JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 12 No. 3S1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v12i3S1.5206

ANALISIS DAN EVALUASI ALGORITMA DBSCAN (*DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE)* PADA TUBERKULOSIS

Iustisia Natalia Simbolon[1*], Prawita Dwi Friskila[2]

[1,2]Informatika; Institut Teknologi Del; Jln. Sisingamangaraja Sitoluama, Laguboti, Toba, Sumatera Utara, Indonesia, 22381; Telp.+62 323-31234

Received: 15 Agustus 2024
Accepted: 5 Oktober 2024
Published: 12 Oktober 2024

Keywords:
TBC, DBSCAN, *Grid-Search, Silhouette, Davies Bouldin Index*

Corespondent Email:
iustisia.simbolon@del.ac.id

Abstrak. Tuberkulosis (TBC) adalah penyakit menular yang disebabkan oleh bakteri *Mycobacterium tuberculosis*. Penyakit ini tetap menjadi ancaman signifikan di Provinsi Sumatera Utara, dengan tingkat kasus yang tinggi menurut Dinas Kesehatan Provinsi Sumatera Utara. TBC merupakan salah satu penyebab kematian di dunia dan terdapat berbagai gejala serta faktor yang dapat meningkatkan risiko infeksi. Tingginya angka kasus TBC di Sumatera Utara menekankan pentingnya identifikasi *cluster* untuk intervensi yang efektif. Penelitian tentang pengelompokan penyakit dengan teknik *clustering* menggunakan algoritma DBSCAN telah dilakukan oleh beberapa peneliti sebelumnya dan menunjukkan hasil yang baik. Oleh karena itu, penelitian ini bertujuan untuk mengelompokkan penyebaran TBC menggunakan DBSCAN. DBSCAN dipilih karena kemampuannya mengidentifikasi *cluster* dengan baik pada data yang memiliki densitas bervariasi. Algoritma DBSCAN diterapkan dengan parameter optimal yang ditentukan melalui *Grid Search*. Hasil optimal menunjukkan nilai *epsilon* 96 dan *minPts* 5, menghasilkan 3 *cluster* utama (tinggi, sedang, dan rendah) dengan *Silhouette Coefficient* 0.41176 dan *Davies-Bouldin Index* 1,194.

Abstract. *Tuberculosis (TBC) is an infectious disease caused by the bacterium Mycobacterium tuberculosis. This disease remains a significant threat in North Sumatra Province, with high case rates according to the North Sumatra Health Department. TBC is one of the leading causes of death globally and there are various symptoms and factors that can increase the risk of infection. The high incidence of TBC in North Sumatra underscores the importance of cluster identification for effective intervention. Research on disease clustering using the DBSCAN algorithm has been conducted by several previous researchers and has shown promising results. Therefore, this study aims to cluster the spread of TB using DBSCAN. DBSCAN is chosen for its ability to effectively identify clusters in data with varying densities. The DBSCAN algorithm is applied with optimal parameters determined through Grid Search. The optimal results show an Epsilon (eps) value of 96 and MinimumPoints (minPts) of 5, resulting in three main clusters (high, medium, and low) with a Silhouette Coefficient of 0.41176 and a Davies-Bouldin Index of 1.194.*

1. PENDAHULUAN

Tuberkulosis (TBC) merupakan infeksi yang sangat umum dan berpotensi mematikan dalam banyak kasus. Penderita tuberkulosis mengalami berbagai gejala seperti demam, penurunan berat badan, dan batuk [1].

Meskipun pengobatan yang tepat dapat mencegah kematian, banyak penderita tuberkulosis tetap menghadapi masalah kesehatan berkelanjutan, serta terdapat bukti peningkatan risiko kecacatan jangka panjang dan kematian dalam populasi ini [1]. Tuberkulosis memiliki keterkaitan yang erat antara manusia dengan lingkungannya [2], terutama di daerah perkotaan dengan populasi dan kepadatan tinggi, sehingga informasi yang akurat mengenai lingkungan perkotaan terkait tuberkulosis sangat penting.

Menurut data dari Dinas Kesehatan Provinsi Sumatera Utara, Laporan Penemuan dan Pengobatan Pasien TBC di Provinsi Sumatera Utara menunjukkan bahwa pada tahun 2022 banyak kota dan kabupaten di wilayah Provinsi Sumatera Utara, terutama Kota Medan, memiliki kasus tertinggi di provinsi tersebut [3]. Banyaknya kasus tuberkulosis sangat berbahaya bagi masyarakat di Indonesia, khususnya di Provinsi Sumatera Utara. Dalam konteks epidemiologi, identifikasi *cluster* penyebaran penyakit merupakan langkah untuk memahami dinamika penyebaran dan untuk merancang intervensi yang efektif [4]. Dengan mengetahui daerah-daerah yang memiliki tingkat infeksi tinggi, sedang, dan rendah, pihak berwenang dapat mengarahkan sumber daya dan upaya pencegahan serta pengobatan dengan lebih efektif. Identifikasi *cluster* penyebaran tuberkulosis di Provinsi Sumatera Utara dapat membantu dalam merancang pemantauan dan evaluasi terhadap TBC.

Penelitian terdahulu berjudul ”Implementasi Metode *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) Dalam Mengelompokkan Penyebaran Tuberkulosis” membahas tentang pengelompokan wilayah penyebaran penyakit tuberkulosis di negara-negara HBC (*High Burden Country*) [5]. Penelitian ini menggunakan *dataset* kasus tuberkulosis dari 28 negara HBC dengan tujuan untuk mengetahui hasil analisis deskriptif penyebaran penyakit tuberkulosis dan melihat hasil pengklasteran penyebaran penyakit tuberkulosis menggunakan metode DBSCAN. Hasil dari penelitian ini menunjukkan variasi proporsi penyakit seperti TB/HIV, TB Paru, EPTB, serta penderita anak-anak, perempuan dan laki-laki di populasi negara India, China dan Indonesia memiliki kasus tertinggi. DBSCAN menghasilkan tiga kelompok utama dan mengidentifikasi satu negara sebagai *noise*. Hal ini menunjukkan bahwa DBSCAN mampu mengelompokkan data secara otomatis tanpa jumlah *cluster* yang ditentukan sebelumnya, serta kemampuannya dalam menangani data *noise*.

DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) merupakan salah satu algoritma *clustering* yang cukup efektif untuk menganalisis penyebaran penyakit seperti tuberkulosis. Algoritma ini mengelompokkan titik-titik data berdasarkan kepadatan di sekitar wilayah tertentu dan mengabaikan data yang dianggap sebagai *noise* atau *outlier* [5]. Keunggulan DBSCAN dalam konteks ini adalah kemampuannya untuk bekerja tanpa memerlukan jumlah *cluster* yang ditentukan sebelumnya, kemampuannya dalam mendeteksi *noise*, dan efektivitasnya pada data dengan kepadatan yang bervariasi [6]. Algoritma DBSCAN juga akan menentukan nilai *epsilon* dan jumlah *minimum points* yang optimal untuk mengelompokkan daerah penyebaran penyakit. Nilai optimal kedua parameter ini akan dicari melalui teknik optimasi, yakni *Grid Search* [7]. Tingkat akurasi dievaluasi menggunakan *Silhouette Coefficient* dan *Davies-Bouldin Index* untuk menilai kinerja algoritma DBSCAN [6]. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam memahami penyebaran tuberkulosis di Provinsi Sumatera Utara dan membantu pihak berwenang dalam mengambil langkah-langkah pencegahan yang lebih efektif.

2. TINJAUAN PUSTAKA

2.1 *Data Preprocessing*

Data preprocessing merupakan tahap awal yang mengubah data mentah menjadi format yang sesuai dengan kebutuhan [8]. Data yang dikumpulkan dapat beragam dan berasal dari berbagai sumber, sehingga mungkin terjadi kesalahan selama pengumpulan data yang menyebabkan data hilang, data salah, atau inkonsistensi. *Data cleaning* melibatkan proses memperbaiki kecacatan atau inkonsistensi data, seperti *missing value*, data ekstrem, dan data duplikat [8].

Missing value terjadi ketika data yang tidak memiliki nilai untuk beberapa atribut. Dengan mengisi nilai hilang secara manual atau menggunakan nilai rata-rata pada atribut adalah hal paling umum untuk dilakukan. Sedangkan, *outlier* merupakan nilai yang menyimpang jauh dari observasi lain yang dapat menyebabkan kesalahan dalam spesifikasi model dan estimasi parameter. Teknik yang digunakan dalam mengatasi nilai *outlier* adalah *Interquartile Range* (IQR) yang memiliki rentang nilai antar kuartil pertama (Q1) dan kuartil ketiga (Q3) [9].

$$Batas\ Bawah = Q1 - 1{,}5 \times IQR \quad \textbf{(1)}$$
$$Batas\ Atas = Q3 + 1{,}5 \times IQR$$
$$IQR = Q3 - Q1$$

2.2 *Exploratory Data Analysis* (EDA)

Analisis Data Eksplorasi digunakan untuk menyelidiki dan merangkum karakteristik data melalui metode visualisasi. EDA membantu menemukan pola, menguji hipotesis, dan memeriksa asumsi [10]. Terdapat empat jenis utama EDA, yaitu Univariat Non-Grafis menganalisis data satu variabel tanpa grafik, Univariat Grafis menggunakan grafik seperti histogram dan *boxplot* untuk menggambarkan data satu variabel, Multivariat Non-Grafis menunjukkan hubungan antara dua atau lebih variabel melalui tabulasi silang atau statistik, dan Multivariat Grafis menampilkan hubungan antar variabel menggunakan grafik *plot* batang. Penggunaan EDA untuk menampilkan statistik deskriptif seperti merangkum data, seperti *mean*, median, kuartil (Q1, Q2, Q3), standar deviasi, minimum dan maksimum.

2.3 *Data Transformation*

Transformasi data mengubah data menjadi format yang sesuai untuk analisis dan mencakup normalisasi, standardisasi, dan diskritisasi [8].

a. Normalisasi menyesuaikan skala data dalam rentang tertentu, seperti 0 hingga 1 [11].
b. Standardisasi mengubah data sehingga memiliki rata-rata nol dan deviasi standar satu.
c. Diskritisasi mengubah data kontinu menjadi data kategorikal dengan membagi rentang nilai menjadi interval.

2.4 Algoritma DBSCAN

DBSAN (*Density-based Spatial Clustering of Applications with Noise*) merupakan algoritma klasterisasi berbasis kepadatan yang dapat menemukan *cluster* dengan bentuk tidak beraturan dalam data yang mengandung *noise* [6]. Algoritma ini mendefinisikan *cluster* sebagai kumpulan maksimal dari titik-titik yang terhubung berdasarkan kepadatan. Konsep utama dari DBSCAN meliputi lingkungan

radius ε (ε-*neighborhood*), objek inti (*core point*) yang memiliki setidaknya *minPts* dalam ε-*neighborhood*, dan objek yang dapat dijangkau berdasarkan kepadatan (*density reachable*) melalui *core point*. Dua objek dikatakan terhubung berdasarkan kepadatan (*density connected*) jika ada objek lain yang dapat menjangkau keduanya.

