PENGELOMPOKAN KABUPATEN/KOTA DI SUMATERA UTARA BERDASARKAN INDIKATOR INFRASTRUKTUR DAN LAYANAN DASAR MENGGUNAKAN PRINCIPAL COMPONENT ANALYSIS DAN K-MEANS CLUSTERING

Raca Hutagalung[1*], Nuramaliyah[2]
[1]Program Studi Statistika, Universitas Terbuka, Medan, Indonesia

**Penulis korespondensi: racahtg@gmail.com*

ABSTRAK

Penelitian ini bertujuan mengelompokkan 33 kabupaten/kota di Provinsi Sumatera Utara berdasarkan indikator infrastruktur dan layanan sosial tahun 2023 untuk melihat variasi kondisi antar wilayah. Proses analisis dilakukan melalui metode *Principal Component Analysis* (PCA) untuk mereduksi dimensi, dan *K-Means Clustering* sebagai teknik pengelompokan. Dua komponen utama dengan nilai *eigen* 4,775 dan 1,034 dipertahankan karena mampu menjelaskan 72,61% keragaman total data. Evaluasi *Silhouette Coefficient* menunjukkan nilai tertinggi pada k= 2, namun konfigurasi k = 3 dipilih karena tetap berada pada kategori baik sekaligus memberikan pemisahan yang lebih stabil dan konsisten dengan struktur skor PCA, dengan iterasi berhenti pada langkah ke-3 dan menghasilkan kelompok beranggotakan 7,4, dan 22 wilayah. Hasil *clustering* menunjukkan tiga pola karakteristik, yaitu wilayah dengan pendidikan relatif baik namun infrastruktur bervariasi, wilayah dengan layanan dasar dan pendidikan yang relatif lemah, serta wilayah dengan kualitas insfrastruktur dan layanan dasar yang lebih baik dibanding klaster lainnya. Temuan ini diharapkan dapat menjadi dasar pendukung perumusan kebijakan pembangunan daerah yang lebih terarah dan berbasis data.

Kata kunci: PCA, *K-Means Clustering*, infrastruktur, layanan dasar.

1 PENDAHULUAN

Pemerataan pembangunan di wilayah Sumatera Utara masih menunjukkan perbedaan yang nyata antara kabupaten dan kota, baik dari aspek sosial, ekonomi, maupun akses layanan dasar. Variasi ini tampak pada akses masyarakat terhadap fasilitas dasar seperti listrik, air minum layak, dan sarana sanitasi, serta layanan sosial, termasuk pendidikan, kesehatan, dan kondisi ekonomi. Ketimpangan tersebut menjadi tantangan bagi perencaaan pembangunan berbasis bukti, karena memerlukan pemahaman yang sistematis mengenai kesamaan dan perbedaan karakteristik setiap wilayah.

Salah satu pendekatakan yang efektif untuk mengindentifikasi pola wilayah dengan kondisi serupa adalah analisis pengelompokan (Saputra, 2025). Melalui metode ini, kabupaten/kota dapat dikelompokkan berdasarkan kesamaan indikator infrastruktur dan layanan sosial. Algoritma *K-Means* banyak digunakan dalam konteks tersebut karena mampu menangani data numerik dengan baik dan menghasilkan kelompok yang mudah dianalisis. Salah satu keunggulan *K-Means* jika dibandingkan dengan clustering lainnya adalah hasil yang bisa direpresentasikan dengan mudah dan juga waktu komputasi yang relatif cepat (Maulana et al., 2023). Namun, hasil *K-Means* sangat bergantung pada jumlah klaster yang ditentukan di awal, serta dapat terpengaruh oleh korelasi antar indikator (Aviliana & Hendikawati, 2025).

Untuk mengatasi keterbatasan tersebut, penelitian ini menerapkan *Principal Component Analysis* (PCA) sebelum melakukan pengelompokan. Secara metodologis, penerapan PCA sebelum *K-Means* membantu mengatasi kolineritas dan mengurangi dimensionalitas sebelum pengelompokan otomatis pada dataset berdimensi tinggi (Sipayung & Hasugian, 2025). Skor

komponen yang diperoleh selanjutnya digunakan input *K-Means*, sehingga proses pengelompokan menjadi lebih efisien dan stabil. Pemilihan jumlah klaster optimal dilakukan dengan mempertimbangkan *silhouette score* > 0,5, yang menunjukkan bahwa klaster yang terbentuk memiliki kualitas pemisahan yang baik (Limbong & Likumahwa, 2025). Pendekatan gabungan antara PCA dan *K-Means* yang dioptimalkan diharapkan menghasilkan pengelompokan wilayah yang mencerminkan kondisi pembangunan daerah yang nyata di Sumatera Utara. Di tingkat nasional, penerapan *K-Means* pada data produksi dan luas panen bawang merah tahun 2021 di Kabupaten Bima menunjukkan kemampuan klasterisasi untuk menggambarkan hasil produksi bawang merah (Maulana et al., 2023).

Berdasarkan hal tersebut, penelitian ini bertujuan untuk mengelompokkan kabupaten/kota di Sumatera Utara menurut indikator infrastruktur dan layanan dasar. Dengan memanfaatkan PCA untuk mereduksi dimensi data dan *K-Means* yang di optimalkan, hasil pengelompokan diharapkan memberikan gambaran yang jelas tentang pola pembangunan antar wilayah serta dapat menjadi acuan bagi pengambilan kebijakan yang lebih tepat sasaran.

2 METODE

Pada penelitian ini, akan menggunakan metode yang dijelaskan pada Gambar 1. Analisis dilakukan dengan bantuan perangkat lunak IBM SPSS Statictics dan Aplikasi R Studio dengan alur sebagai berikut.

Gambar 1. Tahapan Proses Analisis Data

2.1 Jenis dan Sumber Data

Penelitian ini menggunakan data sekunder yang diperoleh dari situs resmi Badan Pusat Statistika (BPS) Provinsi Sumatera Utara. Periode data yang digunakan yaitu Januari hingga Desember 2023, sehingga menggambarkan kondisi terkini infrastruktur dan layanan dasar pada masa pascapandemi COVID-19. Unit analisis penelitian ini adalah seluruh kabupaten/kota di Provinsi Sumatera Utara sebanyak 33 daerah.

