https://jurnal.pustakagalerimandiri.co.id/index.php/pustakadata
DOI : https://doi.org/10.55382/jurnalpustakadata.v5i1.989

Jurnal Pustaka Data
Pusat Akses Kajian Database, Analisa Teknologi, dan Arsitektur Komputer
Vol. 5 No. 1 (2025) 132 – 143
E ISSN: 2809-7483

# Klasterisasi Tren Tuberkulosis Global dengan *Principal Component Analysis* (PCA) dan K-Means

Najibah Aisyah Muhaa[1], Larasati Mya Mulyono[2], Muhammad Rizqi Fadhilah[3], Yuyun Umaidah[4]
[1,2]Sistem Informasi, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang
[3,4]Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang
[1]najibahaisyahibah@gmail.com, [2]larasati.mulyono20@gmail.com, [3]mrfadhilah13@gmail.com,
[4]yuyun.umaidah@staff.unsika.ac.id

***Abstract***

*Tuberculosis (TB) remains a global health issue with over 10 million new cases reported annually. This study aims to identify global epidemiological patterns of TB using Principal Component Analysis (PCA) and K-Means algorithms within the Knowledge Discovery in Databases (KDD) framework. PCA reduced 22 variables into two principal components, while K-Means clustered the data based on incidence, mortality, and other health-related factors. The results revealed three main clusters, including one group of countries with a high TB burden but low treatment success (60.4%), and another with high vaccination coverage (78.1%) despite high incidence rates. Evaluation using the Silhouette Score yielded a value of 0.0694. These findings provide a basis for more targeted TB control strategies. Future research is recommended to incorporate socio-economic and temporal data for a more comprehensive analysis.*

*Keywords: Tuberculosis, PCA, K-Means, Epidemiology, KDD.*

**Abstrak**

Tuberkulosis (TB) masih menjadi masalah kesehatan global dengan lebih dari 10 juta kasus baru setiap tahun. Penelitian ini bertujuan mengidentifikasi pola epidemiologi TB secara global menggunakan *algoritma Principal Component Analysis* (PCA) dan K-Means dalam pendekatan *Knowledge Discovery in Databases* (KDD). PCA mereduksi 22 variabel menjadi dua komponen utama, sedangkan K-Means mengelompokkan data berdasarkan insidensi, mortalitas, dan faktor kesehatan lainnya. Hasil menunjukkan tiga klaster utama, salah satunya negara dengan beban TB tinggi namun keberhasilan pengobatan rendah (60,4%), dan lainnya dengan cakupan vaksinasi tinggi (78,1%) meski insidensi tinggi. Evaluasi dengan *Silhouette Score* menunjukkan nilai 0,0694. Temuan ini memberikan dasar bagi strategi pengendalian TB yang lebih terarah. Penelitian selanjutnya disarankan mempertimbangkan data sosio-ekonomi dan temporal untuk analisis yang lebih mendalam.

Kata kunci: Tuberkulosis, PCA, K-Means, Epidemiologi, KDD.



Submitted : 02-06-2025 | Reviewed : 10-06-2025 | Accepted : 18-06-2025

## 1. Pendahuluan

Tuberkulosis (TB) masih menjadi salah satu penyakit menular paling mematikan di dunia, dengan angka kejadian yang tinggi di banyak negara. Penyakit ini disebabkan oleh *Mycobacterium tuberculosis* dan dapat menyerang berbagai organ tubuh, termasuk paru-paru yang menjadi sasaran utama lokasi paling umum. Menurut data dari *World Health Organization* (WHO) pada tahun 2023 diperkirakan terdapat 10,8 juta kasus baru TB di seluruh dunia, meningkat dari 10,7 juta kasus pada tahun sebelumnya. TB merupakan penyebab kematian utama infeksi bakteri dan termasuk dalam salah satu dari sepuluh besar penyebab kematian global [1].

Pola penyebaran Tuberkulosis (TB) sangat dipengaruhi oleh faktor sosial, ekonomi, dan lingkungan [2]. menyebabkan pola epidemiologi TB dapat bervariasi antar wilayah. Dalam studi Widianingrum et. al (2020) menemukan bahwa penerapan data mining dengan Metode Algoritma K-Means *Clustering* terbukti efektif dalam mengidentifikasi pola transmisi penyakit [3]. Oleh karena itu, penting adanya mengidentifikasi pola-pola penyebaran secara akurat guna mendukung penentuan strategi pengendalian dan pencegahan penyakit yang lebih tepat sasaran.

Klasterisasi merupakan salah satu metode data mining yang dapat digunakan untuk mengelompokkan *record* data berdasarkan kemiripan karakteristik tertentu [4]. Dalam konteks epidemiologi, metode ini dapat digunakan untuk mengelompokkan wilayah berdasarkan tingkat keparahan, penyebaran, atau karakteristik populasi yang terkena TB.

Dalam klasterisasi ada beberapa jenis algoritma yang dapat digunakan salah satunya K-Means. K-Means dikenal efisien secara komputasi sehingga sangat cocok diterapkan pada dataset berukuran besar. Implementasinya pun relatif mudah, dan hasil klaster yang diperoleh dapat diinterpretasikan dengan cukup jelas. Namun demikian, K-Means juga memiliki kekurangan, salah satunya adalah sensitivitas terhadap inisialisasi awal. Hal ini berarti bahwa hasil klaster sangat dipengaruhi oleh pemilihan titik pusat klaster (*centroid*) yang dilakukan secara acak pada awal proses, yang dapat menyebabkan hasil akhir berbeda pada setiap eksekusi [5].

*Principal Component Analysis* (PCA) Untuk meningkatkan kinerja algoritma klasterisasi dan mengurangi kompleksitas data berdimensi tinggi [6], PCA dapat digunakan sebagai teknik reduksi dimensi sebelum proses klasterisasi dilakukan [7], sehingga mampu untuk menyederhanakan dataset yang memiliki banyak *record*. Oleh karena itu perlu ada penyerhanaan tetapi tetap perlu mempertahankan informasi penting, sehingga hasil klaster menjadi lebih akurat dan interpretatif.

Pada penelitian sebelumnya, oleh Ridha Afifa dengan judul "Implementasi *Principal Component Analysis* (PCA) dan *Gap Statistic* untuk *Clustering* Kanker Payudara pada Algoritma K-Means" menunjukkan bahwa kombinasi algoritma PCA dan K-Means yang dioptimalkan dengan *Gap Statistic* mampu menghasilkan klaster yang lebih baik dibandingkan metode K-Means biasa. Hasil terbaik diperoleh saat jumlah klaster ditentukan sebanyak dua, dengan nilai evaluasi *Davies Bouldin Index* (DBI) sebesar 1.195513. Hal ini menunjukkan bahwa reduksi dimensi dengan PCA berhasil mengatasi masalah multikolinieritas, dan *Gap Statistic* efektif dalam menentukan jumlah klaster optimal [8].

Penelitian ini menggunakan PCA dan K-Means, yang hingga saat ini masih jarang diterapkan dalam konteks pemetaan global epidemiologi TB. Pendekatan ini tidak hanya mengurangi kompleksitas data, tetapi juga meningkatkan akurasi dan kejelasan hasil klasterisasi. Dengan demikian, besar harapan penelitian ini memiliki kontribusi penting dalam metode analisis data epidemiologi, terutama dalam upaya memahami pola penyebaran TB secara global secara lebih mendalam dan sistematis.

Dengan memahami pola-pola epidemiologi yang terbentuk, diharapkan hasil penelitian ini dapat menjadi dasar dalam pengembangan strategi pengendalian dan pencegahan TB yang lebih efektif. Selain itu, penelitian ini juga diharapkan dapat memberikan kontribusi dalam pemanfaatan analisis data epidemiologi untuk mendukung pengambilan keputusan di bidang kesehatan masyarakat secara lebih terarah dan strategis.

