HOAQ: JURNAL TEKNOLOGI INFORMASI, Volume 17 Nomor 2, Desember 2026, p362-372
p-ISSN: 2337-5280, e-ISSN: 26207427
DOI: https://doi.org/10.52972/hoaq.vol17no2

SEGMENTASI PROFIL KESEHATAN PROVINSI DI INDONESIA MENGGUNAKAN ANALISIS FAKTOR DAN *K-MEANS CLUSTERING*

Galih Zaky Tristanaya[1], Febriani Yolanda Tessalonika [2], Muhammad Rafy Danadyaksa[3], Shindi Shella May Wara[4], Amri Muhaimin[5]
[1] Program Studi Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Jl. Rungkut Madya, Surabaya – Jawa Timur, Indonesia
Email: [1]24083010088@student.upnjatim.ac.id, [2]240830100037@student.upnjatim.ac.id,
[3]24083010101@student.upnjatim.ac.id

ABSTRAK

Indonesia sebagai negara kepulauan dengan 38 provinsi menyimpan berbagai ketimpangan kondisi kesehatan yang kompleks. Tujuan dari penelitian ini yaitu untuk mengidentifikasi pola kategorisasi provinsi di Indonesia menurut parameter kesehatan melalui pendekatan analisis faktor dan klasterisasi K-Means. Riset ini menggunakan *dataset* yang memuat 12 variabel kesehatan provinsi Indonesia. Kelayakan data dan variabel diverifikasi dengan uji *Kaiser-Meyer-Olkin* (KMO) dan *Bartlett's Test*. Kemudian diekstrak ke faktor-faktor yang menghasilkan 3 faktor. Faktor yang dihasilkan selanjutnya digunakan sebagai input klasterisasi K-Means, dengan jumlah klaster optimal ditentukan melalui *Elbow Method*, *Davies-Bouldin Index*, dan *Calinski-Harabasz Index*, serta divalidasi menggunakan *Silhouette Score*. Hasil penelitian mengangkat empat klaster yang divalidasi dengan *Silhouette Score* sebesar 0,4563. Klaster yang dihasilkan mengungkapkan segmentasi wilayah yang kontras secara detail, di mana klaster 0 yang mencerminkan wilayah berprofil rata-rata atau menengah nasional, klaster 1 yang mencerminkan *outlier* tunggal dengan tingkat kerentanan tertinggi dan kritis, klaster 2 yang mencerminkan pusat demografi dengan kapasitas infrastruktur medis berskala masif, dan klaster 3 yang mencerminkan wilayah terluar yang rentan dan jauh. Analisis ini menegaskan bahwa pendekatan analisis faktor dan klasterisasi K-Means mampu menghasilkan pola-pola dalam kesehatan nasional, sehingga pemerintah mampu mempertimbangkan profil tiap klaster secara spesifik demi intervensi yang efektif dan berkeadilan.

Kata kunci: Analisis Faktor; *K-Means Clustering*; Indikator Kesehatan; Ketimpangan Wilayah; *Silhouette Score*.

ABSTRACT

As an archipelagic nation with 38 provinces, Indonesia faces a range of complex health disparities. The purpose of this study is to identify characteristics of provincial categorization in Indonesia according to health parameters using factor analysis and K-Means clustering. This study uses a dataset containing 16 health variables for Indonesian provinces. The eligibility of the data and variables was verified using the Kaiser-Meyer-Olkin (KMO) test and Bartlett's test. The data was then extracted into factors, yielding three factors. These factors were then used as input for K-Means clustering, with the optimal number of clusters determined using the Elbow Method, the Davies-Bouldin Index, and the Calinski-Harabasz Index, and validated using Silhouette Score. The result identified four clusters, validated by a Silhouette Score of 0.4563. The resulting clusters reveal a detailed segmentation of regions with contrasting characteristics: Cluster 0 represents regions with an average or national-level profile; Cluster 1 represents a single outlier with the highest and most critical vulnerability level; Cluster 2 represents a demographic center with massive medical infrastructure capacity; and Cluster 3 represents vulnerable and remote outlying regions. This analysis confirms that factor analysis and K-means clustering can identify patterns in national health, enabling the government to consider the specific profile of each cluster in order to implement effective and equitable interventions.

Keywords: Factor Analysis; K-Means Clustering; Health Indicators; Regional Inequality; Silhouette Score.

1. PENDAHULUAN

Indonesia merupakan negara kepulauan dengan 38 provinsi yang menghadapi disparitas kondisi kesehatan yang bersifat struktural dan persisten. Fenomena ini mencerminkan disparitas multidimensional yang mencakup aksesibilitas layanan, kapasitas infrastruktur, perilaku masyarakat, dan kondisi sanitasi lingkungan. Studi *Global Burden of Disease* 2019 yang dipublikasikan dalam *The Lancet Global Health* mencatat bahwa selisih angka harapan hidup di Indonesia mencapai 9,9 tahun pada laki-laki dan 13,7 tahun pada perempuan dengan kesenjangan justru semakin melebar sejak 1990 [1]. Fakta ini menunjukkan bahwa pendekatan kebijakan yang bersifat seragam (*one-size-fits-all*) secara empiris gagal menjawab kebutuhan yang sesungguhnya berbeda secara fundamental di setiap klaster wilayah. Oleh karena itu, upaya pengelompokan wilayah berbasis indikator kesehatan di Indonesia