2.2 Karakteristik dan Variabel Penelitian

Penelitian ini menggunakan delapan indikator infrastruktur dan layanan dasar sebagai variabel bebas (X) tanpa melibatkan variabel terikat (Y). Adapun karakteristik variabel yang digunakan disajikan pada Tabel 1.

Tabel 1. Karakteristik variabel penelitian

Variabel	Keterangan
X1	Persentase akses listrik PLN
X2	Persentase akses air minum layak
X3	Persentase sanitasi layak
X4	Persentase penduduk rawat inap
X5	Indeks pembangunan manusia
X6	Angka partisipasi murni SD
X7	Persentasi pemilik fasilitas buang air besar
X8	Persentasi penduduk miskin

Sumber: Data BPS Sumatera Utara 2023

2.3 Prapemrosesan Data

Sebelum memasuki tahap analisis utama, data terlebih dahulu disajikan dalam bentuk statistik deskriptif. Penyajian ini bertujuan utnuk memperoleh gambaran umum mengenai

distribusi data serta mengindentifikasi variabel yang memiliki skala atau variansi yang tidak seimbang (Harmain et al., 2022). Pada tahap berikutnya, dilakukan proses standarisasi guna memastikan seluruh variabel numerik berada pada rentang skala yang seragam sehingga layak dianalisis pada tahap selanjutnya. Standarisasi dilakukan menggunakan metode *min-max*, di mana setiap nilai atribut numerik x_r diubah menjadi nilai terstandarisasi sesuai dengan persamaan berikut (Santoso et al., 2024).

$$x'^{\,r}_{ij} = \frac{x^r_{ij} - \min(x^r_j)}{\max\left(x^r_j\right) - \min(x^r_j)} \tag{1}$$

dengan:

$x'^{\,r}_{ij}$: Nilai atribut numerik hasil standarisasi
x^r_{ij} : Nilai awal sebelum standarisasi
$\min(x^r_j)$: Nilai terendah untuk atribut numerik ke-j
$\max\left(x^r_j\right)$: Nilai tertinggi untuk atribut numerik ke-j

2.4 Analisis Korelasi

Hubungan antarvariabel dianalisis menggunakan Persamaan 2 untuk melihat keeratan hubungan antar variabel (Anwar et al., 2022).

$$r_{a,b} = \frac{n\left(\sum_{i=1}^n z_{ia} z_{ib}\right) - \left(\sum_{i=1}^n z_{ia}\right)\left(\sum_{i=1}^n z_{ib}\right)}{\sqrt{n\left(\sum_{i=1}^n z_{ia}^2\right) - \left(\sum_{i=1}^n z_{ia}\right)^2} \cdot \sqrt{n\left(\sum_{i=1}^n z_{ib}^2\right) - \left(\sum_{i=1}^n z_{ib}\right)^2}}; a = b\ ; 1,2,3, \dots, p \tag{2}$$

dengan:

$r_{a,b}$: Nilai korelasi antara variabel ke-a dan ke-b pada data yang telah distandardisasi

2.5 Principal Component Analysis (PCA)

Analisis PCA diterapkan untuk mengekstraksi informasi utama yang relevan dari kumpulan data dan sekaligus mereduksi kemponen yang dianggap tidak signifikan atau bersifat gangguan (Suhandy & Yulia, 2021). Perhitungan dimulai dari pembentukan matriks kovarians yang selanjutnya digunakan sebagai dasar dalam menentukan nilai *eigen* dan ventor *eigen*. Penentuan nilai *eigen* dilakukan berdasarkan persamaan berikut.

$$\det(\lambda I - R) = 0 \tag{3}$$

Nilai λ yang memenuhi persamaan tersebut menjadi solusi *eigen* dari matriks kovarians. Langkah berikutnya adalah menghitung vektor *eigen* menggunakan relasi (Rosyada & Utari, 2024).

$$R\vec{v} = \lambda\vec{v} \tag{4}$$

dimana λ merupakan nilai *eigen*, $\vec{v}$ *vector eigen*, dan I adalah matriks identitas yang memiliki bentuk sebagai berikut.

$$\mathrm{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \tag{5}$$

Jumlah komponen utama yang dipertahankan ditetapkan menggunakan aturan nilai *eigen*, yaitu hanya komponen dengan nilai $\lambda \geq 1$.

2.5.1 Matriks Loading dan Pembentukan Komponen

Loading digunakan untuk menunjukkan pola struktur data dalam bentuk hubungan korelasi antarvariabel (Enzellina & Suhaedi, 2022). Setiap indikator memiliki nilai *loading* pada masing-masing komponen utama. Besaran nilai *loading* ini menunjukan kuatnya kontribusi variabel terhadap komponen tersebut, sekaligus mencerminkan sejauh mana komponen utama mampu menjelaskan keragaman dari variabel yang bersangkutan (Suhandy & Yulia, 2021). *Loading* diperoleh dari perkalian antara nila *eigen* dan akar kuadrat nilai *eigen*.

$$Loading_{jk} = v_{jk}\sqrt{\lambda_k} \tag{6}$$

dengan:

v_{jk} : elemen ke-j pada nilai *eigen* komponen utama ke-k
λ_k : nilai *eigen* untuk komponen ke-k
$Loading_{jk}$: kontribusi variabel j terhadap PC_k

Secara geometris, nilai *loading* dapat dipandang sebagai kosinus sudut antara suatu variabel dengan komponen utama yang bersangkutan. Semakin kecil sudut yang terbentuk atau dengan kata lain semakin kuat keterkaitan variabel dengan komponen tersebut, maka semakin besar nilai loading yang dihasilkan. Rentang nilai *loading* berada antara -1 hingga +1 (Suhandy & Yulia, 2021) Selanjutnya menghitung transformasi himpunan data baru hasil reduksi dengan menggunakan PCA sesuai persamaan berikut (Putri & Hayati, 2024).

$$PC_{at} = \vec{v}_{at1}Z_1 + \vec{v}_{at2}Z_2 + \cdots + \vec{v}_{ap}Z_{ap} \tag{7}$$

2.6 K-Means Clustering

K-Means merupakan salah satu teknik dalam data mining yang digunakan untuk mengelompokkan objek berdasarkan kemiripan karakteristiknya (Silitonga et al., 2024). Sehingga objek dengan pola serupa berada dalam kelompok yang sama, sedangkan objek dengan ciri yang berbeda akan tergabung dalam kelompok lain (Izzuddin & Wijayanto, 2024).