## 2. Metode Penelitian

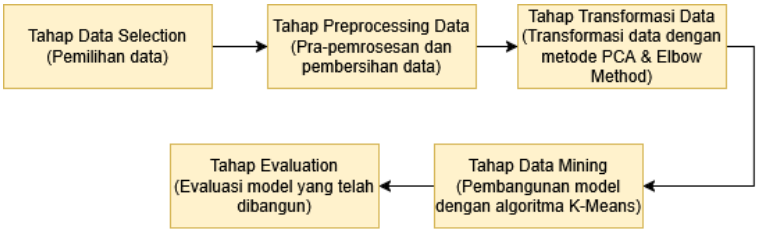
Penelitian ini menerapkan pendekatan kuantitatif dengan orientasi eksploratif guna mengungkap pola-pola epidemiologi Tuberkulosis (TB) global yang tersembunyi dalam data. Pemilihan pendekatan ini didasarkan pada fokus penelitian untuk mengidentifikasi segmentasi baru berdasarkan karakteristik epidemiologi, tanpa menguji hipotesis apriori. Melalui proses eksplorasi ini, diharapkan diperoleh pemahaman yang lebih komprehensif mengenai tren dan dinamika penyebaran TB di berbagai kawasan dunia.

Proses analisis data dilakukan dengan mengadopsi pendekatan *Knowledge Discovery in Databases* (KDD) untuk menganalisis dan mengelompokkan tren penyebaran tuberkulosis secara global menggunakan algoritma K-Means. KDD merupakan pendekatan sistematis yang digunakan untuk menemukan pola-pola signifikan dari kumpulan data yang besar dan kompleks secara bermakna serta mudah dipahami. Salah satu komponen inti dalam metode ini adalah proses data mining, yang memanfaatkan algoritma inferensial untuk menyaring informasi, membangun model, serta mengidentifikasi

Submitted : 02-06-2025 | Reviewed : 10-06-2025 | Accepted : 18-06-2025

pola tersembunyi. Model-model yang dihasilkan berperan penting dalam proses evaluasi, prediksi, dan pemahaman terhadap fenomena yang terdapat dalam data [9]. Pendekatan KDD dipilih sebab kemampuannya yang efektif dalam mengolah data berskala besar untuk memperoleh informasi yang relevan dan berguna [10].

Jenis data yang digunakan dalam penelitian ini adalah data sekunder yang diunduh melalui situs Kaggle, sehingga data tersebut tidak diperoleh secara langsung dari pengamatan atau interaksi dengan objek penelitian [11]. Dataset yang digunakan memuat berbagai atribut yang mencerminkan identitas wilayah (*Country*, *Region*, *Income_Level*), indikator kesehatan terkait tuberkulosis (seperti jumlah kasus, tingkat kematian, tingkat keberhasilan pengobatan, serta kasus tuberkulosis resistan obat), indikator ekonomi (GDP per *capita*, *Health Expenditure per capita*), serta indikator sosial dan layanan kesehatan seperti tingkat urbanisasi, prevalensi merokok, cakupan vaksinasi BCG, dan tes HIV. Secara keseluruhan, terdapat 22 atribut yang merepresentasikan kondisi epidemiologis dan sosial ekonomi di berbagai negara.




Gambar 1. Tahapan KDD

Metodologi KDD secara umum melibatkan lima langkah yang dilakukan secara sistematis dan berurutan [12], antara lain:

1. Pada tahap *Data Selection*, dilakukan pemilihan dataset yang dianggap paling relevan untuk dianalisis sesuai dengan fokus dan sasaran penelitian.
2. Pada tahap *Preprocessing Data*, dilakukan pemeriksaan pada data yang telah dipilih untuk mengidentifikasi dan memperbaiki data inkonsisten, data yang hilang (*missing values*), serta data duplikat. Selain itu, dilakukan pula *Exploratory Data Analysis* (EDA) untuk memahami struktur, distribusi serta hubungan antar variable yang ada pada dataset.
3. Pada tahap *Transformation Data*, dilakukan transformasi bentuk terhadap data yang telah dibersihkan ke dalam bentuk yang lebih sesuai untuk proses *data mining* menggunakan metode *Principal Component Analysis* (PCA) dan *Elbow Method.*
4. Pada tahap *Data Mining*, algoritma K-Means digunakan untuk melakukan klasterisasi terhadap dataset. Tahapan ini merupakan tahapan inti dari pendekatan KDD.
5. Pada tahap *Evaluation*, dilakukan penyesuaian antara pola yang diperoleh dan pengetahuan atau asumsi sebelumnya. Dalam penelitian ini, kualitas klasterisasi diukur menggunakan metode *Silhouette Coefficient.*
6. Pada tahap *Knowledge*, dilakukan peninjauan secara mendalam terhadap pola dan tren yang muncul dari data, tujuannya adalah untuk Menghasilkan kesimpulan yang valid dan keputusan yang didasarkan pada hasil analisis sebelumnya.

## 3. Hasil dan Pembahasan

Hasil dan pembahasan Tuberculosis Global menggunakan metodologi *Knowledge Discovery in Databases* (KDD) dengan metode K-Means.

### 3.1. *Data Selection*

Pada tahapan ini, dilakukan pemilihan atribut-atribut yang dianggap paling relevan dalam analisis penyebaran tuberkulosis secara global. Proses ini bertujuan untuk menyaring fitur yang memiliki kontribusi signifikan terhadap pembentukan pola dan klaster dalam data.

```
[ ]  # Pilih fitur yang relevan untuk analisis TB
     selected_features = [
         'TB_Incidence_Rate',
         'TB_Mortality_Rate',
         'TB_Treatment_Success_Rate',
         'Drug_Resistant_TB_Cases',
         'HIV_CoInfected_TB_Cases',
         'Malnutrition_Prevalence',
         'Smoking_Prevalence',
         'Access_To_Health_Services',
         'BCG_Vaccination_Coverage',
         'HIV_Testing_Coverage'
     ]
```

```
[ ]  # Ambil subset dari dataframe
     df_selected = df[selected_features]
```

Gambar 2. Seleksi Fitur Dataset

Pada gambar di atas terlihat bahwa dari 22 atribut yang tersedia dalam dataset awal, dipilih 11 fitur utama berdasarkan pertimbangan literatur dan relevansi terhadap permasalahan tuberkulosis. Atribut-atribut tersebut merepresentasikan faktor-faktor epidemiologis, kesehatan masyarakat, serta layanan kesehatan yang berpotensi mempengaruhi tingkat penyebaran dan keberhasilan penanganan kasus tuberkulosis di berbagai negara.

Submitted : 02-06-2025 | Reviewed : 10-06-2025 | Accepted : 18-06-2025

3.2. *Data Preprocessing dan Exploratory Data Analysis* (EDA)

Pada tahapan *Data Preprocessing*, dilakukan proses standarisasi data untuk menyetarakan skala antar variabel numerik yang digunakan dalam analisis. Standarisasi ini penting untuk menghindari bias yang disebabkan oleh perbedaan satuan dan skala data.

```
[9]  # Standardisasi data
     scaler = StandardScaler()
     df_scaled = scaler.fit_transform(df_selected)

[ ]  df_scaled
```

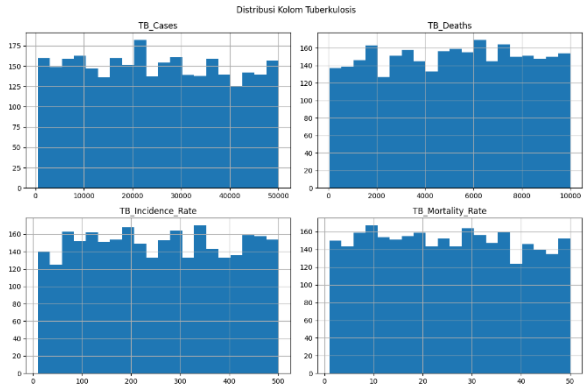
Gambar 3. Proses Standarisasi Data

Berdasarkan gambar di atas, diketahui bahwa proses standarisasi data dilakukan menggunakan metode StandardScaler dari *library scikit-learn*. Metode ini digunakan untuk mengubah setiap fitur agar memiliki nilai rata-rata sebesar 0 dan standar deviasi sebesar 1. Hasil dari proses ini disimpan dalam variabel df_scaled, yang berisi data yang telah terstandarisasi dan siap digunakan pada tahap analisis selanjutnya.

