**LAPORAN KOMPREHENSIF   
SDG 10 BERKURANGNYA KESENJANGAN   
DI MEDAN, SUMATERA UTARA,   
DAN INDONESIA**



**JOVAN TORIO (03082220004**)

**FRONTIER TECHNOLOGY  
PROGRAM STUDI INFORMATIKA  
FAKULTAS TEKNOLOGI INFORMASI  
UNIVERSITAS PELITA HARAPAN  
KAMPUS MEDAN   
2025**

# DAFTAR ISI

[DAFTAR ISI i](#_Toc196668368)

[DAFTAR GAMBAR ii](#_Toc196668369)

[BAB I DEFINISI MASALAH 1](#_Toc196668370)

[1.1. Latar Belakang 1](#_Toc196668371)

[1.2. Rumusan Masalah 1](#_Toc196668372)

[1.3. Formulasi Masalah 1](#_Toc196668373)

[1.4. Justifikasi Pemilihan Pendekatan 2](#_Toc196668374)

[BAB II DATA PREPARATION 3](#_Toc196668375)

[2.1. Sumber Data 3](#_Toc196668376)

[2.2. Pra-pemrosesan Data 3](#_Toc196668377)

[2.3. Selesksi Fitur 5](#_Toc196668378)

[2.4. Kesesuaian Data terhadap Metode Clustering 5](#_Toc196668379)

[BAB III MODEL IMPLEMENTATION 6](#_Toc196668380)

[3.1. Pemilihan Model dan Metode Eksperimen 6](#_Toc196668381)

[3.2. Implementasi Model 7](#_Toc196668382)

[BAB IV INTERPRETATION & SDG INSIGHTS 9](#_Toc196668383)

[4.1. Interpretasi Hasil Clustering 9](#_Toc196668384)

[4.2. SDG Insights 10](#_Toc196668385)

# DAFTAR GAMBAR

[Gambar 2.1. Restrukturisasi Data 5](#_Toc196668278)

[Gambar 2.2. Penggabungan dan Pembersihan Data 6](#_Toc196668279)

[Gambar 2.3. Imputasi Nilai Kosong 6](#_Toc196668280)

[Gambar 3.1. Perhitungan Silhouette Score 8](#_Toc196668293)

[Gambar 3.2. Hasil Perhitungan Silhouette Score 9](#_Toc196668294)

[Gambar 3.3. Implementasi Model 10](#_Toc196668295)

[Gambar 4.1. Visualisasi Hasil Clustering 11](#_Toc196668312)

# BAB I DEFINISI MASALAH

## Latar Belakang

Sustainable Development Goals (SDGs) adalah 17 agenda internasional yang diadopsi oleh anggota PBB untuk mendorong perubahan berdasarkan hak asasi manusia dan kesetaraan pembangunan sosial, ekonomi, dan lingkungan hidup. Di Indonesia, pelaksanaan SDGs atau Tujuan Pembangunan Berkelanjutan dikoordinasikan oleh Kementerian PPN/Bappenas. <https://lestari.kompas.com/read/2023/05/16/153000886/mengenal-tujuan-10-sdgs--berkurangnya-kesenjangan>

Salah satu dari 17 agenda tersebut adalah SDG 10: Pengurangan Ketimpangan. Tujuannya adalah untuk mengurangi kesenjangan dalam hal pendapatan, kesempatan, dan representasi. Indonesia, negara yang memiliki keragaman geografis, ekonomi, dan sosial yang luas, melihat agenda khusus ini sebagai tantangan yang mendesak. Hal ini didukung oleh Maria Lauranti, Direktur Oxfam, yang menyatakan bahwa ketimpangan yang ekstrim terjadi di negara berkembang seperti Indonesia. <https://kbr.id/berita/ragam/laporan-oxfam-2025-memusatkan-perhatian-pada-ketimpangan-untuk-mencapai-sdgs>

Oleh karena itu, penelitian ini dilakukan untuk menganalisis dua indikator SDG 10, yaitu Rasio Gini dan Persentase Penduduk yang hidup di bawah garis kemiskinan nasional. https://sdgs.bappenas.go.id/metadata-indikator-sdgs/ untuk mendapatkan gambaran mengenai pola ketimpangan di setiap daerah.