Proses DBSCAN dimulai dengan memilih titik secara acak, seluruh titik akan dikumpulkan dalam radius ε untuk membentuk *cluster* jika titik tersebut termasuk dalam *core point* dan titik akan ditandai sebagai *noise* jika bukan termasuk *core point*. Kelebihan DBSCAN termasuk kemampuannya dalam mengidentifikasi *outlier*, mengelompokkan data dengan pola tidak beraturan, memindai seluruh data dalam satu iterassi, dan tidak memerlukan penentuan jumlah *cluster* awal [12]. Namun, kekurangannya adalah tidak efektif untuk data dengan kepadatan seragam, tidak cocok untuk data berdimensi tinggi, dan memerlukan penyesuaian parameter (ε dan *minPts*) oleh pengguna [13].

2.5 Matrik Evaluasi

Evaluasi klasterisasi bertujuan untuk mengukur kualitas *cluster* dan interpretasi hasil. Adapun dua jenis matrik evaluasi yang dipakai sebagai berikut [14].

a. *Silhouette Coefficient* mengukur seberapa baik suatu titik berada dalam *cluster* dibandingkan dengan *cluster* tetangga. Nilai mendekati +1 berarti titik jauh dari *cluster* lain, 0 berarti di batas antara *cluster*, dan negatif menunjukkan kemungkinan berada di *cluster* lain. Di mana $a(i)$ adalah rata-rata jarak titik (i) ke titik lain dalam *cluster* yang sama dan $b(i)$ adalah rata-rata jarak ke titik dalam *cluster* terdekat.

$$S(i) = \frac{b(i) - a(i}{\max\left(a(i), b(i)\right)} \quad \textbf{(2)}$$

b. *Davies-Bouldin Index* (DBI) mengukur seberapa terpisah *cluster* satu sama lain. Nilai lebih rendah menunjukkan *cluster* lebih terpisah. Di mana k adalah jumlah *cluster*, σ_i adalah rata-rata jarak titik dalam *cluster* i ke *centroid*, dan d_{ij} adalah jarak antara *centroid cluster i* dan *j*.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max\, j \neq i \left(\frac{\sigma i + \sigma j}{d_{ij}}\right) \quad (3)$$

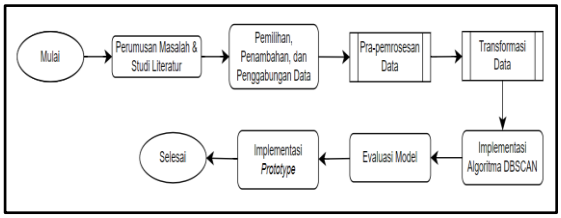
3. METODE PENELITIAN

Dalam penelitian ini diuraikan metode penelitian yang dilakukan. Metode penelitian dapat dijelaskan melalui **Gambar 1.**

3.1. Pengumpulan Data

Data yang digunakan adalah data total kasus tuberkulosis dari 33 tingkatan (kabupaten/kota) di Provinsi Sumatera Utara, yang diperoleh dari Dinas Kesehatan Provinsi Sumatera Utara dan mencakup periode 2016 hingga 2023. Data tersedia dalam delapan *file Excel* terpisah, masing-masing untuk setiap tahun. Setiap *file* berisi 33 baris data dengan 18 variabel berbeda, termasuk informasi tentang jenis pasien, lokasi penyakit, riwayat pengobatan, kelompok usia, jenis kelamin, total kasus, dan tahun pelaporan.

Setiap variabel data dipilih berdasarkan standar WHO untuk memastikan konsistensi dan kemampuan perbandingan dengan data global, termasuk diagnosis (bakteriologis dan klinis), lokasi penyakit (paru dan ekstraparu), riwayat pengobatan, dan status kambuh. *Longitude* dan *latitude* dari *Google Maps* ditambahkan untuk memperkaya analisis dan memungkinkan pembuatan peta interaktif guna memudahkan identifikasi lokasi dengan beban penyakit tinggi. Delapan *file Excel* digabungkan menjadi satu *file* untuk menciptakan sumber data yang terintegrasi dan konsisten, mencakup periode 2016-2023 dan siap untuk tahap *data preprocessing*.



Gambar 1. Metode Penelitian

3.2. *Exploratory Data Analysis* (EDA)

Eksplorasi data dilakukan untuk memahami struktur dan karakteristik data tuberkulosis. Langkah-langkah yang dilakukan mencakup perhitungan statistik deskriptif seperti *mean*, modus, dan standar deviasi untuk variabel kasus pada setiap tipe pasien. Melalui penggunaan teknik EDA, dilakukan analisis dengan *dataset* kasus tuberkulosis di Provinsi Sumatera Utara tahun 2016-2023.

Tabel 1. Beberapa baris pertama *dataframe*

	Tingkatan	Tahun	Longitude	Latitude	Pasien Baru Bakteriologis Paru	Pasien Baru Klinis Paru
0	Nias	2023	97.52	1.127	116.0	56.0
1	Nias	2022	97.52	1.127	132.0	121.0
2	Nias	2021	97.52	1.127	62.0	24.0
3	Nias	2020	97.52	1.127	68.0	129.0
4	Nias	2019	97.52	1.127	271.0	0.0

Data yang ditampilkan menunjukkan beberapa baris pertama dari *dataset* yang digunakan. *Dataset* mencakup informasi tentang total kasus tuberkulosis di 33 kabupaten/kota di Provinsi Sumatera Utara.

Tabel 2. Ringkasan statistik analisis deskriptif

	Tahun	Longitude	Latitude	Pasien Baru Bakteriologis Paru	Pasien Baru Klinis Paru
count	264	264	264	264	264
mean	2019.5	98.890606	2.313788	431.007576	327.621212
std	2.29564	0.740313	0.921597	602.179176	631.497635
min	2016	97.42	0.374	38	0
25%	2017.75	98.45	1.61	161.75	61.75
50%	2019	98.97	2.384	260	172
75%	2021.25	99.44	3.050	424.5	340.25
max	2023	100.18	3.867	3927	4880

Tabel 2 menyajikan ikhtisar statistik yang mencakup rata-rata, median, nilai minimal, dan maksimal untuk setiap kolom pada *dataframe*. Informasi ini memberikan wawasan penting tentang distribusi dan variasi dalam setiap variabel data.