Sedangkan, pada tahapan *Exploratory Data Analysis* (EDA) dilakukan beberapa analisis awal terhadap data untuk memahami pola, distribusi, dan karakteristik dari masing-masing variabel. Berikut adalah hasil analisis awal yang dilakukan pada tahapan ini.

a. Distribusi Kolom Tuberkulosis

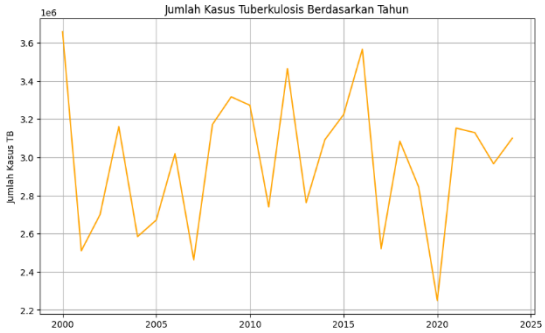
Pada analisis awal ini, dihasilkan empat visualisasi distribusi untuk empat variabel yang berkaitan dengan tuberkulosis, yaitu TB_Cases, TB_Deaths, TB_Incidence_Rate, dan TB_Mortality_Rate.




Gambar 4. Visualisasi Distribusi Kolom Tuberkulosis

Pada grafik TB_Cases terlihat bahwa distribusi jumlah kasus tuberkulosis relatif merata dengan beberapa puncak pada interval tertentu, yang mana ini menunjukkan bahwa kasus tuberkulosis tersebar cukup luas dalam rentang jumlah yang besar. Kemudian, pada grafik TB_Deaths terlihat bahwa sebaran jumlah kematian akibat tuberculosis cenderung merata, namun terlihat adanya beberapa rentang nilai dengan frekuensi yang sedikit lebih tinggi, yang mana ini bisa menunjukkan konsentrasi kematian di kisaran tertentu. Kemudian, pada grafik TB_Incidence_Rate terrlihat bahwa distribusi tingkat insidensi tuberculosis per populasi menunjukkan variasi yang cukup besar, dengan nilai-nilai yang tersebar secara merata namun tetap memiliki beberapa interval dengan frekuensi dominan. Selanjutnya, pada grafik TB_Mortality_Rate terlihat bahwa distribusi tingkat kematian akibat tuberkulosis cenderung menyebar secara merata, namun terdapat beberapa interval yang lebih menonjol, yang mana ini dapat mengindikasikan nilai tingkat kematian yang lebih sering muncul.

b. Jumlah Kasus Tuberkulosis Berdasarkan Tahun

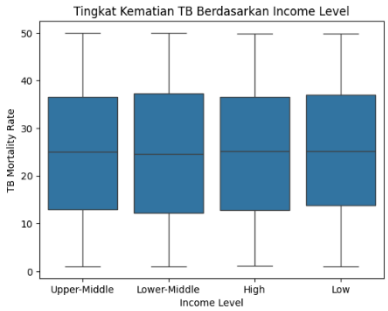



Gambar 5. Visualisasi Jumlah Kasus TB Berdasarkan Tahun

Berdasarkan gambar di atas, terlihat bahwa tren jumlah kasus tuberkulosis dari tahun 2000 hingga 2024. Terlihat bahwa jumlah kasus TB mengalami fluktuasi yang cukup signifikan dari tahun ke tahun. Pada awal tahun 2000, jumlah kasus tuberkulosis berada pada angka tertinggi, kemudian mengalami penurunan drastis pada tahun-tahun berikutnya. Selain itu, terdapat pola naik-turun (volatilitas) yang menunjukkan adanya faktor-faktor eksternal yang mempengaruhi jumlah kasus setiap tahunnya, seperti kebijakan kesehatan, cakupan vaksinasi, sistem pelaporan data, hingga gangguan besar seperti pandemi COVID-19 yang tampak berdampak pada penurunan jumlah kasus secara tajam sekitar tahun 2020. Setelah penurunan tajam tersebut, jumlah kasus TB kembali meningkat secara bertahap pada tahun-tahun berikutnya, yang kemungkinan besar disebabkan oleh normalisasi sistem layanan kesehatan dan pelaporan data setelah pandemi.

Submitted : 02-06-2025 | Reviewed : 10-06-2025 | Accepted : 18-06-2025

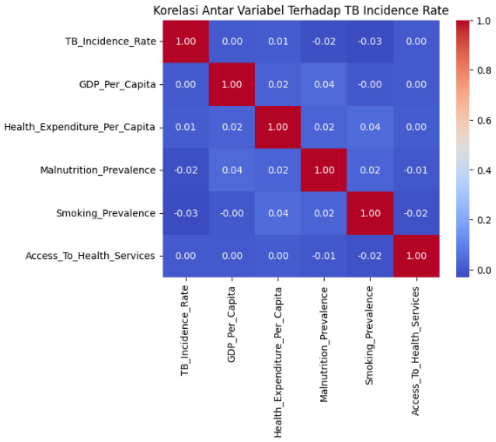
c. Tingkat Kematian Tuberkulosis Berdasarkan *Income Level*




Gambar 6. Visualisasi Tingkat Kematian Berdasarkan Income Level

Berdasarkan gambar di atas, diketahui bahwa sebaran data pada masing-masing kelompok pendapatan terlihat cukup mirip, dengan median tingkat kematian berada pada kisaran yang relatif sama, yakni sekitar angka 25. Hal ini mengindikasikan bahwa perbedaan tingkat pendapatan tidak secara signifikan memengaruhi tingkat kematian akibat tuberkulosis. Selain itu, setiap kelompok pendapatan menunjukkan rentang distribusi yang cukup lebar, yang menandakan adanya variasi besar dalam angka kematian tuberkulosis di dalam tiap kategori. Tidak tampak adanya outlier mencolok, yang menunjukkan bahwa sebagian besar data berada dalam rentang yang wajar. Sehingga, dapat disimpulkan bahwa tingkat pendapatan suatu negara bukan satu-satunya faktor penentu dalam tingkat kematian akibat tuberkulosis, karena faktor lain seperti kualitas layanan kesehatan, program pencegahan, serta kebijakan penanggulangan tuberkulosis juga memiliki peran penting dalam menekan angka kematian akibat penyakit tersebut.

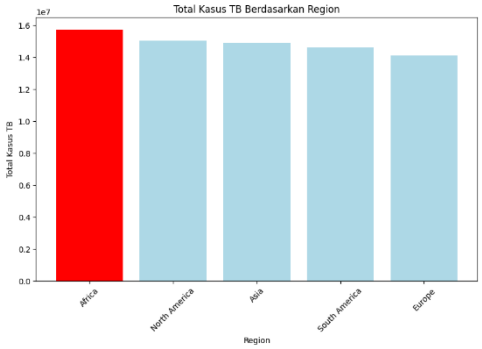
*d.* Korelasi Antar Variabel Terhadap *TB Incidence Rate*




Gambar 7. Visualisasi Korelasi Antar Variabel Terhadap TB Incidence Rate

Berdasarkan gambar di atas, terlihat bahwa hubungan antar variabel terhadap TB Incidence Rate (tingkat kejadian tuberkulosis) cenderung sangat lemah. Korelasi antara *TB Incidence Rate* dengan variabel lainnya seperti GDP Per *Capita*, *Health Expenditure* Per *Capita*, *Malnutrition Prevalence*, *Smoking Prevalence*, dan *Access To Health Services* semuanya memiliki nilai mendekati nol, baik dalam arah positif maupun negatif. Hal ini mengindikasikan bahwa tidak ada variabel yang memiliki hubungan linear yang kuat dengan TB *Incidence Rate* dalam dataset ini. Bahkan, korelasi tertinggi hanya berada di kisaran ±0.04, yang tergolong sangat rendah. Dengan demikian, dapat disimpulkan bahwa faktor-faktor seperti pendapatan per kapita, pengeluaran kesehatan, prevalensi malnutrisi dan merokok, serta akses terhadap layanan kesehatan tidak menunjukkan hubungan linear yang signifikan dengan tingkat kejadian tuberkulosis, setidaknya dalam konteks data yang dianalisis.