## Rumusan Masalah

Penelitian ini bertujuan untuk menganalisis hubungan dan pola antara dua indikator SDG 10, yaitu rasio Gini, sebuah rasio yang mengukur distribusi pendapatan, dan tingkat kemiskinan, di skala Medan, Sumatera Utara, dan Indonesia. Penelitian ini bertujuan untuk memberikan wawasan tentang apa yang harus diprioritaskan Indonesia untuk mencapai SDG 10. Adapun rumusan masalah dari penelitian ini yaitu “Bagaimana pengelompokan provinsi berdasarkan rasio Gini dan tingkat kemiskinan?”

## Kriteria Jenis Masalah

Untuk menentukan metode analisis yang tepat, data yang dikumpulkan dan tujuan penelitian harus selaras dengan metode yang tersedia, yaitu regresi, klasifikasi, *clustering*.

Pendekatan pertama, yaitu regresi, dirancang untuk memprediksi variabel target berdasarkan fitur input. Regresi digunakan untuk mencari pola antar atribut guna memprediksi nilai numerik tertentu, misalnya ketimpangan suatu daerah di beberapa tahun kedepannya.

Sementara itu, jenis permasalahan klasifikasi membutuhkan label kategorikal yang telah ditentukan sebelumnya. Contohnya adalah wilayah-wilayah yang dikumpulkan harus berlabel “ketimpangan tinggi” atau “ketimpangan rendah”. Fitur input dan label ini kemudian digunakan untuk melatih model *machine learning* yang mampu mengklasifikasikan daerah berdasarkan nilai fitur yang diberikan.

Terakhir adalah *clustering* atau pengelompokkan. Metode ini mengidentifikasi pengelompokan alami dalam data tanpa memerlukan variabel target atau label, sehingga dapat dicari suatu kemiripan antar anggota kelompok-kelompok tersebut.

## Justifikasi Pemilihan Metode

Terdapat beberapa pertimbangan yang perlu dilakukan sebelum memilih yang tepat, yaitu:

* *Dataset* yang dikumpulkan dan diproses berupa dua fitur numerik, yaitu rasio Gini dan tingkat kemiskinan daerah, tanpa variabel target atau y. Tidak ada juga label kategorikal.
* Tujuan dari penelitian ini adalah untuk mencari tahu bagaimana pengelompokan provinsi berdasarkan rasio Gini dan tingkat kemiskinan.

Dengan demikian, *clustering* dipilih sebagai metode utama karena selaras dengan struktur data dan tujuan penelitian. *Dataset* yang tidak ada variabel target atau label kategorikal mengartikan bahwa masalah ini merupakan jenis masalah *unsupervised learning*. Selain itu, dengan mengelompokkan provinsi-provinsi dan daerah-daerah yang memiliki profil ketimpangan yang serupa berdasarkan rasio Gini dan tingkat kemiskinan, dapat dilihat daerah-daerah yang membutuhkan perhatian, yang bertumbuh secara merata, dan yang terlalu difokuskan. Sehingga dapat mendukung target SDG 10.

# BAB II DATA PREPARATION

## Sumber Data

Data yang digunakan dalam penelitian ini berasal dari:

* Badan Pusat Statistik (BPS), yang telah berkontribusi dalam menyajikan data persentase penduduk miskin (P0) dan rasio Gini menurut provinsi dan daerah dari tahun 2018 hingga 2024.
* Badan Pusat Statistik Provinsi Sumatera, yang telah berkontribusi dalam menyajikan data rasio Gini Sumatera Utara menurut kabupaten/kota dari tahun 2018 hingga 2024.

Dari sumber yang telah disebutkan, diperoleh 13 dokumen *.csv* yang akan dibersihkan, diproses, dan digabung menjadi 1 (satu) *dataset*.

## Pra-pemrosesan Data

Beberapa tahapan pra-pemrosesan data yang dilakukan meliputi:

