kasus yang sangat tinggi, sementara di sisi lain, mereka juga mengalami keterbatasan fasilitas kesehatan yang cukup parah, dengan jumlah rumah sakit yang sangat minim. Karena itu, wilayah-wilayah ini menjadi prioritas utama untuk mendapatkan intervensi kebijakan kesehatan yang segera dan terfokus, untuk mengurangi beban kasus dan memperbaiki akses terhadap layanan kesehatan yang dibutuhkan.

Dampak Temuan

Temuan dari analisis clustering ini memberikan dampak yang sangat penting dalam proses pengambilan keputusan di tingkat pemerintah dan lembaga terkait. Beberapa dampak utama yang dapat dirasakan adalah:

1. Prioritas: Dengan hasil clustering, pemerintah daerah maupun pemerintah pusat dapat dengan lebih mudah menentukan daerah mana yang perlu mendapatkan perhatian lebih. Wilayah-wilayah yang memiliki tingkat kasus tinggi namun fasilitas pendukung yang terbatas dapat diprioritaskan untuk menerima bantuan lebih cepat, sehingga penanganan dapat dilakukan secara lebih tepat sasaran.

2. Alokasi Sumber Daya: Temuan ini memungkinkan dinas kesehatan untuk merencanakan dan mengalokasikan sumber daya dengan lebih baik. Misalnya, rumah sakit, tenaga medis, atau program sosialisasi dapat dipusatkan di daerah-daerah yang memang membutuhkan lebih banyak bantuan berdasarkan hasil analisis clustering tersebut. Dengan kata lain, pemanfaatan sumber daya menjadi lebih efisien dan tepat.

3. Pembuatan Kebijakan: Hasil clustering juga memberikan dasar yang kuat untuk menyusun kebijakan yang berbasis bukti dan data. Kebijakan yang dibuat tidak hanya berdasarkan jumlah

kasus semata, tetapi juga mempertimbangkan berbagai kondisi lain seperti kapasitas fasilitas, tingkat kesadaran masyarakat, dan berbagai faktor multidimensional lainnya. Dengan begitu, kebijakan yang dihasilkan lebih komprehensif dan relevan dengan kondisi nyata di lapangan.

Dengan demikian, temuan ini tidak hanya memberikan gambaran yang lebih jelas mengenai kondisi di lapangan, tetapi juga memungkinkan pengambilan keputusan yang lebih cerdas dan terarah, guna meningkatkan efektivitas dalam penanganan masalah kesehatan atau isu-isu terkait lainnya.

Pengembangan Sistem Peringatan Dini

Dengan mengidentifikasi wilayah-wilayah yang sangat rentan, pemerintah dapat membangun sistem peringatan dini untuk memantau indikator TBC dan faktor-faktor terkait di klaster-klaster tersebut secara lebih intensif, sehingga intervensi dapat dilakukan lebih cepat saat ada tanda-tanda peningkatan kasus.

Referensi

1. Dinas Kesehatan Provinsi DKI Jakarta. (2024). *Dataset TB DKI Jakarta 2024*. Google Sheets. https://docs.google.com/spreadsheets/d/1UQZU7D7wkznfcLHF3_xHVtuLkdQIvr_496H49UL3sGk
2. Kementerian Kesehatan Republik Indonesia. (2022). *Data Kasus Penderita TBC Tahun 2022*. Portal Data Indonesia. Retrieved from https://katalog.data.go.id/dataset/data-kasus-penderita-tbc-tahun-2022/resource/361ac16c-e637-4a83-9b1e-f18ce943d4b8
3. Kementerian Kesehatan Republik Indonesia. (2022). *Data Kasus Penderita TBC*. Portal Data Indonesia. Retrieved from https://katalog.data.go.id/dataset/data-kasus-penderita-tbc
4. Badan Pusat Statistik Provinsi DKI Jakarta. (2020). *Jumlah rumah sakit umum, rumah sakit khusus, puskesmas, klinik pratama, dan posyandu menurut kabupaten/kota di Provinsi DKI Jakarta tahun 2020*. Retrieved from https://jakarta.bps.go.id/id/statistics-table/3/YmlzemNGUkNVblZLVVhOblREWnZXbkEzWld0eVVUMDkjMw==/jumlah-rumah-sakit-umum--rumah-sakit-khusus--puskesmas--klinik-pratama--dan-posyandu-menurut-kabupaten-kota-di-provinsi-dki-jakarta.html?year=2020
5. Pemerintah Provinsi DKI Jakarta. (2022). *Indeks Standar Pencemaran Udara (ISPU) Tahun 2022*. Satu Data Jakarta. Retrieved from https://satudata.jakarta.go.id/open-data/detail/indeks-standar-pencemaran-udara-ispu-tahun-2022
6. Pemerintah Provinsi DKI Jakarta. (2020). *Data Kepadatan Penduduk Provinsi DKI Jakarta Tahun 2014–2020*. Satu Data Jakarta. Retrieved from https://satudata.jakarta.go.id/open-data/detail/data-kepadatan-penduduk-provinsi-dki-jakarta-tahun-2014-2020
7. Badan Pusat Statistik Provinsi DKI Jakarta. (2022). *Garis Kemiskinan, Jumlah dan Persentase Penduduk Miskin di Daerah Menurut Kabupaten/Kota di Provinsi DKI Jakarta*. Retrieved from https://jakarta.bps.go.id/id/statistics-table/2/NjQ1IzI=/garis-kemiskinan-jumlah-dan-persentase-penduduk-miskin-di-daerah-menurut-kabupaten-kota-di-provinsi-dki-jakarta.html
8. Badan Pusat Statistik Provinsi DKI Jakarta. (2021). *Jumlah rumah sakit umum, rumah sakit khusus, puskesmas, klinik pratama, dan posyandu menurut kabupaten/kota di Provinsi DKI Jakarta tahun 2021*. Retrieved from https://jakarta.bps.go.id/id/statistics-table/3/YmlzemNGUkNVblZLVVhOblREWnZXbkEzWld0eVVUMDkjMw==/jumlah-rumah-sakit-umum--rumah-sakit-khusus--puskesmas--klinik-pratama--dan-posyandu-menurut-kabupaten-kota-di-provinsi-dki-jakarta.html?year=2021
9. Badan Pusat Statistik Provinsi DKI Jakarta. (2022). *Jumlah rumah sakit umum, rumah sakit khusus, puskesmas, klinik pratama, dan posyandu menurut kabupaten/kota di Provinsi DKI Jakarta tahun 2022*. Retrieved from https://jakarta.bps.go.id/id/statistics-table/3/YmlzemNGUkNVblZLVVhOblREWnZXbkEzWld0eVVUMDkjMw==/jumlah-rumah-sakit-umum--rumah-sakit-khusus--puskesmas--klinik-pratama--dan-posyandu-menurut-kabupaten-kota-di-provinsi-dki-jakarta.html?year=2022