Prosedur algoritma *K-Means* diawali dengan menentukan jumlah klaster yang diinginkan serta menetapkan pusat awal, baik secara acak maupun berdasarkan aturan tertentu. Setiap objek data x_i kemudian diukur jaraknya terhadap semua pusat klaster, dan objek tersebut dimasukkan ke klaster dengan pusat terdekat. Setelah seluruh objek teralokasi, posisi pusat klaster diperbaruhi dengan menghitung rata-rata dari anggota klaster masing-masing (Sipayung & Hasugian, 2025). Tahapan penetapan dan pembaruan pusat klaster diulang hingga mencapai kondisi konvergen, ditandai dengan pusat klaster yang stabil dan tidak adanya perpindahan anggota klaster (Mardhatillah et al., 2025). Jarak *euclid* data dengan pusat klaster yang digunakan pada tahap pengelompokan menggunakan persamaan berikut (Anwar et al., 2022).

$$d(x_{a,}c_i) = \sqrt{\sum_{p=1}^{m}(x_{ap} - c_{ip})^2} \tag{8}$$

dengan:
$a = 1,2,3,\ldots,n$ dan $i = 1,2,3,\ldots,k$

Selanjutnya menghitung pusat klaster baru berdasarkan keanggotaan dengan persamaan berikut:

$$\bar{c}_{ip} = \frac{1}{n_i}\sum_{a=1}^{n_i} x_{ap} \tag{9}$$

Penilaian kualitas hasil pengelompokan dapat dilakukan melalui perhitungan nilai *silhouette coefficient*. Proses perhitungannya dilakukan melalui beberapa langkah sebagai berikut (Anwar et al., 2022):

a. Menghitung rata-rata jarak data ke-i dengan semua data pada klaster yang sama

$$a_i = \frac{1}{N_p - 1} \sum_{r=1}^{N_p - 1} d_{i,r}, r \neq i \quad (10)$$

b. Menghitung rata-rata jarak data ke-I dengan semua data pada klaster yang berbeda

$$b_i = \min\{d_i(p)\}, r \neq i \quad (11)$$

c. Menghitung nilai $SC_1(i)$

$$SC_1(i) = \frac{b_i - a_i}{\max\{b_i - a_i\}}, i = 1,2,3, \ldots, \quad (12)$$

d. Menghitung rata-rata nilai $SC_2(p)$

$$SC_2(p) = \frac{1}{N_p} \sum_{x_i \in C_p}^{N_p} SC_1(i) \quad (13)$$

e. Menghitung nilai SC global

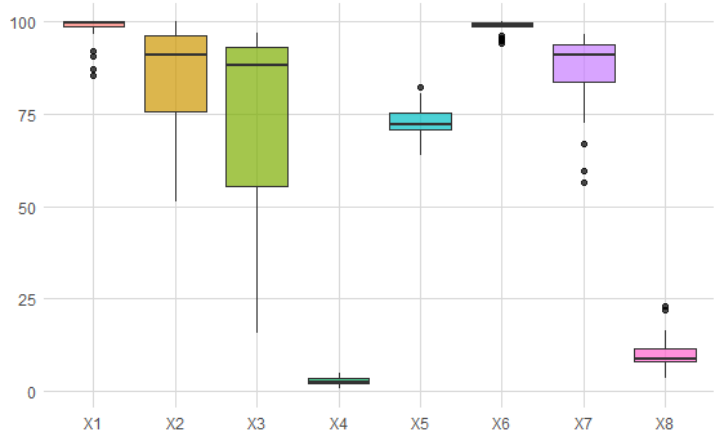
$$SC = \frac{\sum_{p=1}^{K} (N_p \times SC_2(p))}{\sum_{p=1}^{K} N_p} \quad (14)$$

f. Menentukan nilai k optimal berdasarkan nilai *silhouette coefficient* terbesar.

3 HASIL DAN PEMBAHASAN

3.1 Gambaran Umum Data

Distribusi variabel infrastruktur dan layanan dasar pada 33 kabupaten/kota di Sumatera Utara tahun 2023 disajikan pada Gambar 2 berikut.



Gambar 2. Boxplot data infrastruktur dan layanan dasar Sumatera Utara, 2023

Gambar 2 menunjukkan karakteristik sebaran data awal untuk delapan variabel (X_1-X_8) yang memiliki skala dan tingkat keragaman yang berbeda. Variabel X_1dan X_6 memperlihatkan median yang sangat tinggi dengan sebaran yang relatif sempit, menandakan data yang stabil, sedangkan X_3 menunjukkan sebaran paling luas yang mengindikasikan variabilitas data yang tinggi. Variabel X_2, X_5, dan X_7 memiliki variasi sedang dengan beberapa *outlier*, sementara X_4 dan X_8 berada pada nilai rendah, di mana X_4 bersifat sangat homogen dan X_8 masih menunjukkan adanya *outlier*. Secara umum, boxplot ini memperlihatkan ketidaksamaan skala dan keragaman data, sehingga diperlukan standarisasi sebelum analisis lanjutan.

3.2 Standardisasi (Z-Score)

Sebelum analisis PCA dilakukan, seluruh variabel dinormalkan menggunakan pendekatan *z-score* untuk menyeragamkan skala pengukuran. Metode ini memastikan bahwa variabel yang memiliki rentang besar tidak mendominasi proses ekstraksi komponen (Pratama et al., 2023). Metode ini dijelaskan dengan persamaan berikut.