e. Distribusi Kasus Tuberkulosis Berdasarkan Wilayah

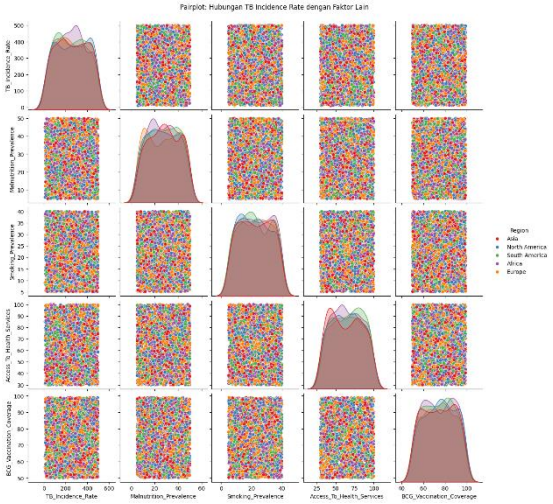



Gambar 8. Visualisasi Distribusi Jumlah Kasus Tuberkulosis Berdasarkan Wilayah

Berdasarkan gambar di atas, dapat disimpulkan bahwa benua Afrika memiliki jumlah kasus TB tertinggi dibandingkan dengan wilayah lain, dengan nilai yang mendekati 15,8 juta kasus. Hal ini ditunjukkan oleh warna batang yang mencolok (merah) yang menandai Afrika sebagai wilayah dengan beban TB terbesar. Sementara itu, wilayah lain seperti North America, Asia, South America, dan Europe juga memiliki total kasus TB yang cukup tinggi, namun sedikit lebih rendah dibandingkan Afrika, dengan kisaran antara 14-15 juta kasus. Meskipun perbedaannya tidak terlalu besar, dominasi Afrika dalam total kasus TB menunjukkan adanya kemungkinan tantangan kesehatan masyarakat yang lebih serius di kawasan tersebut, seperti akses terhadap layanan kesehatan, tingkat kemiskinan, dan prevalensi penyakit penyerta.

Submitted : 02-06-2025 | Reviewed : 10-06-2025 | Accepted : 18-06-2025

f. Perbandingan antara *TB Incidence Rate* dan Faktor Kesehatan Lainnya

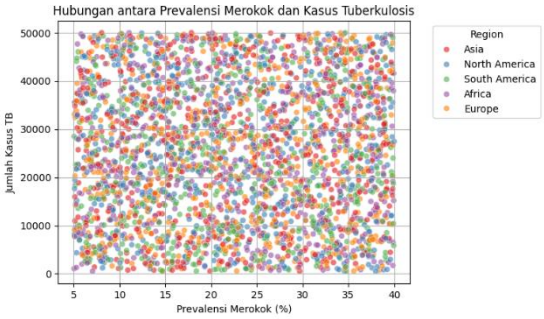



Gambar 9. Visualisasi Perbandingan antara TB Incidence Rate dengan Faktor Kesehatan Lainnya

Berdasarkan gambar di atas, dapat diketahui bahwa kasus TB tersebar di seluruh wilayah dunia, dengan variasi angka kejadian yang cukup besar, terutama pada kisaran 100-300 kasus. Tidak tampak dominasi yang mencolok dari satu wilayah tertentu, meskipun beberapa kawasan seperti Afrika dan Asia menunjukkan konsentrasi data yang cenderung lebih tinggi. Terdapat indikasi bahwa prevalensi malnutrisi memiliki hubungan positif dengan angka kejadian TB, khususnya di wilayah Afrika dan Asia, yang mengisyaratkan bahwa kondisi gizi yang buruk dapat menjadi salah satu faktor risiko utama dalam penyebaran TB.

Sementara itu, prevalensi merokok tidak menunjukkan pola hubungan yang jelas dengan TB incidence rate, menandakan bahwa faktor ini kemungkinan bukan penyebab langsung atau memiliki pengaruh yang lemah terhadap TB. Sebaliknya, akses terhadap layanan kesehatan tampak memiliki korelasi yang lebih relevan. Negara-negara dengan tingkat akses layanan kesehatan yang rendah cenderung mengalami tingkat kejadian TB yang lebih tinggi, yang menegaskan pentingnya ketersediaan fasilitas kesehatan dalam pencegahan dan penanggulangan TB. Untuk cakupan vaksinasi BCG, meskipun sebagian besar negara menunjukkan tingkat vaksinasi yang tinggi (> 70%), angka kejadian TB tetap bervariasi, menunjukkan bahwa vaksinasi saja belum cukup untuk menekan penyebaran penyakit ini. Secara keseluruhan, visualisasi ini mengindikasikan bahwa faktor gizi dan akses layanan kesehatan lebih berpengaruh terhadap penyebaran TB dibandingkan merokok atau cakupan vaksinasi.

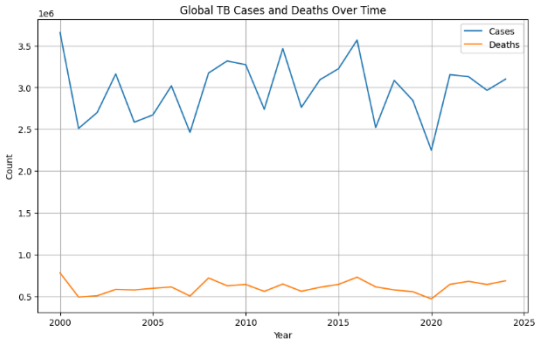
g. Hubungan antara Prevalensi Merokok dengan Kasus Tuberkulosis




Gambar 10. Visualisasi Korelasi antara Prevalensi Merokok dengan Kasus Tuberkulosis

Berdasarkan gambar di atas, terlihat bahwa tidak terdapat korelasi yang signifikan antara kedua variabel. Titik-titik data tersebar secara acak di seluruh grafik, baik untuk wilayah Asia, Afrika, Eropa, Amerika Utara, maupun Amerika Selatan. Hal ini menunjukkan bahwa meskipun ada negara dengan prevalensi merokok yang tinggi, tidak selalu diikuti dengan jumlah kasus TB yang tinggi, dan sebaliknya. Dengan distribusi data yang tidak membentuk tren linear maupun pola tertentu, dapat disimpulkan bahwa prevalensi merokok bukan merupakan faktor dominan yang secara langsung memengaruhi jumlah kasus TB.

h. Tren Jumlah Kasus Kematian Akibat Tuberkulosi




Gambar 11. Visualisasi Tren Jumlah Kasus Kematian Akibat Tuberkulosis

Berdasarkan gambar di atas, dapat diketahui bahwa jumlah kasus TB secara umum berada pada kisaran 2,5 hingga 3,6 juta kasus setiap tahunnya, dengan fluktuasi yang cukup signifikan antar tahun. Puncak jumlah kasus terjadi pada awal dan pertengahan dekade, dengan beberapa penurunan tajam, seperti pada tahun 2001 dan 2020, yang kemungkinan dipengaruhi oleh gangguan pelaporan atau peristiwa global seperti pandemi. Sementara itu, jumlah kematian akibat TB secara konsisten lebih rendah dibandingkan jumlah kasus, dengan angka yang bervariasi antara sekitar 450.000 hingga 800.000 kematian per tahun. Meskipun terdapat fluktuasi, tren

umum kematian tampak relatif stabil, dengan beberapa peningkatan kecil pada tahun-tahun tertentu. Tidak ada penurunan tajam jangka panjang yang signifikan, yang menunjukkan bahwa meskipun jumlah kasus dapat berfluktuasi, penanganan kematian akibat TB tidak menunjukkan peningkatan yang drastis dalam efektivitasnya selama periode ini.

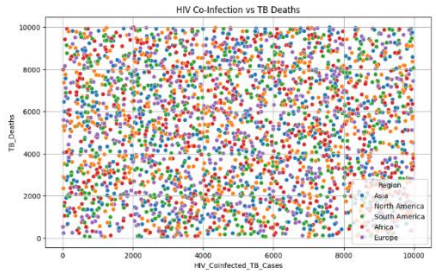
i. Tingkat Keberhasilan Pengobatan Tuberkulosis


Gambar 12. Visualisasi Tingkat Keberhasilan Pengobatan Tuberkulosis

Berdasarkan gambar di atas, dapat diketahui bahwa persentase keberhasilan pengobatan TB berada pada kisaran antara 68,7% - 74,2%, yang mana ini mencerminkan stabilitas relatif dalam efektivitas pengobatan TB selama dua dekade terakhir. Meskipun demikian, terlihat adanya fluktuasi dari tahun ke tahun, dengan penurunan tajam pada tahun 2002 ke titik terendah, dan lonjakan signifikan pada beberapa tahun berikutnya seperti tahun 2004, 2010, dan 2024.