### MemuatData

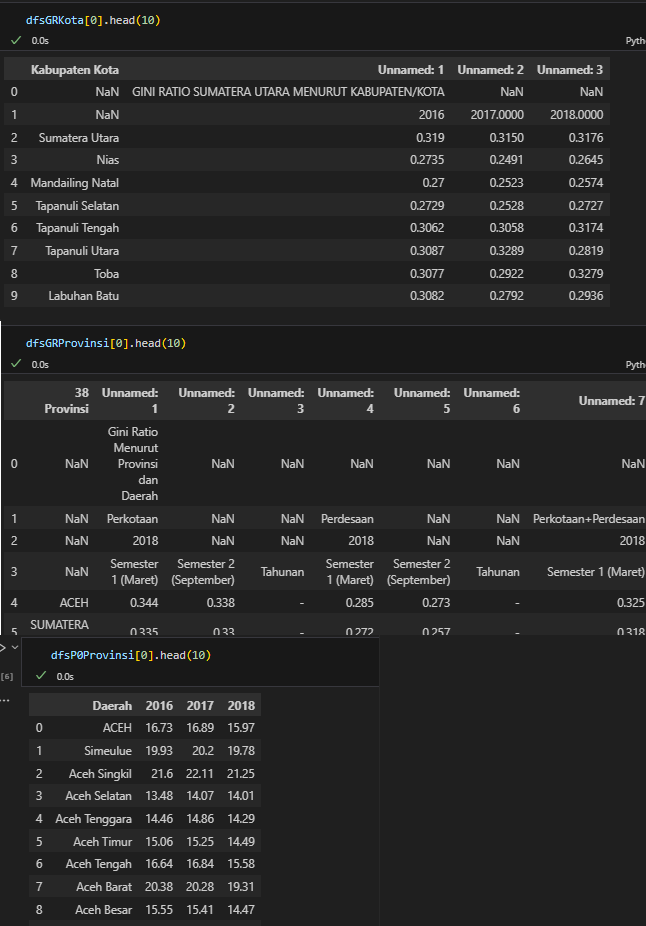
Pertama, data dimuat menggunakan pustaka panda dan os untuk memasukkan data yang berkaitan ke dalam sebuah list.

A screen shot of a computer program

AI-generated content may be incorrect.

Gambar 0.1Fungsi dan kode untuk memuat data

Berikut adalah tampilan data yang perlu dibersihkan:



Gambar 0.2 Data mentah

Dapat dilihat bahwa data mentah sangat tidak rapi, *header* yang tidak beraturan, kolom yang diisi dengan NaN atau “-“, dan masalah lain.

### Pembersihan Data

Pembersihan data akan meliputi beberapa proses, yaitu pembersihan data rasio Gini daerah Sumatera Utara (termasuk Medan), data rasio Gini provinsi di Indonesia, dan tingkat kemiskinan kota dan provinsi di Indonesia.

Untuk data rasio Gini daerah Sumatera Utama, dilakukan beberapa hal seperti pengaturan *header* yang tepat, penghapusan baris yang tidak relevan, penamaan ulang *header*, konversi tipe kolom menjadi int (dari ‘2018.0’ menjadi ‘2018’), penghapusan kolom 2016 dan 2017 karena data yang dibutuhkan adalah tahun 2018 hingga 2024, dan penggabungan semua data tersebut menjadi 1 (satu) *dataframe*.

A screenshot of a computer program

AI-generated content may be incorrect.

Gambar 0.3 *Data Cleaning* rasio Gini skala Kota di Sumatera Utara.

Selanjutnya adalah *data cleaning* untuk rasio Gini skala Provinsi di Indonesia. Pada tahap ini, dilakukan pengaturan *header*, penghapusan baris yang tidak relevan, penggantian “-“ dengan NaN, pengambilan kolom yang relevan, konversi kolom ke bentuk numerik, dan penggantian nama kolom.

Setelah itu, dilakukan imputasi dengan cara pengisian kolom. Terdapat 7 (tujuh) *file,* sehingga dibuat 7 kolom baru yang bernama ‘2018’ hingga ‘2024’ untuk mengisi nilai rasio Gini. Alasan hal ini dilakukan adalah karena *dataset* yang disediakan mengosongkan nilai rasio Gini tahunan. Nilai rasio Gini tahunan dihitung berdasarkan rata-rata dari nilai rasio Gini bulanan yang tersedia. Terakhir, 7 (tujuh) *dataframe* ini digabung menjadi satu.

A screenshot of a computer program

AI-generated content may be incorrect.

Gambar 0.4 *Data cleaning rasio Gini* skala Indonesia

Langkah terakhir untuk membentuk *dataset* rasio Gini adalah menggabungkan *dataset* rasio Gini tingkat Medan, Sumatera Utara, dan Indonesia.

A screenshot of a computer

AI-generated content may be incorrect.

Gambar 0.5 Penggabungan *dataset* rasio Gini menjadi satu

Tahap selanjutnya adalah *data cleaning dataset* tingkat kemiskinan pada skala Medan, Sumatera Utara, dan Indonesia. Pada tahap ini, nilai daerah yang duplikat dihapus, nama kolom tahun dikonversikan ke tipe *integer*, dan kolom tahun 2016 dan 2017 dihapus karena tidak digunakan.