3.3. *Data Preprocessing*

Pra-pemrosesan data merupakan langkah penting dalam analisis data untuk memastikan kualitas dan integritas data sebelum diterapkan pada algoritma *machine learning* seperti klasterisasi DBSCAN [15].

a. Pemeriksaan Nilai *Null*

Dataset yang digunakan mencakup informasi penting seperti jumlah kasus, lokasi (*longitude* dan *latitude*), tingkatan (kabupaten/kota), dan tahun kasus dilaporkan. Berdasarkan hasil pemeriksaan manual yang dilakukan menunjukkan bahwa *dataset* tidak mengandung nilai *null*, sehingga setiap kolom data dapat digunakan secara optimal dan analisis.

b. Pemeriksaan Nilai Duplikat
Nilai duplikat dapat mempengaruhi hasil analisis, sehingga langkah ini penting dilakukan untuk memastikan data yang unik dan bebas dari redundansi. Tahapan dalam identifikasi duplikat meliputi memuat data dengan benar, memeriksa struktur data, dan menghapus nilai duplikat jika ditemukan. Berdasarkan pemeriksaan manual, tidak ada nilai duplikat dalam *dataset* tuberkulosis, dengan memastikan data yang digunakan adalah unik dan setiap entri mempresentasikan kejadian atau entitas yang berbeda, sehingga analisis yang dilakukan akurat dan representatif tanpa bias atau distorsi.

c. Pemeriksaan Nilai *Outlier*
Outlier dapat mengganggu hasil analisis, diidentifikasi menggunakan metode IQR (*Interquartile Range*). Berdasarkan nilai yang ditemukan adalah -480.375 dan 1240.625 untuk masing-masing batas atas dan batas bawah. Nilai batas bawah tidak relevan karena negatif, sedangkan tidak ada data pasien yang melebihi batas atas. Dengan demikian, *dataset* tidak mengandung nilai *outlier* yang signifikan, memastikan data tidak memiliki nilai ekstrem yang dapat mengganggu analisis.

3.4. *Data Transformation*

Transformasi data merupakan langkah penting sebelum menerapkan algoritma klasterisasi seperti DBSCAN. Normalisasi menggunakan *Min-Max Scaler* memastikan semua fitur berada dalam rentang 0 hingga 1, dengan menghindari dominasi fitur dengan skala lebih besar. Misalnya, pada data pasien baru bakteriologis paru dinormalisasi dengan nilai $Min = 62$ dan $Max = 844$, sehingga nilai-nilai asli diubah menjadi rentang yang seragam. Dengan normalisasi, nilai-nilai ekstrem dapat terdeteksi sebagai *outlier* dan analisis data menjadi lebih konsisten.

PCA (*Principal Component Analysis)* juga digunakan untuk mengurangi dimensi data dengan mempertahankan variansi terbesar pada komponen utama. PCA membantu mereduksi kompleksitas data dari 12 fitur menjadi 2 atau 4 komponen utama, yang meningkatkan efisiensi dan akurasi algoritma DBSCAN. Komponen utama yang dipilih mempertahankan fitur dengan kontribusi signifikan terhadap variabilitas data, memastikan informasi penting tetap terjaga dalam analisis.

3.5. Implementasi Algoritma DBSCAN

Seluruh data yang telah diproses akan diimplementasikan dalam algoritma DBSCAN untuk mengklasterisasi berdasarkan kepadatan, dengan parameter utama *epsilon* (ε) dan *minPts*. *Epsilon* menentukan jarak maksimum antar titik dalam satu *cluster*, sedangkan *minPts* merupakan jumlah minimum titik dalam radius *epsilon* untuk membentuk suatu *cluster*. Penentuan *epsilon* dilakukan dengan menghitung jarak antar titik data, memastikan setiap titik memiliki cukup tetangga dalam radius tersebut. Pemilihan *epsilon* yang tepat penting untuk menghindari sensitivitas terhadap *noise* dan *outlier*.

Nilai *minPts* ditetapkan berdasarkan analisis data untuk setiap jenis pasien, umumnya disarankan setidaknya dua kali jumlah dimensi data. Penyesuaian nilai *minPts* dilakukan dengan uji sensitivitas untuk memahami pengaruhnya terhadap jumlah dan ukuran *cluster*. Dengan parameter yang optimal, DBSCAN dapat mengidentifikasi pola spasial dan temporal dalam data tuberkulosis, membantu dalam penanganan dan pencegahan penyakit secara lebih efektif.

4. HASIL DAN PEMBAHASAN

4.1. *Data Preprocessing*

Terdapat beberapa tahapan penting yang dilakukan pada *data preprocessing* untuk memastikan kualitas data diolah sebelum dilanjutkan ke tahap analisis lebih lanjut.

a. Memeriksa Nilai *Null* (*Missing Value*)

```
tingkatan                          0
tahun                              0
longitude                          0
latitude                           0
pasien_baru_bakteriologis_paru     0
pasien_baru_klinis_paru            0
pasien_baru_ekstraparu             0
pasien_tidak_diketahui             0
pasien_kambuh_bakteriologis_paru   0
pasien_kambuh_klinis_paru          0
pasien_kambuh_ekstraparu           0
pasien_diobati_selain_kambuh       0
dtype: int64
Original rows: 264
Rows after dropping nulls: 264
```