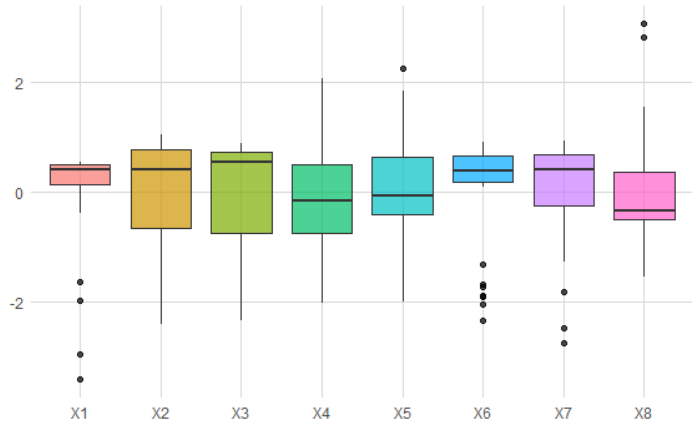
$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j} \quad (15)$$

dengan:
X_{ij} : nilai observasi ke-i pada variable ke-j
$\bar{X}_j$: rata-rata variable ke-j
S_j : simpangan baku variabel ke-j

Tabel 2. Hasil standarisasi

Data	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	-3,424	-2,414	-2,015	-1,398	-1,810	0,730	-1,277	1,234
2	-0,072	-0,444	-1,536	-0,824	-0,888	0,193	-2,491	-0,254
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
33	0,229	-2,414	-0,762	1,711	-0,399	0,3567	0,427	1,158

Transformasi ini menghasilkan matriks data terstandar Z yang menjadi dasar seluruh analisis selanjutnya. Berikut boxplot dari data hasil standarisasi pada Tabel 2.



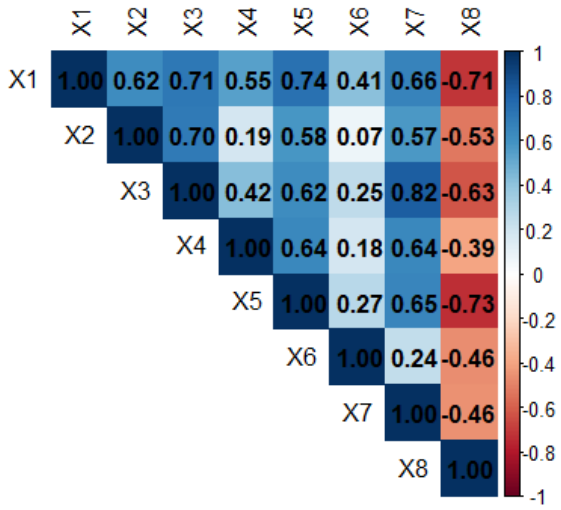
Gambar 3. Boxplot data Z infrastruktur dan layanan dasar Sumatera Utara, 2023

Gambar 3 menunjukkan karakteristik sebaran delapan variabel infrastruktur dan layanan dasar yang menjadi acuan dalam pembentukan klaster kabupaten/kota di Sumatera Utara. Variabel X_1, X_2, dan X_3 menujukkan rentang antar kuartil yang relatif lebih lebar, menandakan variasi kondisi antar kabupaten/kota cukup besar. Beberapa *outlier* pada variabel tersebut menegaskan adanya daerah dengan tingkat infrastruktur atau layanan dasar yang jauh lebih rendah dibanding mayoritas. Variabel X_4 memiliki sebaran yang lebih sempit, mengindikasikan kondisi yang cenderung seragam. Sementara X_5 memperlihatkan satu nilai ekstrem tinggi yang menunjukkan adanya daerah dengan capaian yang berada jauh diatas daerah lainnya. Variabel X_6 menunjukkan banyak *outlier* di sisi bawah, mencerminkan adanya kelompok kabupaten/kota yang tertinggal signifikan dalam indikator tersebut. Variabel X_7 dan X_8 memiliki sebaran sedang dengan variasi antar wilayah yang tidak terlalu ekstrem, meskipun tetap terdapat beberapa nilai menyimpang yang menunjukkan perbedaan kondisi tertentu antar daerah.

Secara keseluruhan, pola pada boxplot memperlihatkan ketidakseragaman sebaran antar variabel, baik dari sisi media, lebar *interquartile range*, maupun jumlah *outlier*. Kondisi ini mendukung perlunya analisis *clustering*, karena perbedaan karakteristik tersebut menunjukkan

bahwa kabupaten/kota memiliki profil infrastruktur dan layanan dasar yang beragam sehingga berpotensi membentuk kelompok yang berbeda secara alami.

3.3 Korelasi Antarvariabel dan Heatmap



Gambar 3. *Correlation heatmap* data infrastruktur dan layanan dasar Sumatera Utara, 2023

Gambar 3 memperlihatkan struktur hubungan yang cukup kuat antar indikator. Variabel X_1, X_2, X_3, X_5, dan X_7 menunjukkan korelasi positif tinggi, dengan sebagian pasangan mendekati nilai 0,70-0,85. Pola ini mengindikasikan adanya multikolinearitas yang signifikan pada kelompok indikator infrastruktur dan layanan dasar. Sementara itu X_8 memiliki korelasi negatif cukup besar terhadap kelompok tersebut, berada pada kisaran -0,65 hingga -0,80. Dua variabel lain, X_4 dan X_6, memperlihatkan korelasi sedang sekitar 0,30 – 0,45. Struktur korelasi ini menunjukkan adanya tumpang tindih informasi, sehingga metode reduksi dimensi PCA diperlukan unutk mengekstraksi komponan utama yang lebih bersifat *orthogonal*.