Selain itu, *dataset* yang disediakan ini mencampurkan setiap provinsi Indonesia beserta kabupaten/kotanya. Karena skala yang digunakan hanya Medan, Sumatera Utara, dan Indonesia, maka setiap kabupaten/kota dihapus kecuali kabupaten/kota Sumatera Utara. Hal ini dilakukan dengan menyimpan kabupaten/kota Sumatera Utara, dan menghapus selebihnya. Terakhir, 3 (tiga) *dataframe* tingkat kemiskinan ini digabung menjadi satu.

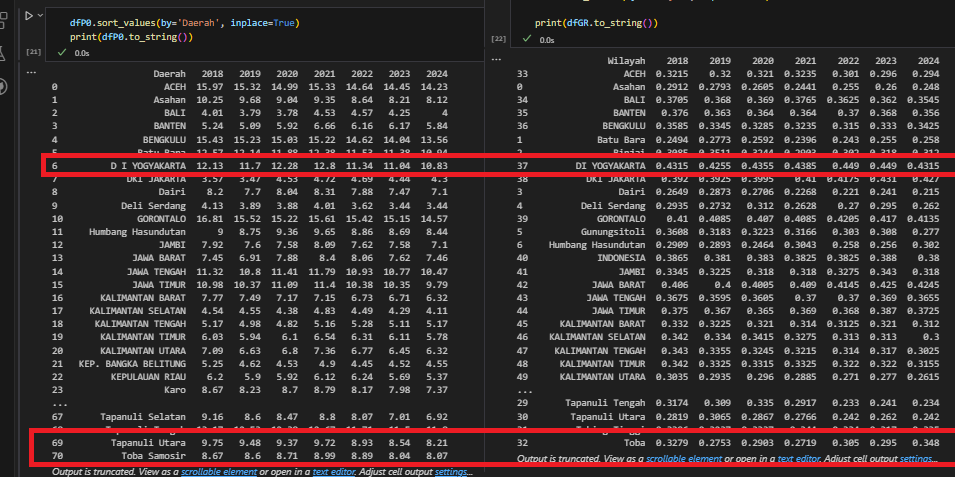
A screenshot of a computer program

AI-generated content may be incorrect.

Gambar 0.6 *Data cleaning* tingkat kemiskinan skala Indonesia.

### Penggabungan Data

Tahap terakhir dalam pra-pemrosesan data adalah penggabungan kedua *dataframe* yang telah dibuat, yaitu *dataframe* rasio Gini dan tingkat kemiskinan (P0). Pertama-tama, *dataframe* yang telah dibuat dieksplorasi.