Gambar 2. Pemeriksaan Nilai *Null*

Berdasarkan **Gambar 2**, tidak ditemukan nilai *null* dalam setiap atribut data tuberkulosis, seperti tingkatan, tahun, *longitude*, *latitude*, dan jenis kasus pasien yang memastikan *dataset* tetap memiliki 264 baris data. Ketiadaan nilai *null* ini menunjukkan integritas data yang tinggi, meningkatkan stabilitas dan hasil DBSCAN, serta menghemat waktu karena tidak perlu menangani nilai *null*. Jika nilai *null* ditemukan, langkah seperti penghapusan baris atau imputasi nilai perlu dilakukan untuk menjaga representativitas dan ukuran *dataset*.

b. Memeriksa Nilai Duplikat

```
Number of duplicate rows: 0
Rows after removing duplicates: 264
```

Gambar 3. Pemeriksaan Nilai Duplikat

Pemeriksaan dan penghapusan nilai duplikat adalah langkah krusial dalam *preprocessing* untuk memastikan integritas *dataset* dan keakuratan analisis. Berdasarkan **Gambar 3** menunjukkan bahwa *dataset* tuberkulosis awal tidak memiliki entri duplikat, yang berarti setiap baris data adalah unik. Ketiadaan duplikat ini menunjukkan data telah diolah dengan baik sejak awal, memastikan analisis lebih akurat dan representatif tanpa risiko bias dari data yang terduplikasi. Meskipun tidak ditemukan duplikat dalam proses ini, pemeriksaan rutin tetap penting untuk mencegah potensi masalah di masa depan dan memastikan kualitas data yang tinggi.

c. Memeriksa Nilai *Outlier*

```
Handled 0 outliers in column 'pasien_baru_bakteriologis_paru'.
Handled 0 outliers in column 'pasien_baru_klinis_paru'.
Handled 0 outliers in column 'pasien_baru_ekstraparu'.
Handled 0 outliers in column 'pasien_tidak_diketahui'.
Handled 0 outliers in column 'pasien_kambuh_bakteriologis_paru'.
Handled 0 outliers in column 'pasien_kambuh_klinis_paru'.
Handled 0 outliers in column 'pasien_kambuh_ekstraparu'.
Handled 0 outliers in column 'pasien_diobati_selain_kambuh'.
```

Gambar 4. Pemeriksaan Nilai *Outlier*

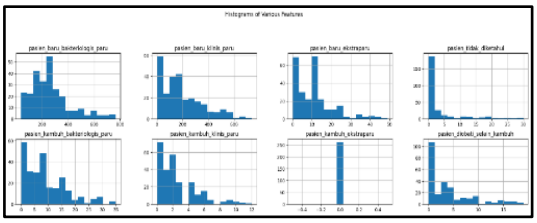
Pengelolaan nilai *outlier* merupakan langkah penting dalam analisis data untuk memastikan kualitas dan akurasi hasil. *Outlier* yang tidak ditangani dengan baik dapat mengganggu hasil statistik dan pemodelan, memberikan gambaran yang salah tentang data. Pada penelitian ini, metode penggantian *outlier* dengan nilai median (IQR) diterapkan pada *dataset* tuberkulosis, kecuali pada atribut tingkatan, *longitude*, dan *latitude*. Berdasarkan **Gambar 4** ditemukan bahwa tidak ada *outlier* yang terdeteksi pada kolom jumlah total kasus tipe pasien, menunjukkan distribusi data yang konsisten tanpa nilai ekstrem. Dengan tidak adanya *outlier*, analisis dapat dilanjutkan tanpa penyesuaian lebih lanjut.

4.2. *Data Transformation*

Terdapat beberapa tahapan penting yang dilakukan pada *data transformation* untuk memastikan kualitas data diolah sebelum dilanjutkan ke tahap analisis lebih lanjut.

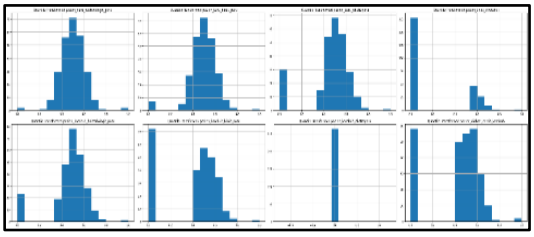
a. Normalisasi Data

Normalisasi data dengan *Min-Max Scaler* digunakan untuk mengubah distribusi data menjadi rentang 0 hingga 1, memperbaiki asumsi model yang memerlukan distribusi normal.



Gambar 4. Bentuk Data Sebelum Transformasi

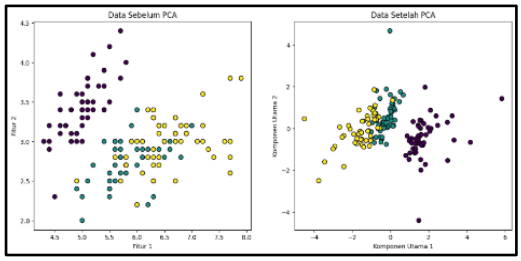
Berdasarkan **Gambar 4**, histogram atribut *dataset* tuberkulosis menunjukkan distribusi yang miring ke kanan dengan mayoritas nilai rendah pada atribut seperti pasien baru bakteriologis paru, pasien baru klinis paru, dan pasien kambuh bakteriologis paru. Beberapa atribut memiliki variasi yang sangat sedikit, seperti pasien kambuh ekstra paru. Secara keseluruhan, distribusi data awal menunjukkan mayoritas pasien memiliki nilai rendah dalam berbagai kategori, dengan frekuensi yang menurun seiring meningkatnya nilai.



Gambar 5. Bentuk Data Setelah Ditransformasi

Setelah dilakukan transformasi, distribusi data pada atribut tipe pasien pada *dataset* tuberkulosis menjadi lebih simetris dan terpusat dalam rentang 0 hingga 1, seperti yang ditunjukkan ada **Gambar 5.** Pada sebagian besar atribut, seperti pasien baru bakteriologis paru, menunjukkan distribusi simetris dengan puncak di sekitar nilai 0.5. Beberapa atribut lain, seperti pasien tidak diketahui dan pasien kambuh ekstraparu, menunjukkan distribusi yang sangat miring ke kiri dengan banyak data pada nilai 0. Transformasi ini meningkatkan kesimetrisan distribusi data, mempermudah analisis statistik lebih lanjut, dan penerapan algoritma DBSCAN pada *dataset*.

b. *Principal Component Analysis* (PCA)
PCA merupakan teknik efektif untuk mereduksi dimensi data dengan menangkap variasi terbesar dalam data ke dalam beberapa komponen utama. Sebelum diterapkan PCA, data tuberkulosis terdiri dari sumbu x dan sumbu y yang mewakili atribut asli dalam *dataset*.



Gambar 6. Penerapan PCA

Pada **Gambar 6**, visualisasi awal menunjukkan penyebaran data berdasarkan dua atribut asli, tetapi penyebaran tidak optimal karena distribusi atribut yang tidak merata dan korelasi yang rumit antar atribut. PCA membantu dengan mengidentifikasi komponen utama yang menangkap variasi terbesar, menyederhanakan struktur data, dan meningkatkan pemahaman pola yang mendasari. Pada grafik hasil PCA, data ditampilkan dalam ruang komponen utama dengan komponen utama 1 dan 2 sebagai sumbu, memberikan gambaran lebih jelas tentang distribusi data setelah ditransformasikan.

```
Number of PCA components: 2
PCA Components:
Component 1:
Top features contributing to this component:
pasien_baru_klinis_paru: 0.747
pasien_baru_bakteriologis_paru: 0.664
pasien_kambuh_bakteriologis_paru: 0.023
pasien_baru_ekstraparu: 0.018
pasien_diobati_selain_kambuh: 0.009

Component 2:
Top features contributing to this component:
pasien_baru_bakteriologis_paru: 0.747
pasien_baru_klinis_paru: -0.664
pasien_baru_ekstraparu: -0.010
pasien_kambuh_klinis_paru: -0.004
pasien_diobati_selain_kambuh: -0.003
```

Gambar 7. Komponen PCA

Berdasarkan **Gambar 7**, data tuberkulosis divisualisasikan dalam dua dimensi menggunakan dua komponen utama hasil PCA, yang menangkap sebagian besar variansi dalam *dataset*. Komponen utama 1 (sumbu x) didominasi oleh variabel pasien baru klinis paru dan pasien baru bakteriologis paru, sementara komponen utama 2 (sumbu y) juga didominasi oleh variabel yang sama, menunjukkan bahwa variasi dalam jumlah pasien baru menjadi faktor dominan. Penerapan PCA ini membantu mereduksi dimensi data, dengan mempertahankan informasi variansi yang relevan, mempermudah visualisasi, dan mengidentifikasi *cluster* yang lebih jelas terstruktur dalam *dataset* yang kompleks.

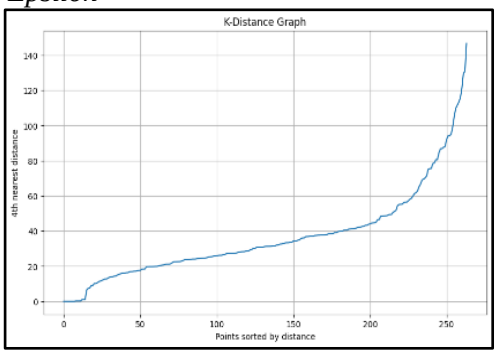
4.3. Implementasi DBSCAN

DBSCAN (*Density-Based Spatial Clustering of Applications with* Noise) adalah algoritma *clustering* berbasis kepadatan yang efektif dalam menangani *outliers* dan tidak memerlukan jumlah *cluster* yang ditentukan di awal. Dalam analisis data tuberkulosis, sebelum membentuk *cluster*, data kategori diubah menjadi variabel *dummy* untuk memenuhi kebutuhan DBSCAN akan data numerik dan menghindari asumsi urutan atau jarak antar kategori.

4.3.1. Parameter DBSCAN

DBSCAN menggunakan parameter *epsilon* dan *minPts* untuk menentukan *cluster*. Titik data dengan tetangga dalam radius *epsilon* yang memenuhi syarat minimum *points* akan membentuk *cluster*, sementara titik yang tidak memenuhi kriteria dianggap sebagai *noise* atau *outlier*.

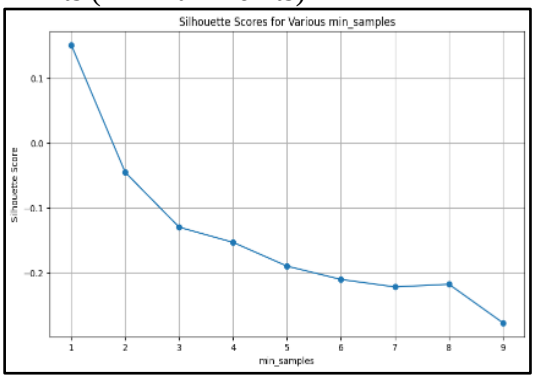
a. *Epsilon*



Gambar 8. Pencarian *Epsilon*

Berdasarkan **Gambar** 8, pencarian nilai *epsilon* menggunakan metode siku (*elbow*) menunjukkan bahwa siku terbentuk sekitar jarak ke-4 tetangga terdekat pada nilai sekitar 200. Setelah titik ini, jarak meningkat tajam, menandakan peralihan ke area dengan densitas rendah atau *outlier*. Nilai *epsilon* optimal dipilih sedikit lebih besar dari titik siku, dengan rentang 30-50 sebagai opsi yang baik untuk menangkap *cluster* utama dan menghindari terlalu banyak *outlier*.