3.4 Analisis Principal Component Analysis (PCA)

3.4.1 Struktur Varians dan Nilai Eigen

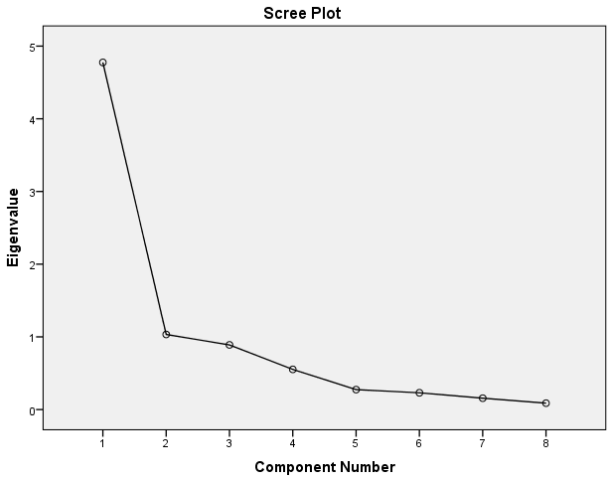
Perhitungan nilai *eigen* dilakukan dengan mengacu pada Persamaan (3). Ringkasan hasil perhitungan nilai *eigen* disajikan pada Tabel 3.

Tabel 3. Nilai Eigen

Notasi	Nilai Eigen	Notasi	Nilai Eigen
λ_1	4, 775	λ_5	0,275
λ_2	1,034	λ_6	0,231
λ_3	0,889	λ_7	0,156
λ_4	0,553	λ_8	0,088

Berdasarkan nilai *eigen* pada Tabel 3, hanya dua komponen yang layak dipertahankan karena memenuhi kriteria $\lambda \geq 1$ yaitu $\lambda_1 = 4{,}775$ dan $\lambda_2 = 1{,}034$. PC_1 menangkap pengaruh variabel infrastruktur dan kesejahteraan, sedangkan PC_2 menggambarkan variasi pada aspek pendidikan. Struktur ini menunjukkan bahwa perbedaan antar kabupaten/kota di Sumatera Utara terutama dipengaruhi oleh kualitas infrastruktur dasar dan tingkat kemiskinan, disusul komponen pendidikan.

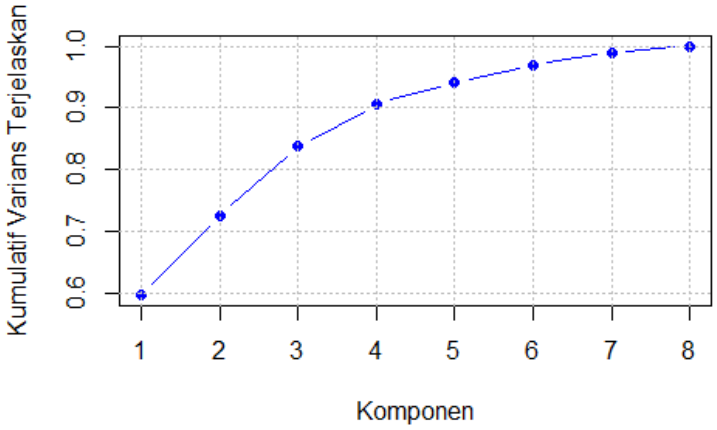
Untuk memperkuat hasil berdasarkan nilai *eigen*, dilakukan visualisasi menggunakan *scree plot* yang ditunjukkan pada Gambar 4.



Gambar 4. Scree plot

Scree plot pada Gambar 4 menampilkan penurunan tajam pada komponen pertama dan kedua, dan mendatar setelah komponen ketiga. Adanya titik siku di komponen kedua mengonfirmasi bahwa jumlah komponen yang optimal untuk dianalisisi adalah dua. Posisi dan kemiripan garis pada *scree plot* konsisten dengan nilai eigen pada Tabel 2, sehingga pemilihan dua komponen pertama sudah tepat.

Selanjutnya, untuk mendukung hasil dari *scree plot*, digunakan grafik kumulatif varians yang mempertimbangkan jumlah komponen yang mampu menjelaskan sekitar 70% hingga 80% dari total varians data (Dewi & Pakereng, 2023).



Gambar 5. Kumulatif varians oleh PCA

Berdasarkan Gambar 5, komponen pertama mampu menjelaskan varians sebesar 0,59 dan kemudian kontribusi varians meningkat menjadi 0,73 setelah penambahan komponen kedua. Dengan demikian, dua komponen sudah mencakup sekitar 72,61% dari keseluruhan varians data. Hasil ini menunjukkan bahawa dua komponen variabel dinilai paling signifikan dan akan digunakan dalam algoritma *K-Means*. Selanjutnya, diperoleh ventor *eigen* dengan persamaan 4 sebagai berikut:

$$\mathrm{I} = \begin{pmatrix} -0{,}408 & 0{,}084 & \cdots & -0{,}066 \\ -0{,}330 & -0{,}392 & \cdots & -0{,}018 \\ \vdots & \vdots & \ddots & \vdots \\ 0{,}368 & -0{,}270 & \ldots & 0{,}360 \end{pmatrix} \tag{16}$$

3.4.2 Loading Matrix dan Pembentukan Dimensi

Setelah jumlah komponen utama ditetapkan berdasarkan nilai *eigen*, *scree plot* dan grafik kumulatif varians, dilakukan analisis *loading matrix* untuk mengidentifikasi kontribusi masing-masing variabel terhadap komponen utama. Nilai *loading* dihitung dengan menggunakan Persamaan 6 dengan ringkasan hasil yang ditampilkan pada Tabel 4.

Tabel 4. Nilai loading

Variabel	PC_1	PC_2
X_1	0,892	-0,086
X_2	0,721	0,399
X_3	0,864	0,214
X_4	0,656	0,20
X_5	0,870	0,007
X_6	0,417	-0,836
X_7	0,841	0,215
X_8	-0,803	0,274

Variabel X_1 memiliki korelasi sebesar 0,892 dengan PC_1 dan -0,086 dengan PC_2. Mengingat bahwa kontribusi absolut terbesar berasal dari PC_1, variabel X_1 dikelompokkan pada komponen tersebut. Kriteria ini konsisten digunakan untuk seluruh variabel.