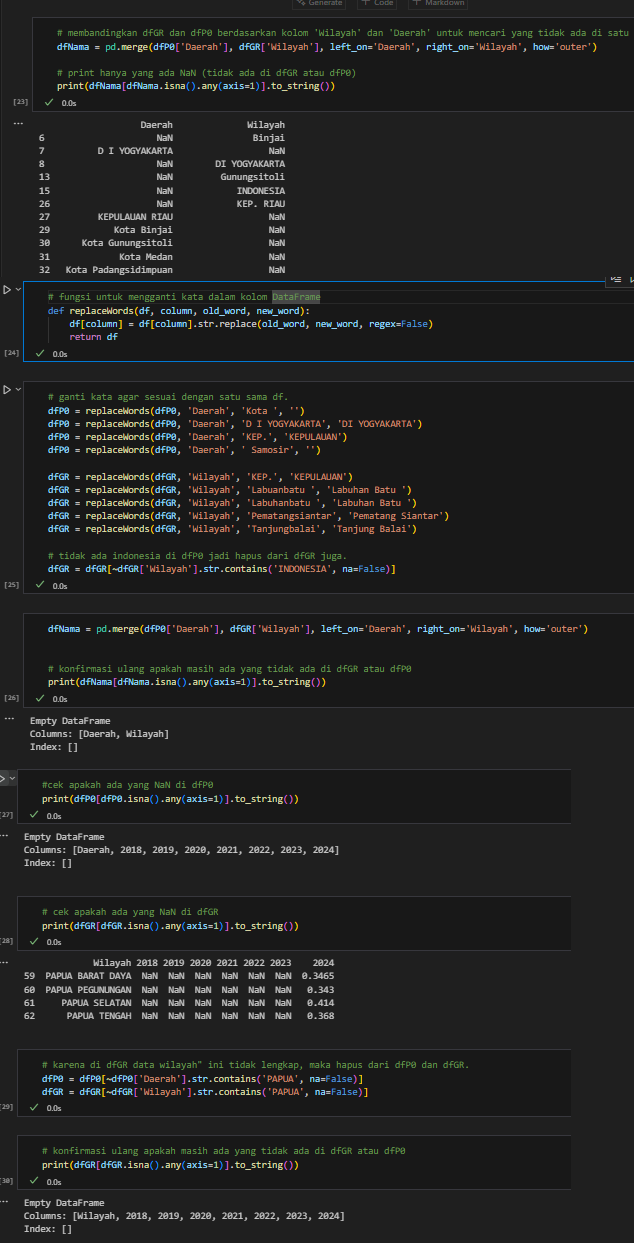


Gambar 0.7 Perbandingan *dataframe* rasio Gini dan tingkat kemiskinan

Sekilas, dapat dilihat bahwa terdapat nama-nama wilayah yang tidak sama, seperti “D I YOGYAKARTA” dan “DI YOGYAKARTA”. Jika kedua data ini digabung dengan kondisi sedemikian, nilai rasio Gini dan tingkat kemiskinan tidak akan berada di baris yang sama.

Untuk mengatasi hal tersebut, pertama harus dilihat nama wilayah apa saja yang di satu data tidak ada ekuivalennya di data yang lain. Kemudian, dengan menggunakan sebuah fungsi, nama-nama wilayah ini diperbaiki agar semua daerah dapat digabung dengan aman.

Selain itu, dicari juga wilayah yang datanya berupa NaN. Dengan data yang kurang, tidak dapat dibuat kesimpulan yang pasti untuk wilayah tersebut, dengan demikian baris tersebut sebaiknya dihapus jika data yang tersedia kurang.



Gambar 0.8 Pembersihan data rasio Gini dan tingkat kemiskinan.

Selanjutnya, data rasio Gini dan tingkat kemiskinan digabung menjadi satu berdasarkan wilayah yang sama, dengan membuat tahun menjadi 1 kolom dan nilai pada tahun tersebut di kolom yang lain. Hasil dataset tersebut di-reset indeksnya dan diurutkan secara abjad dan tahun. Terakhir, data tersebut belum tentu numerik, sehingga dikonversi ke numerik. Tingkat kemiskinan juga dibagi 100 agar berubah dari bentuk persentase menjadi desimal.

A screenshot of a computer program

AI-generated content may be incorrect.

Gambar 0.9 Penggabungan dan Pra-pemrosesan terakhir *dataset*.

## Kesesuaian Data terhadap Metode *Clustering*

Setelah melalui proses pengumpulan dan pra-pemrosesan, *dataset* ini memenuhi kriteria yang sesuai untuk metode *clustering*, yaitu:

* Fitur yang ada bertipe numerik dan kontinu, sehingga fitur yang digunakan, yaitu tingkat kemiskinan dan rasio Gini mendukung perhitungan jarak antar data, misalnya menggunakan *Euclidean distance* atau *Cosine similarity*.
* *Datasest* tidak memiliki nilai target atau label kategorikal, sehingga cocok untuk *unsupervised learning* berbentuk *clustering*.
* *Dataset* yang telah dibersihkan tidak ada nilai kosong yang dapat mengakibatkan distorsi dalam analisis.

Karena memenuhi kriteria tersebut, *dataset* layak digunakan dalam implementasi *clustering*.

# BAB III MODEL IMPLEMENTATION

## Pemilihan Model dan Metode Eksperimen

Untuk menentukan algoritma *clustering* yang tepat, karakteristik dari *dataset* yang ada akan dirujuk pada materi yang tersedia dalam perkuliahan. Untuk memperoleh gambaran yang lebih jelas mengenai karakteristik *dataset*, dilakukan *exploratory data analysis* (EDA) dan juga eksperimen berupa *k-means* *clustering* untuk memperoleh beberapa metrik seperti *silhouette score.*

A screenshot of a computer

AI-generated content may be incorrect.

Gambar 0.1 Karakteristik *dataset*

Melalui tahap awal EDA, dilihat atribut dari *dataset* yang tersedia. Rata-rata dari tingkat kemiskinan (P0) adalah 0.1 atau sekitar 10%, sementara rasio Gini (GR) rata-rata adalah 0.3. P0 terendah dan tertinggi adalah 7% dan 26%, sementara GR terendah dan tertinggi adalah 0.19 dan 0.45. Dapat disimpulkan bahwa Indonesia belum mencapai SDG 10, karena masih terdapat ketimpangan antar daerah. Selain itu, dihitung juga jumlah data berbentuk kota di Sumatera Utara dan provinsi di Indonesia, di mana hasil menunjukkan bahwa jumlahnya cukup setara.