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

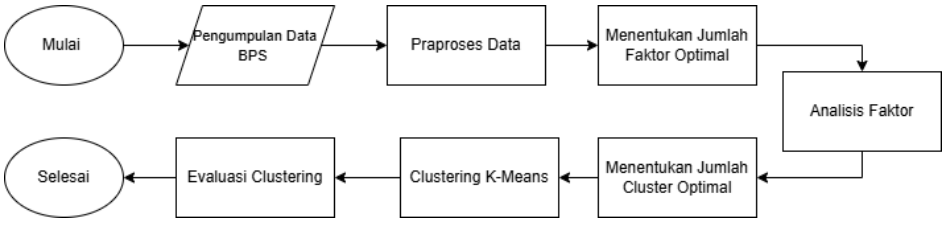
telah dilakukan oleh sejumlah penelitian sebelumnya. Maryani et al. menerapkan K-Means Clustering terhadap indikator kesehatan keluarga di seluruh Indonesia yang menemukan disparitas yang signifikan, khususnya pada dimensi gizi, kesehatan ibu, dan kesehatan anak, dengan Papua dan Papua barat secara konsisten menempati klaster performa rendah [2]. Sementara itu, [3] menerapkan K-Medoids untuk menganalisis kerentanan kesehatan 38 provinsi berdasarkan indikator lingkungan dan mengidentifikasi empat klaster dengan variasi spasial yang jelas [3]. Pada 2025 WHO mencatat rasio kematian ibu di Papua mencapai 565 kematian per 100.000 kelahiran yang berkisar sebelas kali lipat dari angka DKI Jakarta yang mencapai 48, yang menjadi bukti bahwa disparitas kesehatan di Indonesia berdimensi sangat dalam dan tidak dapat diatasi hanya dengan pendekatan analitik yang bersifat deskriptif [4]. Penelitian terdahulu yang dilakukan oleh [5] menunjukkan hasil kombinasi Principal Component Analysis dan K-Means efektif dalam melakukan segmentasi data kesehatan seperti gizi buruk dengan lebih akurat [5].

Berdasarkan metodologis, upaya klasterisasi wilayah berbasis indikator kesehatan telah dilakukan oleh beberapa peneliti, namun umumnya penerapan algoritma pengelompokan secara langsung pada variabel observasional mentah tanpa mempertimbangkan multikolinearitas dan redundansi informasi yang seimbang pada data multivariat berdimensi tinggi [6]. Saat variabel saling berkorelasi tinggi, K-means Konvensional akan memberikan bobot tidak proporsional, sehingga menghasilkan nilai kelompok yang bias. Pendekatan Kualitas klaster yang selanjutnya perlu divalidasi secara menyeluruh, dan penggunaan tiga indeks evaluasi internal secara konvergen pada Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Index terbukti memberikan nilai yang stabil dibanding hanya pada satu indeks tunggal yang akan mudah mengalami bias [7]. Hal ini diperkuat oleh [8] yang mengembangkan pendekatan *ensemble* indeks validitas klaster berbasis integrasi *Silhouette*, Davies Bouldin, dan Calinski Harabasz, dan menemukan bahwa konvergensi dari beberapa indeks secara konsisten menghasilkan estimasi jumlah klaster optimal yang lebih stabil dan handal dibandingkan penerapan indeks tunggal [8]. Temuan serupa dalam analisis komparatif beberapa indeks validitas klaster internal yang menunjukkan bahwa setiap indeks memiliki asumsi matematis dan sensitivitas berbeda terhadap bentuk dan ukuran klaster, sehingga penggunaan tunggal berisiko menghasilkan rekomendasi K yang tidak robust [8]. Selain itu, penelitian pada domain profil beban listrik [9] menemukan bahwa Silhouette Score (SH) cenderung bias terhadap klaster berukuran besar, sedangkan Calinski Harabasz Index (CH) condong pada klaster berbentuk cembung padat, sehingga penggunaan gabungan keduanya bersama Davies Bouldin Index (DBI) dapat saling mengkompensasi kelemahan masing-masing dan menghasilkan keputusan yang lebih konservatif serta tidak mudah terpengaruh oleh karakteristik distribusi data tertentu [10].

Penelitian ini bertujuan mengidentifikasi pola pengelompokan pada provinsi di Indonesia melalui pendekatan struktur laten yang mengintegrasikan analisis faktor berbasis rotasi Varimax dengan K-Means *Clustering* pada ruang faktor. Kelayakan analisis dikonfirmasi melalui uji Kaiser-Meyer-Olkin (KMO) dan Bartlett's Test of Sphericity, sedangkan jumlah klaster optimal ditentukan melalui konvergensi tiga metrik evaluasi yang divalidasi menggunakan Silhouette Analysis. Penelitian ini mendemonstrasikan secara empiris keunggulan pendekatan berbasis struktur laten dibandingkan klasterisasi konvensional dalam konteks data kesehatan multivariat antar provinsi di Indonesia. Secara Praktis, hasil tipologi klaster yang dihasilkan dapat menjadi referensi untuk pihak berwenang untuk merumuskan kebijakan kesehatan terdiferensiasi sesuai profil spesifik tiap klaster, sehingga intervensi yang dirancang lebih tepat sasaran, efisien, dan berkeadilan bagi seluruh provinsi.

2. METODE PENELITIAN

Studi ini disusun melalui proses yang dilakukan secara menyeluruh dan urut dalam menggabungkan metode analisis faktor dan algoritma clustering K-Means. Tahapan penelitian mencakup pengumpulan data, pengolahan data, penerapan metode analisis faktor, menentukan jumlah cluster optimal, klasterisasi K-Means dan diakhiri evaluasi.



Gambar 1. Alur Penelitian

Pengumpulan Data BPS

Studi ini menggunakan data dari berbagai publikasi kesehatan Badan Pusat Statistik (BPS) tahun 2025. Dataset mencakup keselruhan 38 provinsi di Indonesia guna merepresentasikan kondisi kesehatan masyarakat. Meskipun jumlah obeservasi hanya 38 provinsi, angka ini sudah merepresentasikan dengan baik keseluruhan jumlah populasi tingkat provinsi yang ada di Indonesia sehingga dapat meminimalisir adanya bias. Indikator yang dievaluasi secara komprehensif meliputi ketersediaan sumber daya medis, akses pelayanan, kesehatan ibu dan anak, sanitasi lingkungan, perilaku kesehatan, hingga status gizi masyarakat.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Tabel 1. Variabel Penelitian