b. *minPts* (*Minimum Points*)

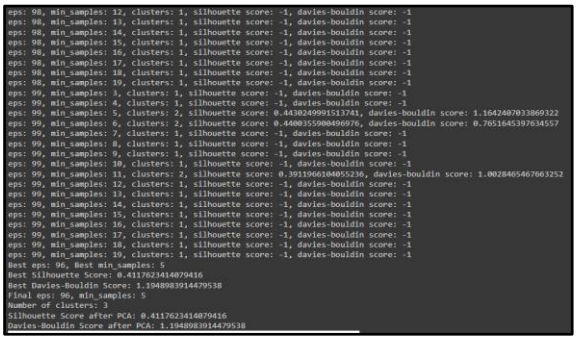


Gambar 9. Pencarian *minPts*

Berdasarkan **Gambar 9**, pemilihan nilai *minPts* menunjukkan bahwa nilai *Silhouette Coefficient* tertinggi sekitar 0.1 saat *minPts* sangat rendah, tetapi menurun tajam setelah *minPts* meningkat dari 1 ke 2. Grafik menunjukkan tren penurunan *Silhouette Coefficient* dengan meningkatnya *minimum points*, mengindikasikan bahwa nilai *minPts* yang lebih tinggi cenderung mengurangi efektivitas *clustering*.

4.3.2. Optimasi dan Evaluasi

Optimasi parameter DBSCAN menggunakan teknik *Grid Search* untuk menemukan kombinasi terbaik dari nilai *epsilon* dan *minPts*, dengan mengevaluasi hasil *clustering* menggunakan *Silhouette Coefficient* dan *Davies-Bouldin Index*.



Gambar 10. Hasil Optimasi Parameter

Berdasarkan **Gambar 10**, hasil optimasi parameter DBSCAN menunjukkan nilai *epsilon* terbaik 96 dan *minPts* terbaik 5. Nilai *Silhouette Score* 0.41176 dan *Davies-Bouldin Index* 1.194 kualitas pengelompokan menunjukkan adanya ruang perbaikan, meskipun *cluster* yang terbentuk cukup baik. Algoritma mengidentifikasi tiga *cluster* utama, sesuai dengan analisis data tuberkulosis.

4.3.3. DBSCAN *Clustering*

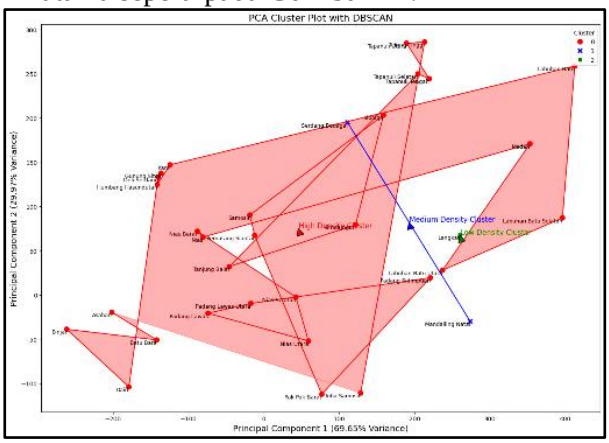
DBSCAN diimplementasikan untuk *clustering* data tuberkulosis dengan parameter optimal *epsilon* dan *minPts*, ditentukan melalui teknik *Elbow* dan *Silhouette*. PCA diterapkan untuk mereduksi dimensi, menghasilkan tiga *cluster* utama (tinggi, sedang, dan rendah). Kualitas *clustering* dievaluasi dengan *Silhouette Coefficient* dan *Davies-Bouldin Index*.

a. DBSCAN *Clustering* Tanpa PCA

DBSCAN diterapkan pada data tuberkulosis tanpa PCA menghasilkan 155 *cluster*, jauh dari tujuan tiga *cluster* utama. Pemilihan parameter *epsilon* dan *minPts* menunjukkan nilai terbaik, namun *clustering* menghasilkan *Silhouette Coefficient* rendah 0.151 dan *Davies-Bouldin Index* rendah

0.266, mengindikasikan *overlap* dan pemisahan *cluster* yang buruk. Tingginya dimensi data dan *noise* menyebabkan banyak *cluster* kecil dan *outlier*.

b. DBSCAN *Clustering* Dengan PCA
DBSCAN diterapkan pada data tuberkulosis yang telah direduksi dimensinya menggunakan PCA, menghasilkan visualisasi *clustering* dalam dua dimensi utama seperti pada **Gambar 11**.



Gambar 11. Visualisasi *Cluster* PCA

PCA mengurangi kompleksitas data berdimensi tinggi menjadi dua komponen utama, dengan sumbu x (PCA 1) menyumbang 69.65% variansi dan sumbu y (PCA 2) menyumbang 29.97%. *Cluster* yang terbentuk mencakup tiga kategori utama, *Cluster* 0 dengan kepadatan tinggi, *Cluster* 1 dengan kepadatan sedang, dan *Cluster* 3 dengan kepadatan rendah.

Rentang nilai untuk setiap *cluster* berdasarkan PCA menunjukkan variasi karakteristik di dalam *cluster* sesuai dengan **Tabel 3**. *Cluster* 0 memiliki rentang nilai yang luas pada PCA 1 dan PCA 2, menunjukkan distribusi yang variatif. *Cluster* 1 dengan rentang nilai yang lebih sempit dan rata-rata positif, menunjukkan data yang lebih terkonsentrasi. *Cluster* 2 memiliki rentang nilai positif pada PCA 1 dan nilai rata-rata yang terdistribusi di sekitar sumbu horizontal pada PCA 2, menunjukkan distribusi yang lebih luas.

Tabel 3. Rentang nilai *cluster*

Cluster	PCA	*Min*	*Max*	*Mean*
0	PCA 1	-287,36	368,82	-30,74
	PCA 2	-240,21	290,1	-1,43
1	PCA 1	353,26	424,72	387,88
	PCA 2	170,83	343,97	246,12
2	PCA 1	374,83	477,62	401.94
	PCA 2	-49.5	87	25.79

Tabel 4. Anggota *cluster*

Tingkat Sebaran	Anggota *Cluster*	Jumlah *Cluster*
Tinggi	Medan, Labuhan Batu, Labuhan Batu Selatan, Labuhan Batu Utara, Padang Sidimpuan, Toba Samosir, Pak Pak Barat, Nias Utara, Nias Selatan, Padang Lawas Utara, Padang Lawas, Batu Bara, Binjai, Asahan, Dairi, Samosir, Tanjung Balai, Nias Barat, Nias, Pematang Siantar, Simalungun, Karo, Gunung Sitoli, Deli Serdang, Humbang Hasundutan, Sibolga, Tapanuli Selatan, Tapanuli Tengah, Tapanuli Utara, Tebing Tinggi	30
Sedang	Serdang Bedagai	2
Rendah	Langkat	1

Evaluasi *clustering* dengan PCA menunjukkan *Silhouette Coefficient* sebesar 0.432 menandakan data cukup terkelompok namun masih ada area perbatasan dan *Davies-Bouldin Index* sebesar 0.491 menunjukkan *cluster* cukup terpisah. DBSCAN dengan PCA berhasil mengidentifikasi *cluster* dengan tingkat penyebaran tinggi, sedang, dan rendah pada tuberkulosis di Provinsi Sumatera Utara.

5. KESIMPULAN

Metode DBSCAN yang dikombinasikan dengan PCA berhasil mengidentifikasi tiga *cluster* utama, yaitu *cluster* tinggi, *cluster* sedang, *cluster* rendah. Hasil *clustering* dengan PCA ditunjukkan dengan nilai *Silhouette Coefficient* sebesar 0.432 dan *Davies-Bouldin Index* sebesar 0.491, yang mengindikasikan bahwa *cluster* yang dihasilkan memiliki pemisahan yang cukup baik.

Sebaliknya, ketika DBSCAN diterapkan tanpa menggunakan PCA, algoritma menghasilkan 155 *cluster* dengan nilai *Silhouette Coefficient* sebesar 0.151 dan *Davies-Bouldin Index* sebesar 0.266. Nilai-nilai ini menunjukkan bahwa kualitas *clustering* yang dihasilkan jauh lebih buruk dibandingkan dengan hasil *clustering* yang menggunakan PCA, mengindikasikan pemisahan yang kurang jelas antara *cluster*.

DAFTAR PUSTAKA

[1] J. Biologi, F. Sains, D. Teknologi, A. Makassar, and K. Mar'iyah, "Patofisiologi Penyakit Infeksi Tuberkulosis," 2021.

[2] E. Hariadi, E. Buston, N. Nugroho, and P. Efendi, "Stigma Masyarakat Terhadap Penyakit Tuberkulosis Dengan Penemuan Kasus Tuberkulosis BTA Positif Di Kota Bengkulu Tahun 2022," *2023*.

[3] Dinas Kesehatan Provinsi Sumatera Utara, "TBC Report 2022 of North Sumatera," 2022.

[4] Y. Puspita Sari *et al.*, "Implementasi Algoritma K-Means untuk Clustering Penyebaran Tuberkulosis di Kabupaten Karawang," vol. 5, no. 2, p. 2020.

[5] A. W. Sulistyo, "Implementasi Metode Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Dalam Mengelompokkan Penyebaran Tuberkulosis," Aug. 2020.

[6] Andreas. C. Muller and S. Guido, *Introduction to Machine Learning with Python*. O'Reilly Media, Inc, 2017.

[7] P. Yuli Utami, S. Agustian Hudjimartsu, T. Aurilia Viona, H. Sharfina, J. A. Yani No, and K. Barat, "Optimasi Parameter Algoritma DBSCAN untuk Mendeteksi Titik Panas Kebakaran Hutan dan Lahan," *Jurnal Edukasi dan Penelitian Informatika*, vol. 9, 2023.

[8] J. Ha, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*. Elsevier, 2011. doi: 10.1016/C2009-0-61819-5.

[9] R. Efendi, A. Junaidi, and A. M. Rizki, "Penentuan Pusat Klaster Secara Otomatis Pada Algoritma Density Peaks Clustering Berbasis Metode Inter Quartile Range," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4997.

[10] I. R. Management Association, Ed., *Machine Learning: Concepts, Methodologies, Tools and Applications*. IGI Global, 2012. doi: 10.4018/978-1-60960-818-7.

[11] V. V. Starovoitov and Yu. I. Golub, "Data Normalization in Machine Learning," *Informatics*, vol. 18, no. 3, pp. 83–96, Sep. 2021, doi: 10.37661/1816-0301-2021-18-3-83-96.

[12] E. Retnoningsih and R. Pramudita, "Mengenal Machine Learning dengan Teknik Supervised dan Unsupervised Learning Menggunakan Python," *Bina Insan ICT Journal*, vol. 7, no. 2, p. 156, Dec. 2020, doi: 10.51211/biict.v7i2.1422.

[13] D. Deng, "DBSCAN Clustering Algorithm Based on Density," in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, IEEE, Sep. 2020, pp. 949–953. doi: 10.1109/IFEEA51475.2020.00199.

[14] E. Hopkins, "Machine Learning Tools, Algorithms, dan Techniques," *Journal of Self-Governance and Management Economics*, vol. 10, no. 1, pp. 43–55, 2022.

[15] O. Maimon and L. Rokach, "Data Mining and Knowledge Discovery Handbook (Second Edition)," 2010.