Tren ini mengindikasikan bahwa meskipun program pengobatan TB secara umum cukup efektif, terdapat tahun-tahun tertentu di mana keberhasilan pengobatan mengalami penurunan, yang mungkin disebabkan oleh faktor-faktor seperti resistensi obat, gangguan pada sistem pelayanan kesehatan, atau masalah kepatuhan pasien terhadap pengobatan. Namun, peningkatan pada tahun-tahun setelah penurunan menunjukkan adanya pemulihan atau perbaikan dalam penanganan TB.

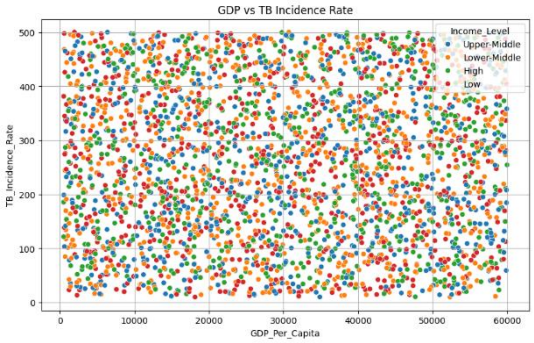
j. Korelasi antara Jumlah Kasus Tuberculosis yang Terinfeksi HIV dengan Jumlah Kematian Akibat Tuberkulosis




Gambar 13. Visualisasi Korelasi antara Jumlah Kasus TB yang Terinfeksi HIV dengan Jumlah Kematian Akibat TB

Berdasarkan gambar di atas, dapat terlihat bahwa sebaran data sangat luas dan tidak menunjukkan korelasi yang kuat secara visual antara banyaknya kasus TB yang terinfeksi HIV dengan jumlah kematian akibat TB. Titik-titik data tersebar hampir merata di seluruh rentang sumbu X dan Y, menandakan bahwa jumlah kematian akibat TB tidak selalu sebanding dengan jumlah kasus TB yang juga mengidap HIV. Ini mengindikasikan adanya faktor lain yang memengaruhi kematian akibat TB selain hanya jumlah kasus dengan ko-infeksi HIV, seperti efektivitas pengobatan, akses ke layanan kesehatan, status gizi, dan faktor sosial-ekonomi. Namun, dari sudut pandang regional, terlihat bahwa wilayah Afrika (titik merah) memiliki konsentrasi yang relatif lebih tinggi pada area dengan angka kasus HIV/TB dan kematian yang lebih tinggi, yang konsisten dengan laporan epidemiologis bahwa Afrika merupakan salah satu wilayah dengan beban TB-HIV tertinggi di dunia. Sementara wilayah lain seperti Eropa dan Amerika Utara cenderung memiliki sebaran yang lebih merata dengan jumlah kasus yang cenderung lebih rendah.

k. Korelasi antara GDP per Kapita dengan tingkat insidensi Tuberkulosis




Gambar 14. Visualisasi Korelasi antara GDP dengan Tingkat Insidensi TB

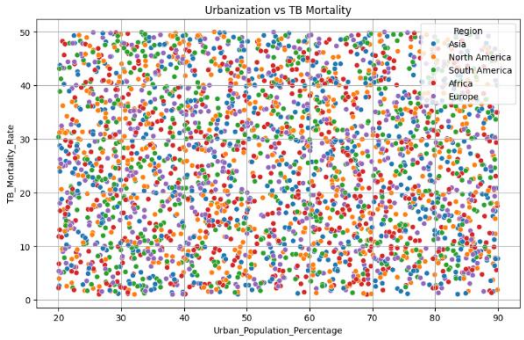
Berdasarkan gambar di atas, dapat terlihat adanya kecenderungan umum bahwa negara dengan GDP per kapita yang lebih tinggi cenderung memiliki tingkat insidensi TB yang lebih rendah. Titik-titik yang mewakili negara-negara berpendapatan tinggi (warna hijau) sebagian besar tersebar pada bagian kanan grafik (GDP tinggi) dan sebagian besar berada di bawah, menunjukkan tingkat insidensi TB yang relatif rendah. Sebaliknya, negara-negara berpendapatan rendah (warna merah) banyak terkonsentrasi pada sisi kiri grafik (GDP rendah) dan tersebar di seluruh rentang vertikal, yang menunjukkan tingginya beban TB pada kelompok pendapatan rendah. Namun, terdapat juga variasi yang cukup besar di setiap kelompok pendapatan. Beberapa negara dengan pendapatan menengah atau bahkan rendah memiliki tingkat insidensi TB yang relatif rendah, dan sebaliknya, beberapa negara berpendapatan menengah atas menunjukkan tingkat

Submitted : 02-06-2025 | Reviewed : 10-06-2025 | Accepted : 18-06-2025

insidensi yang tinggi. Hal ini menunjukkan bahwa meskipun tingkat GDP per kapita berperan dalam memengaruhi tingkat insidensi TB, faktor lain seperti sistem kesehatan masyarakat, kebijakan pengendalian TB, dan faktor sosial-ekonomi lainnya juga memiliki pengaruh signifikan.

l. Korelasi antara Persentase Populasi Perkotaan dengan Tingkat Mortalitas Akibat Tuberkulosis

Berdasarkan gambar di atas, dapat diketahui bahwa

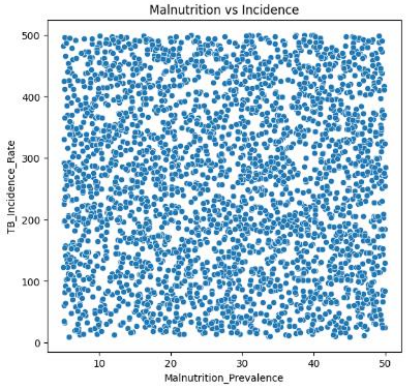



Gambar 15. Visualisasi Korelasi Persentase Populasi Perkotaan dengan Tingkat Moralitas Akibat TB

tidak tampak adanya korelasi yang kuat atau pola linear yang jelas antara tingkat urbanisasi dan tingkat kematian akibat TB. Titik-titik data tersebar cukup merata di seluruh rentang sumbu horizontal (20% - 90% urbanisasi) dan sumbu vertikal (0-50 tingkat mortalitas), menunjukkan bahwa baik negara dengan populasi sangat urban maupun negara yang masih mayoritas pedesaan dapat mengalami tingkat kematian TB yang tinggi maupun rendah.

Penyebaran warna yang merata juga menunjukkan bahwa perbedaan regional (misalnya antara Afrika, Eropa, atau Asia) tidak sepenuhnya menentukan pola hubungan antara urbanisasi dan mortalitas TB. Meski demikian, pada kisaran urbanisasi tinggi (>70%), tampaknya beberapa wilayah seperti Eropa dan Amerika Utara cenderung memiliki lebih banyak titik pada tingkat mortalitas TB yang lebih rendah, yang bisa mengindikasikan kontribusi sistem kesehatan yang lebih baik di kawasan tersebut.

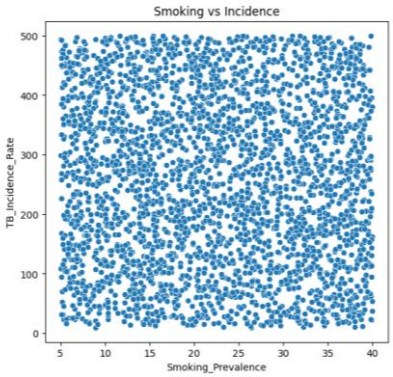
*m. Malnutrition vs Incidence*




Gambar 18. Visualisasi Korelasi antara Jumlah Dokter dengan Tingkat Keberhasilan Pengobatan TB

Berdasarkan gambar di atas, dapat terlihat bahwa titik-titik data tersebar merata di seluruh rentang sumbu horizontal (prevalensi malnutrisi antara 5 hingga 50%) dan sumbu vertikal (insidensi TB antara 0-500 kasus). Hal ini mengindikasikan bahwa tidak terdapat pola hubungan yang kuat atau korelasi linier yang jelas antara prevalensi malnutrisi dan tingkat insidensi TB. Meskipun malnutrisi secara teori dapat melemahkan sistem imun dan meningkatkan risiko TB, data dalam grafik ini tidak menunjukkan hubungan yang konsisten secara visual.