Variabel	Simbol	Satuan	Skala
Jumlah Nakes	X_1	Orang	Rasio
Jumlah Fasyankes	X_2	Unit	Rasio
Jumlah Penduduk	X_3	Ribu Jiwa	Rasio
Umur Harapan Hidup Saat Lahir	X_4	Tahun	Rasio
Proporsi Rumah Tangga dengan Sanitasi Layak	X_5	%	Rasio
Proporsi Rumah Tangga dengan Sumber Air Minum Air Layak	X_6	%	Rasio
Proporsi Anak Umur 0-59 Bulan Mendapat Imunisasi Campak	X_7	%	Rasio
Proporsi Bayi Usia < 6 Bulan yang Mendapat ASI Eksklusif	X_8	%	Rasio
Persentase Melahirkan di Fasilitas Kesehatan	X_9	%	Rasio
Persentase Penduduk yang Mempunyai Keluhan Kesehatan	X_{10}	%	Rasio
Rata-rata Konsumsi Kalori per Kapita per Hari	X_{11}	Kkal/Kapita/Hari	Rasio
Rata-rata Konsumsi Protein per Kapita per Hari	X_{12}	Gram/Kapita/Hari	Rasio

Praproses Data

Data yang telah dihimpun dari BPS selanjutnya memasuki tahap pra proses. Langkah pertama yang dilakukan adalah menghapus variabel non-numerik, seperti kolom "Provinsi", untuk murni menyisakan data kuantitatif. Setelah itu, dilakukan standarisasi guna menyeragamkan skala pengukuran antar variabel. Kemudian diakhiri dengan seleksi variabel menggunakan evaluasi *Measure of Sampling Adequacy* (MSA).

Standarisasi Data

Standardisasi menggunakan metode Z-Score untuk menyetarakan skala data. Nilai terstandarisasi pada Z_{ij} didapatkan dengan mengurangi nilai observasi ke-i pada variabel ke-j X_{ij} terhadap rata-rata variabel X_j^-, yang kemudian dibagi dengan standar deviasinya S_j, berdasarkan persamaan berikut.

$$Z_{ij} = \frac{X_{ij} - \underline{X_j}}{S_j}$$

Uji Kelayakan Analisis Faktor

Data terstandarisasi selanjutnya dievaluasi untuk mengetahui kelayakannya untuk analisis faktor menggunakan *Bartlett's Test* dan *Kaiser-Meyer-Olkin* (KMO). *Bartlett's Test* untuk membuktikan signifikansi korelasi antarvariabel di dalam matriks dengan syarat kelayakan *p-value* < 0,05. Sementara itu, uji KMO mengukur kecukupan sampel sekaligus kesesuaian variabel untuk direduksi menjadi faktor, di mana nilainya harus melampaui ambang batas 0,50 dan diinterpretasikan semakin baik jika mendekati angka 1.

1) *Kaiser-Meyer-Olkin* (KMO)

Pengujian KMO dilakukan guna mengevaluasi sampel untuk mengukur kekuatan pola korelasi antarvariabel di dalam *dataset*. Nilai KMO yang tinggi mengindikasikan bahwa korelasi antarvariabel yang kuat sehingga layak dilakukan analisis faktor. Secara matematis, nilai KMO dihitung dengan berdasarkan persamaan berikut.

$$KMO = \frac{\sum_{i \neq j} \; r_{ij}^2}{\sum_{i \neq j} \; r_{ij}^2 + \sum_{i \neq j} \; q_{ij}^2} \quad (1)$$

2) *Bartlett's Test*

Pengujian *Bartlett's Test* dilakukan untuk mengevaluasi kelayakan ekstraksi faktor dengan memastikan bahwa matriks korelasi R berbeda signifikan dari matriks identitas. Semakin kecil nilai signifikansi yang dihasilkan, semakin kuat pola hubungan antarvariabel di dalam data sehingga proses reduksi dimensi layak untuk dilakukan. Secara matematis, nilai statistik uji Bartlett χ^2 dihitung berdasarkan jumlah observasi n, jumlah variabel penelitian p, serta nilai determinan dari matriks korelasi $|\ R\ |$, menggunakan persamaan berikut.

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right) ln\ ln\ |R| \quad (2)$$

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Seleksi Variabel Menggunakan MSA

Seleksi kelayakan tiap variabel menggunakan *Measure of Sampling Adequacy*. Apabila nilai pengujian variabel MSA_i memperoleh skor di bawah 0,50, variabel tersebut akan dieliminasi secara bertahap karena tidak memiliki korelasi parsial yang memadai. Secara matematis, kalkulasi MSA dilakukan dengan membandingkan kuadrat korelasi r_{ij} terhadap korelasi parsialnya q_{ij} antara variabel ke-i dan variabel ke-j, berdasarkan persamaan berikut.

$$MSA_i = \frac{\sum_{j \neq i} \quad r_{ij}^2}{\sum_{j \neq i} \quad r_{ij}^2 + \sum_{j \neq i} \quad q_{ij}^2} \tag{3}$$

Menentukan Jumlah Faktor Optimal

Penentuan jumlah faktor optimal dilakukan melalui perhitungan nilai eigenvalue dari matriks korelasi serta divalidasi secara visual menggunakan Scree Plot guna menghasilkan model merepresentasikan mayoritas informasi data. Faktor dipertahankan dalam model adalah faktor yang memiliki eigenvalue lebih besar dari satu. Secara matematis, nilai eigenvalue ini diperoleh dengan menyelesaikan persamaan determinan yang melibatkan matriks identitas berikut.

$$|R - \lambda I| = 0 \tag{4}$$

Analisis Faktor

Proses reduksi dimensi diawali dengan menghitung nilai eigenvalue untuk menentukan jumlah faktor optimal pembentuk model. Kemudian model tersebut kemudian diproses ulang menggunakan rotasi orthogonal Varimax agar setiap variabel X_i terdistribusi dan mengelompok secara tegas pada faktor umum F_j sesuai dengan bobot factor loading-nya a_{ij} tanpa saling tumpang tindih. Secara matematis, hubungan linier pada struktur Analisis Faktor yang turut memperhitungkan galat spesifik ε_i dari masing-masing variabel ini dirumuskan melalui persamaan berikut.