Selanjutnya, beberapa grafik dibuat seperti:

* Grafik *scatter plot* antara P0 dan GR yang menunjukkan bahwa kota di Sumatera Utara cenderung memiliki P0 dan GR yang rendah , dengan beberapa *outlier* yang memiliki P0 tinggi. Data berskala provinsi cenderung P0 yang rendah, dengan GR yang menengah ke atas. Dapat dilihat juga beberapa *outlier* yang memiliki GR yang sangat tinggi. Kesimpulannya adalah dari grafiktersebut terdapat ketimpangan yang tampak karena titik-titik tersebut menyebar.

A screenshot of a computer screen

AI-generated content may be incorrect.

Gambar 0.2 Grafik *scatter plot* hubungan P0 dan GR berdasarkan skala.

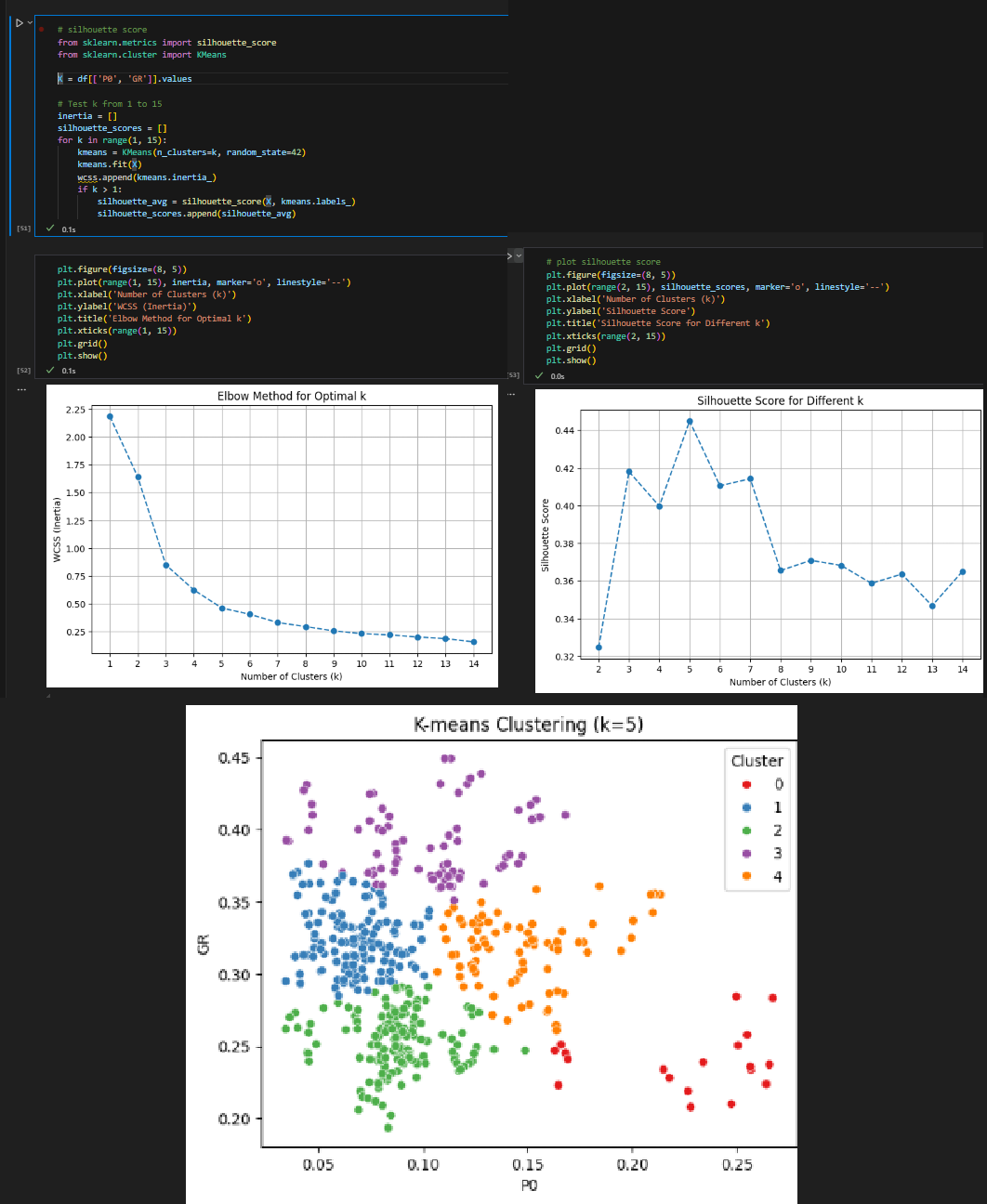
* Grafik matriks korelasi antara P0 dan GR yang menunjukkan bahwa korelasi antar fitur sangat kecil, yaitu sebesar -0.082. Hal ini mendukung pernyataan bahwa *regression* kurang cocok untuk digunakan pada *dataset* ini.