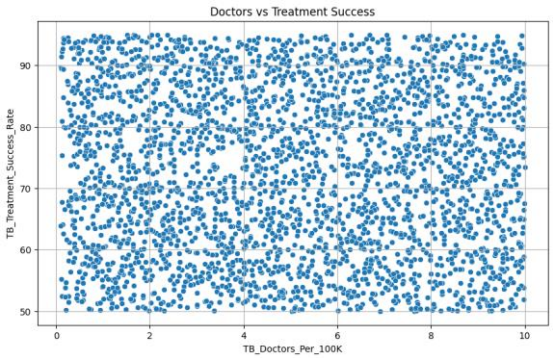
n. Korelasi *Smoking vs Incidence*




Gambar 17. Visualisasi Korelasi Smoking vs Incidence

Berdasarkan gambar di atas, dapat diketahui bahwa distribusi data pada grafik di atas menunjukkan penyebaran data yang acak tanpa pola korelasi yang menonjol antara prevalensi merokok dan insidensi TB. Rentang prevalensi merokok dari sekitar 5% hingga 40% tidak menunjukkan peningkatan atau penurunan yang jelas pada tingkat TB. Artinya, dari visualisasi ini, sulit untuk menyimpulkan bahwa merokok secara langsung berkorelasi dengan insidensi TB pada tingkat populasi dalam dataset ini.

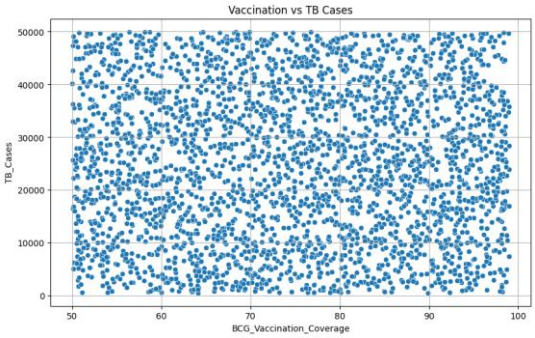
o. Korelasi antara Jumlah Dokter Tuberkulosis dengan Tingkat Keberhasilan Pengobatan Tuberkulosis




Berdasarkan gambar di atas, dapat diketahui bahwa tidak tampak pola hubungan linier yang kuat antara jumlah dokter TB dan tingkat keberhasilan pengobatan. Meskipun kita mungkin mengharapkan bahwa semakin banyak dokter TB akan

meningkatkan keberhasilan pengobatan, grafik ini menunjukkan bahwa tingkat keberhasilan cenderung tetap tinggi di berbagai tingkat kepadatan dokter. Dengan kata lain, bahkan negara-negara dengan jumlah dokter TB yang rendah dapat memiliki tingkat keberhasilan pengobatan yang tinggi, dan sebaliknya. Penyebaran yang luas dari titik-titik ini mengindikasikan bahwa faktor lain kemungkinan memainkan peran penting dalam menentukan keberhasilan pengobatan TB, seperti sistem kesehatan yang efisien, ketersediaan obat yang memadai, kepatuhan pasien terhadap pengobatan, dukungan sosial, serta program nasional untuk TB.

p. Korelasi antara Vaksinasi BCG dengan Jumlah Kasus Tuberkulosis

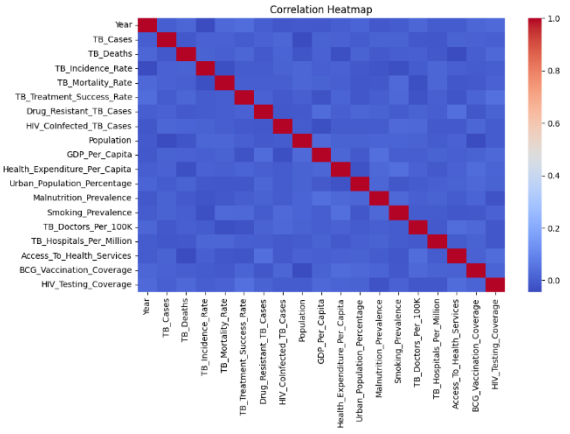



Gambar 19. Visualisasi Korelasi antara Vaksinasi dengan Jumlah Kasus TB

Berdasarkan gambar di atas, dapat diketahui bahwa tidak tampak adanya korelasi yang jelas atau pola hubungan linier yang kuat antara cakupan vaksinasi BCG dan jumlah kasus TB. Negara dengan cakupan vaksinasi tinggi (>90%) tetap bisa memiliki jumlah kasus TB yang tinggi, sementara beberapa negara dengan cakupan lebih rendah juga dapat memiliki kasus yang relatif rendah.

Analisis ini menunjukkan bahwa cakupan vaksinasi BCG secara umum mungkin tidak secara langsung menurunkan jumlah kasus TB secara keseluruhan. Hal ini dapat terjadi karena vaksin BCG lebih efektif dalam mencegah bentuk TB yang parah pada anak-anak, namun perlindungannya terhadap infeksi TB paru pada orang dewasa lebih terbatas, serta terdapat factor lain seperti kepadatan penduduk, kemiskinan, malnutrisi, akses ke layanan kesehatan, dan pengendalian infeksi yang juga memainkan peran penting dalam tingkat penularan TB.

q. Korelasi antar Variabel dalam Dataset




Gambar 20. Visualisasi Korelasi antar Variabel dalam Dataset

Berdasarkan gambar di atas, dapat diketahui bahwa korelasi yang ditampilkan menunjukkan hubungan antar variabel dalam dataset tuberkulosis (TB). Warna merah pada diagonal mencerminkan korelasi sempurna (nilai 1) antara setiap variabel dengan dirinya sendiri, yang merupakan hasil yang diharapkan dan menunjukkan keakuratan perhitungan korelasi. Di luar diagonal, sebagian besar warna yang muncul adalah biru tua hingga biru keabu-abuan, yang mengindikasikan bahwa hubungan antar variabel umumnya lemah atau tidak linier. Misalnya, korelasi antara cakupan vaksinasi BCG dan jumlah kasus TB, maupun antara prevalensi merokok dan tingkat insidensi TB, tampak sangat lemah.

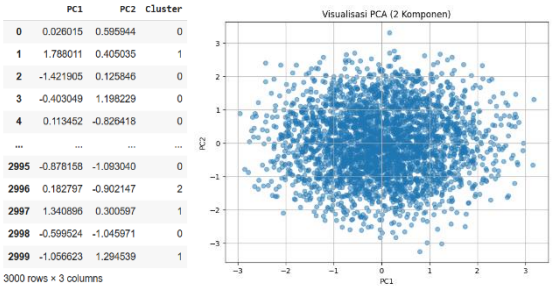
Kemudian, beberapa variabel ekonomi dan kesehatan seperti PDB per kapita (GDP_Per_Capita) dan pengeluaran kesehatan per kapita (Health_Expenditure_Per_Capita) menunjukkan korelasi positif sedang, yang mencerminkan bahwa negara dengan ekonomi yang lebih baik cenderung memiliki sistem kesehatan yang lebih kuat. Sementara itu, variabel terkait TB seperti jumlah kasus, kematian, dan tingkat insidensi TB cenderung memiliki korelasi yang sedikit lebih tinggi satu sama lain, meskipun tidak terlalu kuat. Hal ini cukup logis karena tingginya jumlah kasus biasanya berdampak langsung pada jumlah kematian dan tingkat kejadian.

Selain itu, faktor sosial seperti akses terhadap layanan kesehatan, persentase populasi urban, dan jumlah dokter TB per 100 ribu penduduk tidak menunjukkan hubungan yang signifikan dengan tingkat keberhasilan pengobatan TB. Ini mengindikasikan bahwa hubungan antar faktor dalam pengendalian TB bersifat kompleks dan kemungkinan tidak linier.