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i \tag{5}$$

Selanjutnya mengevaluasi nilai komunalitas untuk mengukur seberapa besar proporsi informasi dari variabel asli yang berhasil diekstrak oleh model akhir. Nilai komunalitas yang tinggi mengindikasikan bahwa variabel tersebut terwakili secara optimal, yang selanjutnya akan ditransformasikan menjadi metrik baru berupa skor faktor terstandarisasi. Secara matematis, komunalitas untuk variabel ke-i ini dikalkulasikan dengan menjumlahkan kuadrat dari bobot muatan faktornya a_{ij} pada keseluruhan jumlah faktor yang terbentuk m, berdasarkan persamaan berikut.

$$h_i^2 = \sum_{j=1}^{m} \quad a_{ij}^2 \tag{6}$$

Proses ini diakhiri dengan pembentukan skor faktor untuk mengonversi data awal menjadi metrik baru terstandarisasi dengan dimensi yang lebih ringkas tanpa menghilangkan informasi esensial. Nilai skor faktor ini dikalkulasikan melalui koefisien skor faktor b_{ij} dengan nilai variabel yang telah distandarisasi Z_i untuk seluruh jumlah variabel penelitian p. Metrik baru inilah yang selanjutnya diimplementasikan sebagai data masukan utama pada algoritma K-Means Clustering untuk mengelompokkan observasi provinsi, yang secara matematis dirumuskan melalui persamaan berikut.

$$F_j = b_{1j}Z_1 + b_{2j}Z_2 + \cdots + b_{pj}Z_p \tag{7}$$

Menentukan Jumlah *Cluster* Optimal

Sebelum melakukan clustering perlu untuk menentukan jumlah klaster optimal yang ditentukan melalui Davies-Bouldin Index (DBI) untuk mengukur perbandingan antara tingkat kohesi dan pemisahan antar kelompok. Semakin kecilnya nilai DBI menandakan bahwa jumlah grup yang tercipta memiliki struktur anggota yang semakin baik. Secara matematis, metrik ini dikalkulasikan dengan mencari rasio maksimal dari penjumlahan rata-rata jarak anggota ke centroid pada cluster S_i dan klaster S_j dibagi dengan jarak antar-centroid kedua klaster tersebut M_{ij}, berdasarkan persamaan matematis berikut.

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \quad \left(\frac{S_i + S_j}{M_{ij}}\right) \tag{8}$$

Calinski-Harabasz Index digunakan untuk mengevaluasi seberapa baiknya kualitas suatu klaster berdasarkan rasio antara tingkat variasi antar-klaster B_k dengan variasi di dalam klaster W_k. Perolehan nilai CHI yang semakin tinggi mengindikasikan bahwa kelompok observasi memiliki homogenitas internal yang sangat baik. Oleh karena itu, pembentukan jumlah cluster optimal dari total keseluruhan observasi didasarkan pada capaian skor CH tertinggi, yang secara matematis dikalkulasikan melalui persamaan berikut.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

$$CH = \frac{\frac{B_k}{K-1}}{\frac{W_k}{n-K}} \quad (9)$$

***Clustering* K-Means**

K-Means merupakan sebuah metode *unsupervised learning* yang memisahkan data kedalam grup - grup berdasarkan tingkat kemiripan karakteristiknya. Proses pembagian anggota klaster dilakukan dengan mengevaluasi kedekatan setiap data terhadap titik pusat terdekat menggunakan jarak Euclidean $d(x_i, x_j)$. Secara matematis, metrik ini dikalkulasikan dengan mengakumulasikan selisih kuadrat antara nilai observasi ke-i x_{ik} dan observasi ke-j x_{jk} pada keseluruhan jumlah variabel penelitian, berdasarkan persamaan berikut.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} \quad (x_{ik} - x_{jk})^2} \quad (10)$$

Setelah seluruh data terkelompokkan, titik *centroid* akan diperbarui dengan menghitung titik tengah atau nilai rata-rata dari seluruh anggota di klaster tersebut. Perhitungan tersebut akan terus dilakukan hingga mencapai titik konvergen yang dimana posisi *centroid* tidak ada lagi data yang berpindah grup.

Evaluasi Hasil *Clustering*

Hasil clustering yang telah terbentuk dievaluasi dengan Silhouette Score serta divisualisasikan melalui Silhouette Plot. Metode ini menghitung nilai evaluasi untuk setiap observasi ke-i $S(i)$ berdasarkan rasio perbandingan antara tingkat kohesi, yakni rata-rata jarak suatu data dengan anggota di cluster yang sama $a(i)$, dengan tingkat separasinya, yakni rata-rata jarak terhadap klaster tetangga terdekat $b(i)$.Secara matematis, dirumuskan melalui persamaan berikut.

$$S(i) = \frac{b(i)-a(i)}{(a(i),b(i))} \quad (11)$$

3. HASIL DAN PEMBAHASAN

Uji Kelayakan Data dan Variabel

Hasil uji untuk memenuhi asumsi dan kelayakan sebelum dianalisis menggunakan metode analisis faktor dilakukan dengan menggunakan *Kaiser Meyer Olkin*, *Bartlett's Test*, dan *Measure of Sampling Adequacy*.

Tabel 2. Uji Kelayakan Analisis Faktor

Pengujian	Indeks	Kriteria
Kaiser Meyer Olkin (KMO)	0.7479	> 0.5
Pengujian	P-Value	Kriteria
Bartlett's Test (Sig.)	1.0124e-43	< 0.05

Berdasarkan Tabel 2 diperoleh p-value KMO 0.7479 itu menunjukkan bahwa dataset memiliki kecukupan sampel sehingga bisa digunakan dalam analisis faktor. Kemudian hasil *Bartlett's Test* menunjukkan *p-value* 0.0000 itu memperlihatkan adanya hubungan kuat antar variabel penelitian yang menunjukkan bahwa dataset bisa dilakukan analisis faktor.

Tabel 3. Seleksi Variabel MSA

Variabel	MSA	Variabel	MSA
X_1	0.791730	X_7	0.672055
X_2	0.594157	X_8	0.595624
X_3	0.598050	X_9	0.874688
X_4	0.827223	X_{10}	0.847634
X_5	0.720639	X_{11}	0.685673
X_6	0.798335	X_{12}	0.787536

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Berdasarkan Tabel 3 seluruh variabel yang digunakan untuk penelitian masih berada di atas batas minimum kelayakan sebesar 0.5 sehingga tidak terdapat variabel yang perlu dieliminasi dari proses analisis.