3.4.3 Persamaan Komponen Utama

Berdasarkan hasil komputasi ventor *eigen*, representasi persamaan komponen utama dirumuskan kembali sesuai dengan persamaan 7.

$$PC_{1,1} = -0{,}408Z_{1,1} - 0{,}3301Z_{1,2} - 0{,}395Z_{1,3} - 0{,}300Z_{1,4} - 0{,}398Z_{1,5} - 0{,}191Z_{1,6} - 0{,}385Z_{1,7} + 0{,}368Z_{1,8} \quad (17)$$

$$PC_{1,2} = 0{,}084Z_{2,1} - 0{,}392Z_{2,2} - 0{,}210Z_{2,3} + 0{,}020Z_{2,4} - 0{,}007Z_{2,5} + 0{,}823Z_{2,6} - 0{,}211Z_{2,7} - 0{,}270Z_{2,8} \quad (18)$$

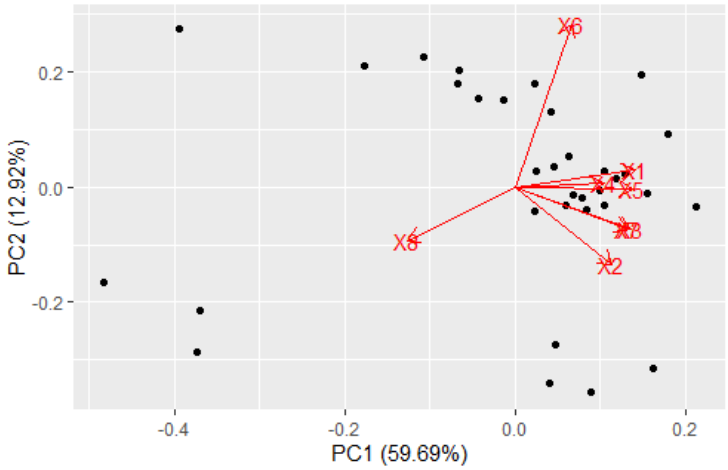
Persamaan ini menunjukkan kombinasi linier dari variabel yang telah di standarisasi (Z-*score*) dalam membentuk PC_1 dan PC_2, serta menggambarkan besarnya kontribusi relatif masing-masing variabel terhadap kedua komponen utama tersebut.

3.4.4 Transformasi Skor Komponen

Skor PCA diperoleh dengan menghitung persamaan komponen utama menggunakan data yang telah distandarisasi. Skor PC_1 dan PC_2 kemudian digunakan sebagai representasi baru yang menggambarkan struktur utama data dan menjadi input analisis *K-Means*. Proses transformasi data baru melalui PCA secara rinci disajikan pada Tabel 5.

Tabel 5. Transformasi Skor Komponen

No	PC1	PC2
1	-2,259	-1,577
2	-1,013	-1,214
3	-613	-1,291
4	38	-1,167
:	:	:
:	:	:
33	0,38	-1,037



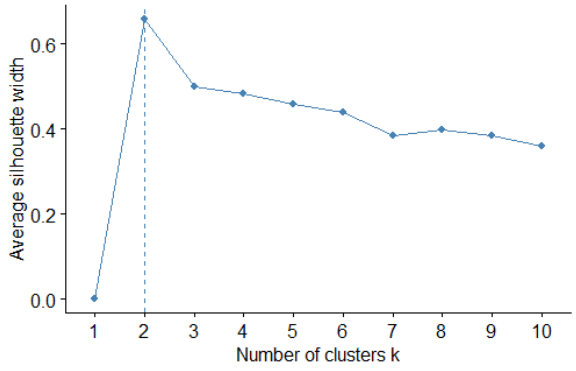
Gambar 6. Biplot dua komponen utama

Biplot dua komponen pada Gambar 6, menunjukkan bahwa PC_1 dan PC_2 mampu menjelaskan 72,61% variasi data. Variabel X_1, X_5, X_3 dan X_2 searah dan dominan pada PC_1. Hal ini menandakan korelasi positif serta peran utama dalam membedakan objek. Sebaliknya X_6 berkontribusi besar pada PC_2 dengan arah yang relatif berlawanan, menunjukkan hubungan lemah hingga negatif terhadap variabel pada PC_1. Sebaran objek yang tidak seragam mencerminkan heterogenitas karakteristik dan mendukung potensi pengelompokan berdasarkan kemiripan variabel.

3.5 Analisis K-Means

3.5.1 Penentuan Jumlah Cluster

Penentuan jumlah klaster dilakukan menggunakan metode *silhouette coefficient* (sc). Nilai rata-rata *silhouette* dari k=2 hingga k=5 ditunjukkan pada Gambar 7.



Gambar 7. Sillhouette coefficient plot

Nilai rata-rata *silhouette* pada Gambar 7 menunjukkan bahwa puncak koefisien diperoleh pada k =2, namun pemilihan jumlah klaster tidak hanya mempertimbangkan nilai maksimum. Menurut (Limbong & Likumahwa, 2025), pemilihan jumlah klaster optimal dilakukan dengan mempertimbangkan *silhouette score* > 0,5, yang menunjukkan bahwa klaster yang terbentuk memiliki kualitas pemisahan yang baik. Pada k=3, nilai *silhouette* masih berada pada kategori baik (sekitar 0,52), dan pola grafik menunjukkan penurunan stabil setelah k=3, menandakan klaster tambahan tidak meningkatkan pemisahan signifikan. Selain itu, pola sebaran pada biplot PCA sebelumnya menunjukkan adanya tiga kelompok alami. Dengan demikian, jumlah klaster optimal ditetapkan sebesar k=3, seimbang antara kualitas pemisahan dan representasi struktur data yang lebih rinci.

3.5.2 Penentuan Pusat Klaster (v_{cj}) Awal Secara Serentak

Tahap awal pembentukan klaster dilakukan dengan menentukan pusat klaster menggunakan metode *trial and error*. Nilai pusat klaster yang diperoleh dari proses tersebut ditampilkan pada Tabel 6.