### 3.3. Transformasi Data

Pada tahap transformasi data, dilakukan menggunakan metode *Principal Component Analysis* (PCA) dan *Elbow Method*. PCA digunakan untuk menyederhanakan data yang berdimensi tinggi menjadi dua komponen utama, yaitu PC1 dan PC2,

dan *Elbow Method* dilakukan dengan memplot nilai inertia terhadap jumlah klaster (k) dari 1 hingga 10, dengan tujuan menentukan jumlah klaster yang optimal.



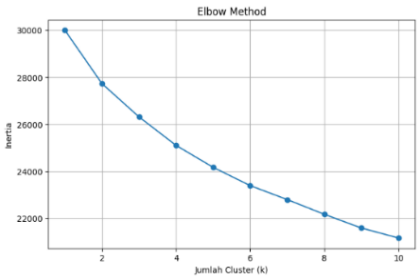
| | PC1 | PC2 | Cluster |
|---|---|---|---|
| 0 | 0.026015 | 0.595944 | 0 |
| 1 | 1.788011 | 0.405035 | 1 |
| 2 | -1.421905 | 0.125846 | 0 |
| 3 | -0.403049 | 1.198229 | 0 |
| 4 | 0.113452 | -0.826418 | 0 |
| ... | ... | ... | ... |
| 2995 | -0.878158 | -1.093040 | 0 |
| 2996 | 0.182797 | -0.902147 | 2 |
| 2997 | 1.340896 | 0.300597 | 1 |
| 2998 | -0.599524 | -1.045971 | 0 |
| 2999 | -1.056623 | 1.294539 | 1 |

3000 rows × 3 columns



Gambar 21. PCA

Dari visualisasi tersebut terlihat bahwa data untuk PCA tersebar secara merata, menunjukkan bahwa proses reduksi dimensi berhasil dilakukan dengan baik dan siap digunakan untuk proses klasterisasi selanjutnya seperti K-Means.

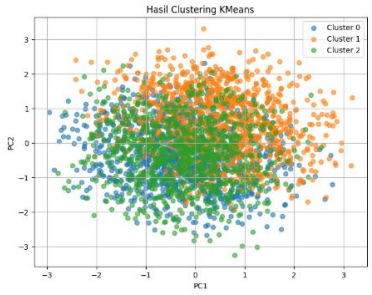



Gambar 22. Visualisasi Elbow Method

Selanjutnya dilakukan *Elbow Method* untuk menentukan jumlah klaster yang optimal. Pada grafik menunjukan bahwa terdapat penurunan nilai inertia yang cukup signifikan pada k1 – k3, dan mengalami perlambatan pada k4 – k10.

3.4. Data Mining

Pada tahapan ini kami menggunakan algoritma K-Means (*Clustering*), yang dimana mengelompokkan data berdasarkan kemiripan pola dari sejumlah variabel terkait tuberkulosis (TB), seperti jumlah kasus TB, tingkat insidensi TB, tingkat kematian akibat TB, keberhasilan pengobatan TB, kasus TB resistan obat, kasus TB yang terinfeksi HIV, dll.




Gambar 23. Visualisasi Clustering - KMEANS

Visualisasi ini menunjukan bahwa masing-masing titik pada plot merepresentasikan hasil pengelompokan ke dalam *Cluster* 0 (biru), *Cluster* 1 (oranye), dan *Cluster* 2 (hijau). Sumbu X dan Y diberi label PC1 dan PC2, yang mengindikasikan bahwa data telah direduksi dimensinya menggunakan teknik *Principal Component Analysis* (PCA) untuk memproyeksikan data ke dalam dua dimensi agar dapat divisualisasikan. Namun, distribusi titik pada visualisasi PCA menunjukkan bahwa klaster belum sepenuhnya terpisah secara jelas. Hal ini mengindikasikan kemungkinan tumpang tindih antar kelompok, yang menjadi salah satu keterbatasan dari metode K-Means saat menangani data yang tidak terpisah secara tegas.

Hasil *clustering* kemudian dibentuk kembali kedalam Jumlah data per klasternya agar mendapatkan *insight* lebih lanjut.

```
Jumlah data per cluster:
Cluster
0    1058
1    1041
2     901
Name: count, dtype: int64

Insight berdasarkan Cluster:
```

| Cluster | Year | TB_Cases | TB_Deaths | TB_Incidence_Rate | TB_Mortality_Rate |
|---|---|---|---|---|---|
| 0 | 2011.842155 | 25088.876181 | 5038.776938 | 230.693582 | 25.477817 |
| 1 | 2012.356388 | 24861.951009 | 5069.303554 | 220.371258 | 25.609894 |
| 2 | 2011.936737 | 24415.639290 | 5176.785794 | 326.366437 | 24.131188 |

Gambar 24. Data Cluster

Dari gambar terlihat bahwa jumlah data pada k0 = 1058, k1 = 1041, k2 = 901. Dan mendapatkan *insight* yakni pada

Tabel 1. Jenis Cluster

| | |
|---|---|
| **Cluster 0** | Jumlah Data: 1058 |
| | Jumlah Negara: 10 |
| | Negara: Indonesia, Nigeria, Russia, Bangladesh, USA, India, Pakistan, Brazil, China, South Africa … |
| | Region Unik: Asia, South America, Africa, Europe, North America |
| **Cluster 1** | Jumlah Data: 1041 |
| | Jumlah Negara: 10 |
| | Negara: Brazil, USA, Indonesia, Nigeria, South Africa, China, Russia, Pakistan, Bangladesh, India ... |
| | Region Unik: North America, Africa, South America, Asia, Europe |
| **Cluster 2** | Jumlah Data: 901 |
| | Jumlah Negara: 10 |

Submitted : 02-06-2025 | Reviewed : 10-06-2025 | Accepted : 18-06-2025

Negara: Russia, China, South Africa, Bangladesh, India, Brazil, Pakistan, Indonesia, USA, Nigeria …

Region Unik: South America, Africa, North America, Asia, Europe

*Cluster* 0 menunjukkan kelompok negara dengan beban tuberkulosis (TB) yang sangat tinggi, ditandai dengan angka insidensi sebesar 230 kasus per 100.000 penduduk dan tingkat kematian mencapai 25,5 per 100.000. Menariknya, klaster ini memiliki akses terhadap layanan kesehatan yang tergolong cukup baik dengan skor rata-rata 70,7, tetapi justru mencatat tingkat keberhasilan pengobatan yang rendah, yaitu hanya 60,4%. Ketidaksesuaian ini mengindikasikan bahwa akses layanan saja belum cukup menjamin efektivitas penanganan TB. Kemungkinan penyebabnya mencakup: resistensi obat, kualitas layanan yang tidak merata, atau kegagalan dalam pengawasan dan kepatuhan pasien terhadap pengobatan. Klaster ini mencerminkan kondisi negara-negara yang secara struktural sudah siap, namun masih menghadapi tantangan dalam implementasi layanan yang efektif.

*Cluster* 1 memiliki karakteristik yang lebih baik dalam hal keberhasilan pengobatan, dengan rata-rata 84,4%, meskipun angka insidensi TB masih tinggi (220 kasus per 100.000) dan mortalitas mencapai 25,6. Yang menarik, klaster ini mencatat prevalensi HIV tertinggi, yaitu 58,5%, namun tetap mampu menjaga efektivitas pengobatan. Hal ini mungkin disebabkan oleh keberadaan program pengendalian TB-HIV yang lebih terintegrasi dan berjalan baik. Selain itu, cakupan vaksinasi BCG di klaster ini juga cukup baik, yaitu 74%, yang turut berperan dalam menekan laju keparahan penyakit. Meskipun akses layanan kesehatan di klaster ini lebih rendah dari klaster 0 (65,1), hasilnya justru menunjukkan efisiensi sistem kesehatan yang mungkin lebih fokus pada penatalaksanaan dan intervensi yang tepat. Oleh karena itu, klaster ini berpotensi menjadi model percontohan bagi strategi intervensi TB di negara lain.