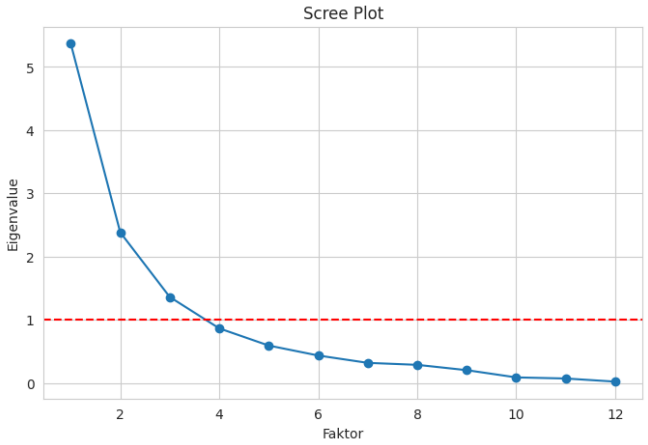
Menentukan Jumlah Faktor Optimal

Dalam menetapkan faktor optimal dilakukan dengan memakai nilai eigenvalue yang kemudian akan divisualisasikan ke dalam scree plot. Berikut adalah output dari perhitungan eigenvalue dan visualisasi scree plotnya.

Tabel 4. Nilai Eigenvalue Jumlah Faktor Optimal

Faktor	Eigenvalue
1	5.369366
2	2.382947
3	1.361093
…	…
11	0.022214

Berdasarkan Tabel 4 perhitungan eigenvalue terdapat tiga faktor mempunyai eigenvalue lebih besar daripada 1, yaitu faktor 1 5.369366, faktor 2 2.382947, dan faktor 3 1.361093. Faktor pertama memiliki nilai eigenvalue paling tinggi dibandingkan faktor lainnya, yang menunjukkan bahwa faktor tersebut mampu menjelaskan proporsi variasi data terbesar, sehingga tiga faktor pertama tersebut dianggap telah mampu mewakili sebagian besar informasi dari seluruh variabel.



Gambar 2. Scree Plot Faktor Optimal

Berdasarkan Gambar 2, memperlihatkan scree plot terdapat penurunan eigenvalue sangat tajam pada faktor pertama hingga ketiga. Setelah faktor ketiga, grafik mulai menunjukkan pola yang lebih landai dan cenderung stabil, pola menunjukkan bahwa kontribusi faktor setelah faktor ketiga dalam menjelaskan keragaman data menjadi semakin kecil. Titik perubahan grafik dari curam menjadi landai sering disebut sebagai *elbow point* menunjukkan batas jumlah faktor yang layak dipertahankan. Dengan demikian, tiga faktor utama dinilai telah cukup untuk merepresentasikan sebagian besar informasi dari variabel penelitian.

Analisis Faktor

Analisis faktor diawali dengan proses loading faktor dari setiap variabel penelitian terhadap jumlah faktor optimal yang telah ditentukan. *Loading* faktor untuk mengidentifikasi kekuatan hubungan variabel dengan faktor tertentu.Dalam penelitian ini batas loading yang digunakan pada kelompok loading faktor adalah $|loading| \geq 0{,}50$. Hasil *loading* faktor memiliki 2 jenis nilai yaitu postif dan negatif. Nilai positif bermakna bahwa sebuah variabel memiliki kekuatan hubungan berbanding lurus dengan faktor yang terbentuk, sedangkan tanda negatif bermakna bahwa variabel tersebut memiliki hubungan yang berbanding terbalik dengan faktor yang terbentuk.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Tabel 5. Matriks *Loading*

Variabel	Faktor 1	Faktor 2	Faktor 3	Variabel	Faktor 1	Faktor 2	Faktor 3
X_1	0.002	-0.509	-0.537	X_7	0.520	-0.034	-0.061
X_2	0.096	0.982	0.171	X_8	-0.074	0.209	0.373
X_3	0.140	0.924	0.177	X_9	0.787	0.134	0.337
X_4	0.680	0.274	0.424	X_{10}	0.445	0.252	0.647
X_5	0.185	0.793	0.088	X_{11}	0.182	0.046	0.872
X_6	0.809	0.225	-0.063	X_{12}	0.770	-0.056	0.557

Berdasarkan Tabel 5 dapat diketahui pada faktor 1 dibentuk oleh lima variabel yang memiliki korelasi positif kuat, yaitu X_6 dengan 0,809, X_9 dengan 0,787, X_{12} dengan 0,770, X_4 dengan 0,680, dan X_7 dengan 0,520. Kemudian faktor 2 dibentuk oleh tiga variabel dengan korelasi positif yang sangat kuat, yaitu X_2 dengan 0,982, X_3 dengan 0,924, dan X_5 dengan 0,793. Faktor 3 didominasi oleh variabel X_{11}, X_{10}, dan X_1. Faktor 3 dibentuk oleh tiga variabel, yaitu X_{11} dengan 0,872, X_{10} dengan 0,647), dan X_1 yang memiliki nilai *loading* negatif terbesar pada faktor ini, yaitu -0,537. Nilai negatif pada variabel X_1 mengindikasikan bahwa besarnya tingkat kerentanan keluhan kesehatan berbanding terbalik dengan ketersediaan jumlah tenaga kesehatan di faktor tersebut. Berdasarkan hasil ketiga matriks loading dapat diketahui ketiga kelompok tersebut didefinisikan sebagai Kualitas Kesehatan Dasar pada faktor 1, Skala Demografi dan Infrastruktur pada faktor 2, dan Tingkat Kerentanan Wilayah pada faktor 3.