A screenshot of a computer screen

AI-generated content may be incorrect.

Gambar 0.3 Matriks korelasi antara P0 dan GR.

* Grafik *elbow method* dan *silhouette score* untuk masing-masing *k,* untuk melihat jumlah *k* yang cocok jika melakukan *clustering* dengan *k-means*. Dengan grafik, dapat dilihat bahwa *inertia* mulai *converge* setelah *k*=5. Hal ini didukung oleh *silhouette score* yang menunjukkan bahwa *k=*5 merupakan jumlah *cluster* yang optimal. Grafik *scatter plot* untuk k=5 dibuat untuk mengilustrasikan *cluster* yang telah dibuat.



Gambar 0.4 Grafik *elbow method, silhouette score,* dan *k-means clustering* dengan *k*=5.

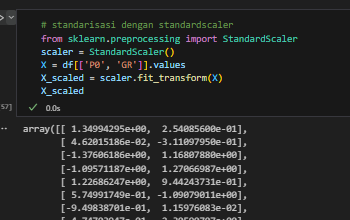
* *Dataset* ini memiliki beberapa karakteristik dengan algoritma yang cocok yaitu: *dataset* kecil (*hierarchical clustering* dan DBSCAN), dimensi rendah (cocok untuk sebagian besar algoritma), tipe data numerik (cocok untuk *k-means*, *hierarchical*, dan gaussian mixture), *cluster* berbentuk *arbitrary* (DBSCAN, OPTICS, dan Mean Shift), jumlah *cluster* tidak pasti karena *k*=5 menunjukkan *silhouette score* yang rendah (DBSCAN, *hierarchical,* OPTICS), memerlukan sensitivitas pada *outlier* (K-Medoids dan DBSCAN).

Berdasarkan uraian ini, algoritma *clustering* yang berpotensial baik dan cocok digunakan adalah *k-means* dan DBSCAN. Hasil dari kedua algoritma ini akan dibandingkan.

## Implementasi Model

Pada tahap ini, dilakukan implementasi *model clustering* menggunakan algoritma DBSCAN untuk menganalisis *dataset* tingkat kemiskinan dan rasio Gini pada skala Medan, Sumatera Utara, dan Indonesia. Langkah-langkah yang dilakukan adalah sebagai berikut:

1. Standarisasi *dataset* dengan menggunakan StandardScaler dari pustaka sklearn karena DBSCAN sensitif terhadap *outlier*.



Gambar 0.5 Standarisasi *dataset*

1. Pemilihan *hyperparameter* berupa eps (jarak maksimum antara dua sampel untuk dianggap sebagai bertetangga) dan min\_sample (jumlah sampel minimum dalam sebuah kelompok untuk dianggap sebagai sebuah *cluster*) untuk DBSCAN. Untuk min\_sample, disarankan sebesar 2\*dimensi, sehingga min\_sample yang ditetapkan adalah 4 (empat), mengingat dimensi dari *dataset* ini adalah 2 (dua).

Sementara itu, eps ditentukan menggunakan menggunakan *k-nearest neighbours* dan metode yang mirip dengan *elbow method*, di mana jumlah eps yang dipilih adalah skor jarak terurut yang paling drastis.

A screen shot of a graph

AI-generated content may be incorrect.

Gambar 0.6 Pencarian eps optimal.

Melalui metode yang telah dikerjakan, dapat dilihat bahwa jumlah eps yang optimal adalah 0.2-0.3.

1. Implementasi model *k-means* (dengan *k=*5) dan DBSCAN menggunakan pustaka sklearn.

A screenshot of a computer

AI-generated content may be incorrect.

Gambar 0.7 Perbandingan ilustrasi *c;ustering* DBSCAN dan *k-means clustering*.