Tabel 6. Pusat Klaster Untuk k=3

x_i	Kab/Kota	Pusat klaster (v_{cj})	Variabel	
			PC_1	(PC_2
1	Nias	1	2.77158	0.9483
8	Asahan	2	0.50497	2.03788
7	Labuhan Batu Selatan	3	0.84595	-1.11864

3.5.3 Penentuan Jarak Euclid Untuk Setiap Objek Terhadap Pusat Klaster

Dalam memperoleh kedekatan objek dengan pusat klaster, digunakan ukuran jarak *euclid* berdasarkan persamaan 8. Rincian hasil perhitungan tersebut tercantum pada Tabel 7.

Tabel 7. Pusat Klaster Untuk k=3

Data ke-i	Jarak Euclid Data terhadap Pusat Klaster			Alokasi Klaster
	v_1	v_2	v_3	
1	2.2619861	5.459108	6.3911423	1
2	0.6278379	2.833995	3.8591514	1
.	.	.	.	.
33	3.0488047	1.016376	2.2401841	2
34	1.9316124	1.668625	2.7519721	2

3.5.4 Pembaruan Titik Pusat Klaster

Setelah semua objek dialokasikan, pusat klaster diperbarui berdasarkan rata-rata nilai objek dalam masing-masing klaster menggunakan persamaan 9. Proses ini di ulang hingga tidak terjadi perpindahan objek antar klaster. Algoritma *K-Means* mencapai konvergensi pada iterasi ke-3 sehingga pusat klaster akhir ditunjukkan pada Tabel 8 berikut.

Tabel 8. Pusat Klaster Baru Untuk k=3

Klaster	PC_1	PC_2
1	-3.979	0.043
2	0.0367	0.996
3	1.181	-0.362

Dengan demikian, proses pembentukan klaster telah menghasilkan tiga kelompok wilayah yang memiliki karakteristik serupa berdasarkan komponan utama hasil PCA.

3.5.5 Hasil Akhir Pengelompokan

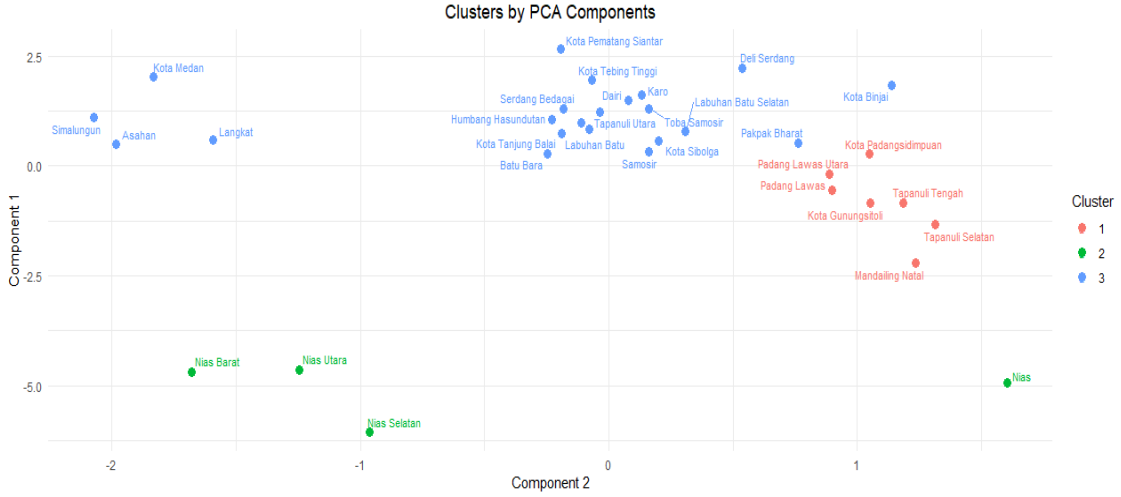
Ringkasan hasil pengelompokan untuk k=3 disajikan pada Tabel 9.

Tabel 9. Ringkasan hasil klaster final

Klaster	Jumlah Anggota	Wilayah
k_1	7	Mandailing Natal, Tapsel, Tapteng, Padang Lawas Utara, Padang Lwas, Kota Padangsidimpian, Kota Gunungsitoli

Klaster	Jumlah Anggota	Wilayah
k_2	4	Nias, Nias Sleatan, Nias Utara, Nias Barat
k_3	22	Taput, Toba Samosir, Labuhan Batu, Asahan, Simalungung, Dairi, Karo, Deli Serdang, Langkat, Humbang Hasundutan, Pakpak Bharat, Samosir, Serdang bedagai, Batu Bara, Lab Batu Selatan, Lab Batu Utara, K. Sibolga, K. Tanjung Balai, K. P. Siantar, K. Tebing tinggi, K.Medan, K.Binjai

Visualisasi hasil klaster berdasarkan PCA ditampilkan pada Gambar 8 berikut.



Gambar 8. Persebaran kabupaten/kota di Sumatera Utara tahun 2023

Scatterplot hasil pengelompokan menunjukkan pemisahan tiga klaster yang relatif tegas pada ruang PC_1- PC_2. Wilayah dalam klaster 1 terkonsentrasi pada skor PC_2 yang tinggi dan PC_1 yang sedang, mengindikasikan capaian pendidikan yang relatif baik namun infrastruktur bervariasi. Klaster 2 berada pada skor PC_1 rendah dan PC_2 rendah, mencerminkan wilayah dengan kondisi layanan dasar dan pendidikan yang lemah. Klaster 3 terletak pada skor PC_1 tinggi dengan variasi PC_2 yang lebih lebar, menggambarkan wilayah dengan kualitas infrastruktur dan pelayanan dasar yang lebih baik dibanding dua kelompok lainnya.

3.5.6 Interpretasi Klaster

Berdasarkan pusat klaster dan distribusi objek, karakteristik masing-masing klaster dapat di interpretasikan sebagai berikut.

Tabel 10. Hasil pengelompokan kabupaten/kota menggunakan 3 klaster

Klaster	Karakteristik Utama	Interpretasi
k_1	Nilai PC_1 sedang, PC_2 tinggi	Wilayah pada klaster ini menunjukkan capaian pendidikan yang relatif baik namun infrastruktur bervariasi
k_2	Nilai PC_1 dan PC_2 rendah	Kelompok ini mencerminkan wilayah dengan kondisi layanan dasar dan pendidikan yang lemah.
k_3	Nilai PC_1 tinggi, variasi PC_2 luas	Klaster ini menggambarkan wilayah dengan kualitas infrastruktur dan pelayanan dasar yang lebih baik dibanding dua kelompok lainnya

Hal ini menunjukkan bahwa ketiga klaster memiliki pola kondisi data yang berbeda satu sama lain dan pembagian klaster sudah mencerminkan struktur data.