Tabel 6. Skor Faktor

Data Entri	Skor Faktor 1	Skor Faktor 2	Skor Faktor 3
0	0.015	0.209	0.286
1	0.176	1.611	0.070
2	-0.619	-0.003	0.883
3	0.443	-0.247	-0.277
4	-0.049	-0.360	0.175
…	…	…	…

Setelah melakukan *loading* faktor. Kemudian menghitung matriks koefisien skor faktor untuk mengestimasi bobot dari masing-masing variabel. Matriks koefisien ini selanjutnya dikalikan dengan nilai data asli yang telah distandarisasi untuk menghasilkan Hasil Skor Faktor. Berdasarkan Tabel 6 dapat diketahui pada data entri 0 pada skor Faktor 2 dengan skor mencapai 0.209 ini mengindikasikan bahwa wilayah ini adalah wilayah metropolitan atau provinsi padat penduduk dengan jumlah fasilitas kesehatan yang masif. Sementara itu, faktor 1 dan faktor 2 sedikit di atas rata-rata. Interpretasi ini terus berlanjut hingga data entri terakhir.

Tabel 7. Skor *Communality*

Variabel	*Communality*	Variabel	*Communality*
X_1	0.548654	X_7	0.276098
X_2	1.003104	X_8	0.188772
X_3	0.905886	X_9	0.752285
X_4	0.718565	X_{10}	0.681519
X_5	0.935080	X_{11}	0.797315
X_6	0.709678	X_{12}	0.908892

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Langkah terakhir dalam melakukan analisis faktor adalah dengan melakukan evaluasi terhadap nilai komunalitas untuk mengukur proporsi informasi dari variabel yang berhasil diwakili oleh tiga faktor yang telah tercipta. Pada tabel 7 dapat diketahui variabel X8 memiliki skor 0.188772 yang dimana itu merupakan skor yang sangat rendah. Di sisi lain pada variabel X2 memiliki nilai komunalitas yang mencapai lebih dari 1 yaitu 1.003104 yang dimana secara teori melebihi batas yaitu 1. Hal ini mengindikasikan adanya anomali atau bias yang diakibatkan oleh ukuran sample yang kecil yang dimana pada penelitian ini hanya 38. Meskipun hasil skor communality melebihi batas, nilai komunalitas tetap dibatasi pada nilai maksimum yaitu 1,0 yang menegaskan bahwa varians pada variabel X2 berhasil ditangkap secara utuh dan maksimal oleh struktur faktor yang terbentuk di dalam model.

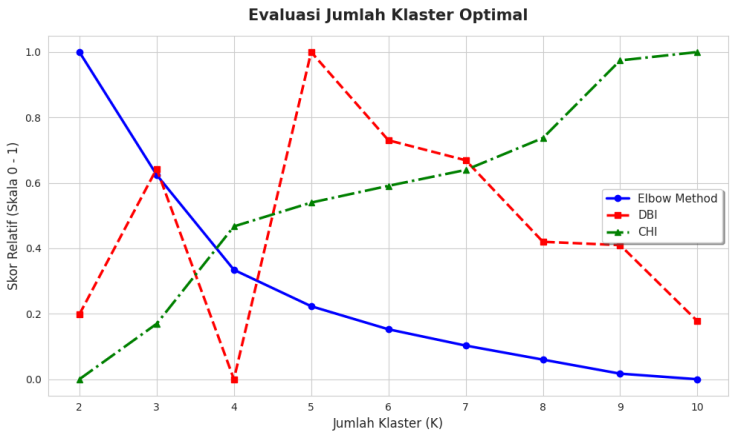
Menentukan Jumlah Cluster Optimal

Menentukan jumlah cluster optimal dilakukan menggunakan beberapa metode diantaranya *Elbow Method*, *Calinski Harabasz*, *Davies-Bouldin Index,* dan *Silhouette Score*.

Tabel 8. Penentuan Claster Optimal

K	*Elbow Method*	DBI	CHI	*Silhouette Score*
2	77.7987	0.6047	14.6408	0.4813
3	51.2144	0.6684	19.8952	0.5156
4	30.6542	0.5763	29.1278	0.4563
5	22.7798	0.7194	31.3845	0.3692
6	17.7880	0.6808	32.9753	0.3390
7	14.2621	0.6721	34.4792	0.3392
8	11.2219	0.6364	37.5095	0.3627
9	8.1814	0.6350	44.8650	0.4008
10	6.9815	0.6017	45.6574	0.3732

Pada Tabel 8 dapat ketahui bahwa pada *Elbow Method* terjadi penurunan nilai yang sangat drastis pada K = 2 hingga K = 4. Setelah K = 4 penurunan nilai mulai melandai. Titik patahan yang terbentuk pada K = 4 mengindikasikan bahwa penambahan klaster setelah titik tersebut tidak lagi memberikan efisiensi pengelompokan yang baik. Pemilihan klaster dengan K = 4 ini juga didukung oleh nilai DBI, di mana nilai optimal ditunjukkan pada skor terendah dengan skor 0,5763. Di sisi lain, CHI menunjukkan tren yang terus meningkat seiring bertambahnya jumlah K, dengan nilai maksimal pada K = 10 dengan nilai 45,6574. Namun, karena pemilihan cluster optimal mengharuskan adanya kesamaan dari mayoritas metode yang digunakan, maka K = 4 dipilih sebagai jumlah klaster yang paling optimal. Keputusan ini selanjutnya juga dikonfirmasi ulang stabilitasnya melalui metrik *Silhouette Score* pada K = 4 yang memperoleh skor 0,4563. Berikut adalah visualisasi jumlah klaster optimal.