*Cluster* 2 ditandai dengan angka insidensi TB tertinggi, yaitu 326 kasus per 100.000 penduduk, namun tingkat kematian relatif lebih rendah (24,1). Selain itu, kasus TB resistan juga tercatat paling rendah dibanding dua klaster lainnya. Padahal, akses terhadap layanan kesehatan di klaster ini tergolong rendah (56,8). Namun, tingkat keberhasilan pengobatan mencapai 72,9%, dan cakupan vaksinasi BCG tertinggi di antara semua klaster, yaitu 78,1%. Kondisi ini menunjukkan bahwa peran imunisasi dan deteksi dini sangat signifikan dalam mengendalikan dampak TB, bahkan pada wilayah yang terbatas secara infrastruktur kesehatan. Klaster ini menekankan pentingnya strategi preventif dan proteksi biologis, seperti vaksinasi massal dan edukasi masyarakat, untuk mengimbangi keterbatasan fasilitas medis.

3.2. Evaluasi

Berdasarkan hasil perhitungan menggunakan *Silhouette Score*, nilai k terbaik untuk pengelompokan data adalah k = 3 dengan skor sebesar 0.0694. *Silhouette Score* digunakan untuk mengukur seberapa baik data dikelompokkan, dengan rentang nilai antara -1 hingga 1. Nilai yang mendekati 1 menunjukkan bahwa data berada dalam *cluster* yang tepat, sedangkan nilai mendekati 0 menunjukkan bahwa data berada di batas antara dua cluster. Dari semua nilai k yang dicoba (2 hingga 10), k = 3 tetap memberikan skor tertinggi, sehingga dianggap sebagai pilihan terbaik untuk jumlah cluster.

Meskipun secara umum hasil klasterisasi mampu memberikan gambaran yang bermakna, perlu dicatat bahwa kualitas pemisahan antar klaster masih belum sepenuhnya optimal. Visualisasi menggunakan PCA menunjukkan adanya tumpang tindih antar kelompok, dan hasil evaluasi dengan *Silhouette Score* menghasilkan nilai yang cukup rendah. Hal ini dapat disebabkan oleh beberapa keterbatasan, seperti rendahnya korelasi antar variabel, kompleksitas hubungan non-linier, serta kemungkinan bahwa beberapa indikator penting seperti kualitas kebijakan atau tingkat kepatuhan tidak terekam dalam data kuantitatif yang tersedia.

Oleh karena itu, hasil klasterisasi ini tetap perlu ditafsirkan dengan mempertimbangkan konteks sosial dan sistem kesehatan masing-masing negara. Selain itu, pendekatan lanjutan seperti penggunaan algoritma klasterisasi non-linear seperti DBSCAN atau t-SNE dapat menjadi arah penelitian berikutnya untuk meningkatkan ketajaman pemodelan.

## 4. Kesimpulan

Berdasarkan hasil penelitian, penerapan metodologi *Knowledge Discovery in Databases* (KDD) dengan metode K-Means berhasil mengelompokkan data tuberkulosis global ke dalam tiga klaster utama. Proses transformasi data melalui *Principal Component Analysis* (PCA) dan penentuan jumlah klaster optimal menggunakan *Elbow Method* serta evaluasi dengan *Silhouette Score* menghasilkan nilai optimal pada k = 3 dengan skor 0,0694.

klaster pertama terdiri dari negara-negara dengan insidensi TB 230 per 100.000, akses layanan kesehatan 70,7, tetapi hanya mencatat keberhasilan pengobatan 60,4%, menunjukkan tantangan dalam efektivitas sistem meskipun infrastruktur tersedia. Klaster kedua memiliki keberhasilan pengobatan tertinggi (84,4%), dengan akses layanan lebih rendah (65,1) namun tetap mampu mengendalikan mortalitas, bahkan dengan prevalensi HIV tertinggi (58,5%), menandakan adanya efisiensi sistem

Submitted : 02-06-2025 | Reviewed : 10-06-2025 | Accepted : 18-06-2025

pengobatan. Klaster ketiga menunjukkan angka insidensi tertinggi (326) namun tingkat kematian lebih rendah (24,1) dan keberhasilan pengobatan yang cukup baik (72,9%), dengan cakupan vaksinasi tertinggi (78,1%), menegaskan peran proteksi biologis dalam pengendalian TB meski keterbatasan layanan medis.

Hasil ini menunjukkan bahwa keberhasilan pengendalian TB tidak hanya bergantung pada akses layanan kesehatan, tetapi juga pada efektivitas pengobatan dan perlindungan biologis seperti vaksinasi. Temuan ini dapat menjadi dasar strategi intervensi yang lebih tepat sasaran, dan ke depan disarankan menambahkan variabel sosio-ekonomi dan data longitudinal untuk analisis yang lebih komprehensif.

## Daftar Rujukan

[1] D. Sitanggang, L. Simangunsong, G. F. Sundah, R. Hutahaean, and I. Indren, "Application of Data Mining for Tuberculosis Disease Classification Using K-Nearest Neighbor," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 13, no. 3, pp. 318–322, Nov. 2024, doi: 10.32736/sisfokom.v13i3.2218.

[2] A. Kusumaningrum, G. Wulandari, and A. Kautsar, "Tuberkulosis di Indonesia: Apakah Status Sosial-Ekonomi dan Faktor Lingkungan Penting?," *Jurnal Ekonomi dan Pembangunan Indonesia*, vol. 23, no. 1, pp. 1–14, Jan. 2023, doi: 10.21002/jepi.2023.01.

[3] P. A. Kusuma and A. U. Firmansyah, "Deteksi Penyebaran Penyakit Tuberkulosis dengan Algoritma K-Means Clustering Menggunakan Rapid Miner," *Jurnal Teknologi Informatika dan Komputer*, vol. 8, no. 2, pp. 41–54, Sep. 2022, doi: 10.37012/jtik.v8i2.1173.

[4] I. Nyoman and M. Adiputra, "CLUSTERING PENYAKIT DBD PADA RUMAH SAKIT DHARMA KERTI MENGGUNAKAN ALGORITMA K-MEANS," *INSERT: Information System and Emerging Technology Journal*, vol. 2, no. 2, p. 99, 2021.

[5] S. Mustrofin, T. Wicaksono, and A. Murtadho, "Perbandingan Kinerja Algoritma Kmeans dengan Kmeans Median pada Deteksi Kanker Payudara," *Jurnal Informasi dan Teknologi*, vol. 5, 2023.

[6] N. H. Alfajr and S. Defiyanti, "PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN METODE RANDOM FOREST DAN PENERAPAN PRINCIPAL COMPONENT ANALYSIS (PCA)," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3S1, Oct. 2024, doi: 10.23960/jitet.v12i3S1.5055.

[7] S. Manullang *et al.*, "ANALISIS FAKTOR PENYEBAB PENYAKIT JANTUNG MENGGUNAKAN METODE PRINCIPAL COMPONENT ANALYSIS (PCA)," *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 5, no. 3, pp. 1568–1588, Dec. 2024, doi: 10.46306/lb.v5i3.732.

[8] R. Afifa *et al.*, "Sistemasi: Jurnal Sistem Informasi Implementasi Principal Component Analysis (PCA) dan Gap Statistic untuk Clustering Kanker Payudara pada Algoritma K-Means Implementation of Principal Component Analysis (PCA) and Gap Statistic for Breast Cancer Clustering in the K-Means Algorithm," 2024. [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[9] M. Annas and S. N. Wahab, "Data Mining Methods: K-Means Clustering Algorithms," *International Journal of Cyber and IT Service Management*, vol. 3, no. 1, pp. 40–47, Mar. 2023, doi: 10.34306/ijcitsm.v3i1.122.

[10] H. Hardika, M. Martanto, A. R. Dikananda, and M. Mulyawan, "ALGORITMA REGRESI LINIER UNTUK MENINGKATKAN MODEL PREDIKSI PENJUALAN PADA TOKO DEVANJAYABAN," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 2, Apr. 2025, doi: 10.23960/jitet.v13i2.6357.

[11] R. A. Fadilla and P. A. Wulandari, "LITERATURE REVIEW ANALISIS DATA KUALITATIF: TAHAP PENGUMPULAN DATA," *MITITA JURNAL PENELITIAN* , vol. 1, 2023.

[12] F. Salsabila, T. Ridwan, and H. H, "ANALISA VOLUME PENYEBARAN SAMPAH DI KARAWANG MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, Apr. 2024, doi: 10.23960/jitet.v12i2.4226.

Submitted : 02-06-2025 | Reviewed : 10-06-2025 | Accepted : 18-06-2025