4 KESIMPULAN

Penelitian ini berhasil mengidentifikasi variasi pembangunan infrastruktur dan layanan dasar pada 33 kabupaten/kota di Sumatera Utara melalui pendekatan *Principal Component Analysis* (PCA) dan *K-Means clustering*. Berdasarkan hasil analisis, ditemukan bahwa pengelompokan terbaik terbentuk dalam tiga klaster, yang mencerminkan wilayah dengan pendidikan relatif baik namun infrastruktur bervariasi, wilayah dengan kondisi layanan dasar dan pendidikan yang lemah, serta wilayah dengan kualitas infrastruktur dan layanan dasar yang lebih baik dibanding klaster lainnya. Temuan ini menunjukkan bahwa ketimpangan pembangunan memang terstruktur dan dapat dipetakan secara statistik melalui pola kemiripan antar indikator. Hasil pengelompokan ini memiliki relevansi praktis karena dapat digunakan sebagai dasar perumusan kebijakan yang lebih terarah, terutama dalam prioritas intervensi pembangunan dan pemerataan akses layanan publik.

UCAPAN TERIMAKASIH

Penulis menyampaikan terima kasih kepada Universitas Terbuka atas dukungan akademik dan fasilitas yang diberikan. Penulis juga mengapresiasi seluruh pihak yang telah berkontribusi dalam pengumpulan data dan kelancaran penelitian ini.

DAFTAR PUSTAKA

Anwar, K., Goejantoro, R., & Prangga, S. (2022). Pengelompokan kabupaten/kota di pulau kalimantan berdasarkan indikator indeks pembangunan manusia tahun 2020 menggunakan optimasi K-Means cluster dengan principle component analysis. *Jurnal Eksponensial*, *13*(2), 131–140.

Aviliana, F., & Hendikawati, P. (2025). Algoritma K-Means dan analisis komponan utama untuk mengatasi multikolineritas pada pengelompokan kabupaten tertinggal. *Jurnal Informatika Sunan Kalijaga, 10*(3), 294-306. https://doi.org/10.14421/jiska.2025.10.3.294-306

Enzelline, G., & Suhaedi, D. (2022). Pengunaan metode principal component analysis dalam menentukan factor domnan. *Jurnal Riset Matematika, 2*(2), 101-110. https://doi.org/10.29313/jrm.v2i2.1992

Harmain, A., Paiman, P., Kurniawan, H., Kusrini, & Maulina, D. (2022). Normalisasi data untuk efisiensi K-Means pada pengelompokan wilayah berpotensi kebakaran hutan dan lahan berdasarkan sebaran titik panas. *Teknimedia: Teknologi Informasi dan Multimedia, 2*(2), 83-89. https://doi.org/10.46764/teknimedia.v2i2.49

Izzuddin, K. H., & Wijayanto, A. W. (2024). Pemodelan clustering Ward, K-Means, DIANA, dan PAM dengan PCA untuk karakterisasi kemiskinan Indonesia tahun 2021. *Jurnal Komputika, 13*(1). https://doi.org/10.34010/komputika.v13i1.10803

Limbong, J. J. A., & Likumahwa, F. M. (2025). Implementation of the K-Means algorithm for clustering students' web programming course grades using silhouette Score. *Journal of System and Computer Engineering, 6*(3), 271–280.

Mardhatillah, A. N., & Permana, D. (2025). Pemetaan ketimpangan fasilitas pendidikan di Kabupaten Padang Pariaman menggunakan K-Means. *Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika, 6*(1), 345–351.

Maulana, A. A., Rafii, A. W., Anjelina, Y. A., & Widodo, E. (2023). Pengelompokkan kecamatan di Kabupaten Bima berdasarkan jumlah produksi dan luas panen bawang merah tahun 2021 menggunakan K-means clustering. *Jurnal Statistika, 16*(1), 442–451.

Putri, N. S., & Hayati, M. N. (2024). Pengelompokan kabupaten/kota di Kalimantan berdasarkan indikator pendidikan menggunakan metode K-Means dengan optimasi principal component analysis. *Jurnal Eksponensial, 15*(2), 128–138. https://doi.org/10.30872/eksponensial.v15i2.1373

Rosyada, I.A., & Utari, D.T. (2024). Penerapan principal component analysis untuk reduksi pada algoritma K-Means clustering. *Jambura Journal of Probability and Statistics, 5*(1), 6-13.

Santoso, A. B., Kusuma, A. C., & Nooraeni, R. (2024). Development of a hybrid fuzzy geographically weighted K-prototype clustering and genetic algorithm for enhanced spatial analysis: Application to rural development mapping. *Jurnal Aplikasi Statistika dan Komputasi Statistik, 16*(2), 122-139.

Saputra, N.R. (2025). Pengelompokan wilayah Indonesia berdasarkan komponen indeks pembangunan manusia dengan pendekatan algoritma K-Means clustering. *Jurnal SKANIKA: Sistem Komputer dan Informatika, 17*(1), 45-56.

Silitonga, A. I., Nabila, Z. A., Rizkia, C., Lubis, Z., & Safitri, N. (2024). Klasterisasi gizi buruk dan stunting di Provinsi Sumatera Utara menggunakan K-Means clustering. *Jurnal Methodika, 10*(2), 13–18.

Sipayung, S. P., & Hasugian, P. M. (2025). Integrating PCA and K-Means for evidence-based staple food segmentation: An Indonesian food policy approach. *Jurnal Teknik Informatika, 9*(4), 3175–3189.

Suhandy, D., & Yulia, M. (2021). *Tutorial Analisis Data Spektra Menggunakan The Unscrambler* (2nd ed.). Graha Ilmu.