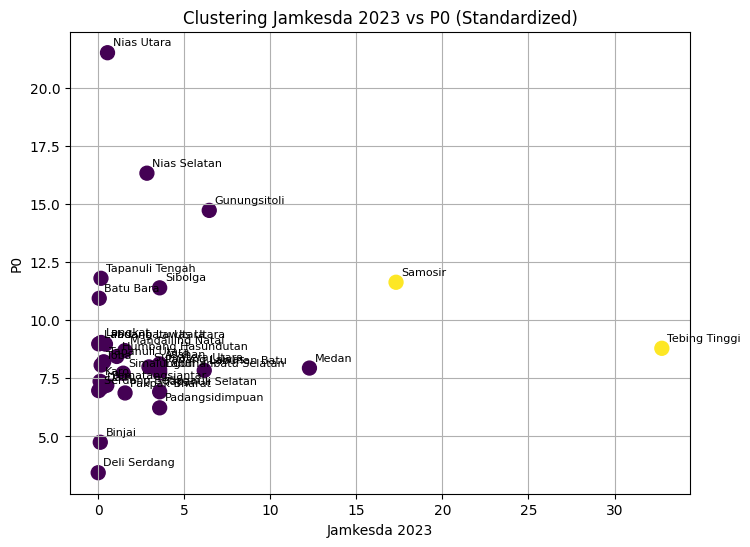
1. Evaluasi menggunakan tiga metode validasi internal yaitu *silhouette score,* indeks Davies-Bouldin*,* dan indeks Calinski-Harabasz.

A screen shot of a computer

AI-generated content may be incorrect.

Gambar 0.8 Hasil evaluasi menggunakan *silhouette score,* indeks Davies-Bouldin, dan indeks Calinski-Harabasz

# BAB IV INTERPRETATION & SDG INSIGHTS



Gambar 4.1. Visualisasi Hasil Clustering

## Interpretasi Hasil Clustering

Visualisasi hasil *clustering* antara Persentase Peserta Jamkesda tahun 2023 dan P0 (Persentase Penduduk Miskin) menunjukkan bahwa kabupaten/kota di Provinsi Sumatera Utara terbagi menjadi dua kelompok utama:

1. Cluster 1 (Warna Ungu):

Mayoritas kabupaten/kota tergabung dalam *cluster* ini. Karakteristik utamanya adalah:

* Memiliki nilai persentase Jamkesda yang relatif lebih rendah dibandingkan dengan *cluster* lain.
* Variasi P0 lebih besar, mencakup daerah dengan tingkat kemiskinan rendah hingga tinggi.
* Sebagian besar daerah memiliki keterkaitan kuat antara rendahnya cakupan Jamkesda dengan tingkat kemiskinan moderat.

1. Cluster 2 (Warna Kuning):

Terdiri dari daerah yang lebih sedikit, yaitu Samosir dan Tebing Tinggi. Karakteristik:

* Memiliki nilai persentase peserta Jamkesda yang jauh lebih tinggi dibandingkan *cluster* lainnya.
* Tingkat kemiskinan (P0) bervariasi, namun relatif lebih moderat dibandingkan daerah-daerah *cluster* ungu dengan kemiskinan ekstrem.
* Menunjukkan adanya strategi jaminan kesehatan yang sangat agresif meskipun tingkat kemiskinan tidak setinggi daerah lain.

## SDG Insights

Hasil clustering ini memberikan wawasan penting terkait pencapaian Sustainable Development Goals (SDGs), terutama:

* **SDG 1: No Poverty**

Pengelompokan daerah berdasarkan tingkat kemiskinan (P0) membantu mengidentifikasi wilayah yang masih memiliki tingkat kemiskinan tinggi. Daerah-daerah seperti Nias Utara dan Nias Selatan yang masuk dalam *cluster* dengan cakupan Jamkesda rendah dan P0 tinggi menandakan perlunya intervensi lebih lanjut untuk mengatasi kemiskinan ekstrem.

* **SDG 3: Good Health and Well-Being**

Tingginya partisipasi dalam program jaminan kesehatan daerah (Jamkesda) di wilayah seperti Tebing Tinggi dan Samosir (*cluster* kuning) menunjukkan upaya positif dalam meningkatkan akses layanan kesehatan. Pola ini menekankan bahwa perluasan cakupan jaminan kesehatan berpotensi berkontribusi terhadap kesejahteraan masyarakat dan mempercepat pencapaian target kesehatan universal.

Integrasi hasil clustering ini dapat mendukung perumusan kebijakan yang lebih tepat sasaran untuk kedua tujuan pembangunan berkelanjutan tersebut.