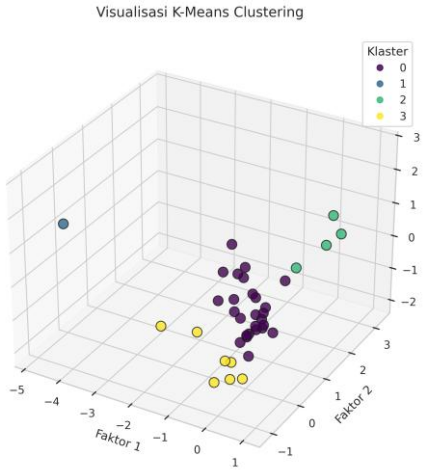
Gambar 3. Evaluasi Jumlah Cluster Optimal

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Berdasarkan Gambar 3 kurva *Elbow Method* menunjukkan adanya patahan yang cukup tajam dari K=4, kemudian mulai melandai setelah K=5. Hal menunjukkan jumlah K setelah 4 tidak memberikan penurunan variasi intra-klaster yang signifikan. Selain itu pada DBI menunjukkan bahwa nilai indeks terendah berada pada K=4. Nilai terendah DBI tersebut menunjukkan kualitas klaster yang baik. Terakhir *Calinski-Harabasz Index* memperlihatkan tren yang terus naik. Nilai CHI maksimal didapat pada K=10, yang menunjukkan bahwa pemisahan antar-klaster semakin besar dibandingkan variasi di dalam klaster. Sehingga dari ketiga metode tadi, *elbow method* dan DBI memiliki nilai K=4 sedangkan nilai CHI K=10 maka jumlah cluster optimal yang dipilih adalah K=4.

K-Means Clustering

K-Means dilakukan setelah menetapkan jumlah cluster optimal yaitu 4. Proses segmentasi ini mempartisi 38 provinsi di Indonesia ke dalam empat kelompok terpisah berdasarkan tingkat kemiripan karakteristik kesehatan, skala demografi, dan tingkat kerentanan yang direpresentasikan oleh matriks skor faktor.

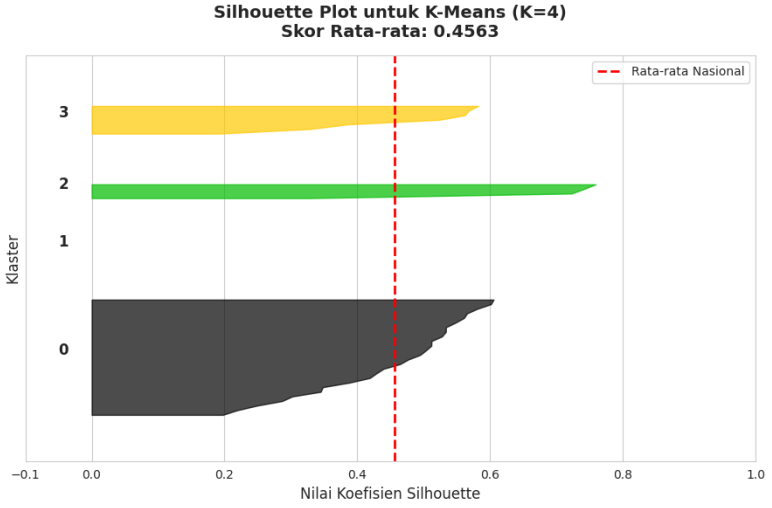


Gambar 4. Visualisasi K-Means *Clustering* dengan 3 Faktor

Berdasarkan Gambar 4 dapat terlihat hasil visualisasi dari K-Means Clustering dimana dapat terlihat bahwa proses clustering telah berhasil dilakukan di mana model dapat membagi *dataset* kedalam 4 kelompok berbeda yang sesuai dimana pada cluster 0 rata - rata berpusat di tengah dengan jumlah data paling banyak diantara cluster lainnya. Selain itu cluster 1 hanya memiliki 1 anggota saja dan posisinya paling jauh dari *cluster - cluster* lainnya.

Evaluasi Hasil *Clustering*

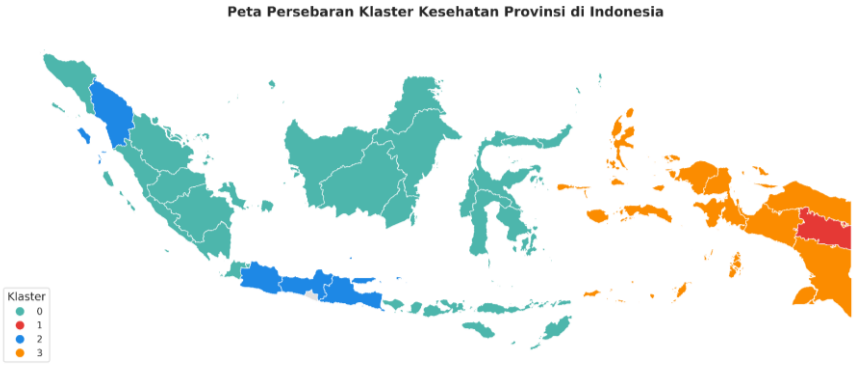
Setelah mengetahui hasil clustering, perlu dilakukan evaluasi dengan menggunakan *Silhouette Score* untuk menentukan tingkat kedekatan data dalam kelompok yang sama serta tingkat pemisah antar *cluster*.



Gambar 5 *Silhouette* Plot K=4

Berdasarkan visualisasi pada gambar 5, keempat klaster yang terbentuk mayoritas telah melampaui jumlah rata - rata nasional kecuali klaster 1. Secara keseluruhan K=4 mampu menghasilkan kualitas klaster berada pada kategori cukup baik dan mengelompokkan observasi sesuai karakteristik dengan *silhouette score* sebesar 0.4563.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi



Gambar 6. Peta Klaster Kesehatan

Berdasarkan Gambar 6, dapat diketahui profil dari masing-masing klaster. Klaster 0 merupakan profil menengah nasional yang mencakup sebagian besar wilayah Sumatera, Kalimantan, Sulawesi, hingga Bali dan DKI Jakarta dengan total jumlah 26 provinsi. Klaster 0 ini tidak memiliki permasalahan infrastruktur yang terlalu tertinggal, namun juga belum mencapai skala masif seperti provinsi-provinsi padat di Pulau Jawa. Kemudian Klaster 1 merupakan profil sangat kritis dimana hanya beranggotakan 1 anggota saja yaitu papua pegunungan dimana memiliki keterbatasan tenaga medis dan infrastruktur yang paling absolut dibandingkan seluruh provinsi lain, terpisahnya satu observasi menjadi klaster tersendiri menunjukkan adanya *outlier* data yang ekstrim. Klaster 2 Merupakan profil pusat demografi dan infrastruktur masif yang beranggotakan Jawa Barat, Jawa Timur, Jawa Tengah, dan Sumatera Utara, klaster mengelompokkan empat provinsi dengan jumlah populasi terbesar di Indonesia dan juga tingkat kepadatan penduduk yang ekstrem di wilayah ini diikuti oleh ketersediaan fasilitas kesehatan yang sangat masif, membedakannya secara signifikan dari provinsi lainnya. Klaster 3 Merupakan profil wilayah rentan dan terluar yang dimana laster ini mengelompokkan wilayah-wilayah di kawasan Timur Indonesia, yakni Maluku, Maluku Utara, Papua, Papua Barat, Papua Barat Daya, Papua Selatan, dan Papua Tengah. Klaster ini memiliki skala demografi populasi yang kecil dan memiliki tingkat kerentanan keluhan kesehatan di atas rata-rata dan memerlukan prioritas percepatan pemerataan fasilitas medis.

4. SIMPULAN

Berdasarkan analisis tersebut dapat disimpulkan bahwa analisis ini berhasil mengelompokkan 38 provinsi di Indonesia dengan indikator kesehatan multidimensional dengan menggunakan pendekatan analisis faktor dengan menggunakan *K-Means clustering.* Uji kelayakan menggunakan *Bartlett's Test* yang p-valuenya didapat bernilai 0,0000 yang bisa dikatakan cukup kuat korelasi antar variabelnya kemudian melakukan *KMO-Test* yang didapat bernilai 0,7479. Dari 12 variabel tersebut yang lolos MSA, di ekstrak lagi dengan tiga faktor laten optimal berdasarkan kaiser dan scree plot. Ketiga faktor yang ada adalah yaitu faktor kualitas layanan dan kesehatan dasar, faktor kapasitas infrastruktur dan demografi yang artinya integrasi skor faktor sebagai input K-Means telah terbukti mengatasi masalah multikolinearitas.

Penentuan klaster dengan metode *Elbow Method, Davies-Bouldin Index,* dan *Calinski Harabasz Inde*x telah ditetapkan dengan jumlah klaster sebesar K=4 yang telah divalidasi dengan silhouette score sebesar 0.4563. Keempat klaster tersebut berhasil mencerminkan gradasi kapasitas sistem kesehatan antar provinsi di Indonesia. Klaster 0 terdiri dari 26 provinsi dengan profil kesehatan menengah nasional, klaster 1 adalah outlier tunggal yaitu Papua Pegunungan dengan tingkat kerentanan dan keterbatasan infrastruktur paling kritis, klaster 2 terdiri dari Jawa Barat, Jawa Timur, Jawa Tengah, dan Sumatera Utara sebagai pusat demografi beserta kapasitas fasilitas kesehatan masif, dan klaster 3 mengelompokkan wilayah terluar kawasan Timur Indonesia dengan kerentanan kesehatan di atas rata-rata nasional. Dapat disimpulkan bahwa pendekatan berbasis struktur laten dapat menemukan pola tersembunyi yang tidak dapat dilakukan oleh klasterisasi konvensional, sehingga hasil dari temuan ini dapat menjadi landasan ilmiah bagi pemerintah untuk merancang kebijakan kesehatan yang tepat sasaran dan berkeadilan sesuai jenis masing-masing klaster.

DAFTAR PUSTAKA

[1] GBD 2019 Indonesia Subnational Collaborators, "The State of Health In Indonesia's Provinces, 1990–2019: a Systematic Analysis for the Global Burden of Disease Study 2019," *The Lancet Global Health*, vol. 10, no. 11, pp. e1632–e1645, Nov. 2022, doi: 10.1016/S2214-109X(22)00371-0.

[2] H. Maryani, A. Rizkianti, and N. Izza, "Classification of Healthy Family Indicators in Indonesia Based on a K-Means Cluster Analysis," *Journal of Preventive Medicine and Public Health*, vol. 57, no. 3, pp. 234–241, May 2024, doi: 10.3961/jpmph.23.497.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

[3] O. Arnilia, S. I. Ishak, T. Widodo, and I. G. N. A. B. Tatwa, "Clustering of Provincial Health Vulnerability Levels in Indonesia Using the K-Means Method," *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 7, no. 1, pp. 26–34, Mar. 2026, doi: 10.33365/jatika.v7i1.1469.

[4] World Health Organization, "New WHO Publication Highlights Importance of Health Inequality Monitoring to Advance Health Equity," *WHO Indonesia*. [Online]. Apr. 4, 2025. Available: https://www.who.int/indonesia/news/detail/04-04-2025-new-who-publication-highlights-the-importance-of-health-inequality-monitoring-to-advance-health-equity. [Accessed: May 30, 2026].

[5] G. F. Ibanez, G. W. Wiriasto, and Rosmaliati, "Kombinasi Principal Component Analysis dengan Algoritma K-Means untuk Klasterisasi Data Stunting," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 5, no. 1, 2024, doi: 10.30865/klik.v5i1.1977.

[6] A. K. Jain, "Data Clustering: 50 Years Beyond K-Means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011

[7] M. Shutaywi and N. N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering," *Entropy*, vol. 23, no. 6, p. 759, Jun. 2021, doi: 10.3390/e23060759.

[8] B. Sowan, T.-P. Hong, A. Al-Qerem, and C.-H. Chen, "Ensembling Validation Indices to Estimate the Optimal Number of Clusters," *Applied Intelligence*, vol. 53, pp. 9933–9957, Aug. 2022, doi: 10.1007/s10489-022-03939-w.

[9] A. Karanikola, C. M. Liapis, and S. Kotsiantis, "A comparative Study of Validity Indices on Estimating the Optimal Number of Clusters," *in Proc. 2021 12th Int. Conf. Information, Intelligence, Systems and Applications (IISA)*, IEEE, 2021, pp. 1–8, doi: 10.1109/IISA52424.2021.9555497.

[10] S. Haben, C. Singleton, and P. Grindrod, "Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 136–144, Jan. 2016, doi: 10.1109/TSG.2014.2325587.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi