Science and Education Journal
Vol 5 No 3 (2026)
E-ISSN : 2962-9713
P-ISSN : 2963-928X

Pengelompokan Status Gizi Balita Di Indonesia Menggunakan Algoritma Gaussian Mixture Model

Nur Fadila Syaus[1], Fahrezal Zubedi[2]
[12]Program Studi Statistika, Universitas Negeri Gorontalo, Indonesia,
nurf4dila5yaus@gmail.com

Abstrak

Status gizi balita adalah indikator penting dalam menilai kualitas kesehatan masyarakat dan keberhasilan pembangunan di Indonesia. Kondisi ini masih menunjukkan adanya perbedaan antarprovinsi yang dipengaruhi oleh berbagai faktor determinan. Perbedaan karakteristik ini menunjukkan perlunya pendekatan analisis yang mampu mengelompokkan wilayah berdasarkan kemiripan karakteristik indikator status gizi balita. Penelitian ini bertujuan menentukan jumlah cluster optimal dan mengidentifikasi karakteristik pengelompokan provinsi di Indonesia berdasarkan indikator status gizi balita menggunakan metode *Gaussian Mixture Model* (GMM) dengan validasi indeks. Data yang digunakan berasal dari publikasi Survei Status Gizi Indonesia (SSGI) Tahun 2024 yang mencakup 36 provinsi di Indonesia. Hasil penelitian menunjukkan bahwa jumlah *cluster* optimal adalah sebanyak dua cluster (K=2), berdasarkan hasil validasi *Davies-Bouldin Index* (1,065), *Silhouette Score* (0,327), serta didukung oleh nilai *Bayesian Information Criterion* (BIC) tertinggi sebesar (−2104,791). *Cluster* 1 terdiri dari 8 provinsi dengan kondisi pendukung status gizi balita yang relatif baik namun masih perlu peningkatan pada aspek sanitasi, sedangkan *cluster* 2 terdiri dari 28 provinsi dengan kondisi sanitasi yang relatif baik namun masih memiliki keterbatasan dalam aspek pendukung status gizi balita. Hasil ini menunjukkan bahwa GMM mampu mengidentifikasi karakteristik pengelompokan provinsi yang dapat menjadi dasar penyusunan kebijakan intervensi gizi.

Kata Kunci: *Gaussian Mixture Model*, Validasi Indeks, Pengelompokan, Status Gizi Balita.

Abstract

The nutritional status of toddlers is a key indicator for assessing public health and the success of development in Indonesia. This condition still shows disparities among provinces, influenced by various determining factors. These differences in characteristics highlight the need for an analytical approach capable of grouping regions based on similarities in the characteristics of toddler nutritional status indicators. This study aims to determine the optimal number of clusters and identify the characteristics of provincial groupings in Indonesia based on indicators of the nutritional status of infants and toddlers using the Gaussian Mixture Model (GMM) method with index validation. The data used were derived from the 2024 Indonesian Nutrition Status Survey (SSGI) publication, which covers 36 provinces in Indonesia. The results indicate that the optimal number of clusters is two (K=2), based on the Davies-Bouldin Index (1,065) and Silhouette Score (0,327), and supported by the highest Bayesian Information Criterion (BIC) value of (−2104,791). Cluster 1 consists of 8 provinces with relatively good conditions supporting the nutritional status of toddlers but still requiring improvement in sanitation, while Cluster 2 consists of 28 provinces with relatively good sanitation but still facing limitations in aspects supporting the nutritional status of toddlers. These results indicate that GMM is capable of identifying provincial clustering characteristics that can serve as a basis for policy formulation.

Keywords: *Gaussian Mixture Model, Index Validation, Clustering, Nutritional Status of Toddlers*.

E-ISSN : 2962-9713
P-ISSN : 2963-928X

Corresponding Author:
Nur Fadila Syaus
Universitas Negeri Gorontalo; nurf4dila5yaus@gmail.com

PENDAHULUAN

Implementasi statistika kesehatan memiliki peran strategis dalam mendukung pembangunan kesehatan masyarakat. Melalui analisis dan interpretasi data kesehatan dapat menjadi informasi yang akurat dan berbasis bukti. Informasi yang dihasilkan dari proses statistik dapat digunakan untuk mengidentifikasi permasalahan kesehatan, memantau kondisi kesehatan masyarakat, mengevaluasi program kesehatan, serta mendukung perencanaan dan pengambilan keputusan yang lebih efektif di bidang kesehatan (Saskirana et al., 2025). Pemantauan status gizi pada balita sangatlah krusial dilakukan guna memastikan bahwa balita memperoleh kebutuhan gizi secara seimbang dan memadai, dengan demikian proses perkembangan dan pertumbuhan seorang anak dapat berlangsung secara lancar dan bebas dari risiko kekurangan gizi (Jannah & Rosha, 2024) Status gizi balita dinilai berdasarkan indikator antropometri yang mencerminkan kondisi pertumbuhan anak berusia 0–60 bulan. Indikator tersebut meliputi *stunting*, *wasting*, *underweight*, dan *overweight* (Kementerian Kesehatan Republik Indonesia, 2020).

Berdasarkan data dari Survei Status Gizi Indonesia (SSGI) 2024, prevalensi *stunting* pada balita di Indonesia tercatat sebesar 19,8%, sedangkan prevalensi *underweight* sebesar 16,8%, *wasting* 6,2% dan *overweight* sebesar 3,4%. Selain itu, terdapat ketimpangan antar provinsi dalam pencapaian indikator gizi misalnya, prevalensi *stunting* di Nusa Tenggara Timur mencapai 37%, sedangkan di Bali kurang dari 10%, prevalensi *underweight* di Nusa Tenggara Timur mencapai 34,4%, sedangkan di Bali 7,2%, prevalensi *wasting* di di Kalimantan Timur 5,2%, sedangkan di Nusa Tenggara Timur mencapai 12,3%, prevalensi *overweight* di Bali mencapai 6,4%, sedangkan di Kalimantan Barat hanya 1,4% (Kementerian Kesehatan Republik Indonesia, 2024). Ketimpangan antar provinsi ini dapat disebabkan oleh berbagai faktor. Ketimpangan prevalensi antarprovinsi mencerminkan perbedaan yang nyata dalam pencapaian indikator status gizi balita di Indonesia, sehingga diperlukan analisis yang lebih mendalam untuk memahami pola dan karakteristik permasalahan gizi di setiap wilayah. Pengelompokan wilayah administratif di Indonesia berdasar pada indikator status gizi balita dilakukan untuk mengidentifikasi kelompok wilayah yang memiliki kesamaan karakteristik gizi maka hal tersebut mampu mendukung perancangan strategi intervensi dan kebijakan yang lebih tepat sasaran.

Beberapa penelitian sebelumnya telah dilakukan pengelompokan wilayah berdasarkan status gizi balita menggunakan analisis *cluster*. Penelitian yang dilakukan oleh (Nikmah et al., 2025) , yaitu mengelompokkan provinsi berdasarkan status gizi anak balita menggunakan Algoritma *K-Medoids*, penelitian ini menghasilkan 2 *cluster* terbaik berdasarkan evaluasi *davies-bouldin index* (DBI) dimana *cluster* 1 terdiri dari wilyah dengan status gizi yang memerlukan perhatian lebih sedangkan, *cluster* 2 memiliki status gizi balita yang lebih baik. Selain itu, (Jannah & Rosha, 2024) juga melakukan penelitian untuk mengelompokkan provinsi berdasar pada indikator status gizi balita menggunakan metode *cluster ensemble*, menghasilkan 2 *cluster* terbaik dengan nilai validitas *Silhouette Coefficient* dimana *cluster* 1 mencakup wilayah dengan permasalahan gizi kurang (17 provinsi)

E-ISSN : 2962-9713
P-ISSN : 2963-928X

sedangkan, *cluster* 2 mencakup wilayah dengan permasalahan gizi lebih (17 provinsi). Meskipun penelitian-penelitian tersebut memberikan gambaran awal mengenai pola pengelompokan provinsi, metode yang digunakan masih memiliki sejumlah keterbatasan dalam menangani karakteristik data status gizi balita. Baik metode *K-Medoids* maupun analisis *cluster ensemble* merupakan metode *cluster* berbasis jarak (*distance-based clustering*) yang menghasilkan pengelompokan secara tegas (*hard clustering*), sehingga setiap provinsi ditempatkan secara mutlak pada satu *cluster* tanpa mempertimbangkan tingkat probabilitas atau kedekatan relatif antar *cluster*. Pendekatan ini kurang mampu menangkap variasi dan kemiripan karakteristik antarprovinsi secara lebih fleksibel, terutama ketika pola data tidak sepenuhnya terpisah atau memiliki distribusi yang saling tumpang tindih (Vidyananta & Dermawan, 2025).

Mempertimbangkan keterbatasan metode hard clustering dalam merepresentasikan variasi dan tumpang tindih karakteristik antarprovinsi, *Gaussian Mixture Model* (GMM) dipilih sebagai alternatif karena dapat memberikan probabilitas keanggotaan pada setiap *cluster* dan mampu memodelkan distribusi data yang tidak terpisah secara tegas. GMM merupakan pendekatan *model-based clustering* yang memodelkan distribusi data sebagai campuran beberapa *gaussian*, sehingga mampu menangkap struktur *cluster* yang kompleks, serta memberikan probabilitas keanggotaan pada setiap provinsi (Riyono et al., 2022). Penelitian Sebelumnya yang dilakukan oleh (Setiawan & Kurniawan, 2023) menunjukkan bahwa metode *gaussian mixture model* memiliki kinerja pengelompokan yang lebih baik dibandingkan beberapa metode *distance-based* lainnya. Dalam penelitian tersebut, GMM dibandingkan dengan DBSCAN dan *agglomerative hierarchical cluster* dalam pengelompokan rumah sakit di Jakarta, dan hasilnya menunjukkan bahwa GMM memberikan nilai *Davies–Bouldin Index* terendah, sehingga dianggap sebagai model dengan kualitas *cluster* terbaik dengan menghasilkan 3 *cluster*. Selain itu, penelitian yang dilakukan (Prabiantissa & Yuliastuti, 2021) yaitu membandingkan GMM dengan *K-means* diperoleh hasil bahwa GMM menunjukan performa yang lebih unggul daripada *K-means*. Temuan ini mendukung bahwa pendekatan berbasis model memiliki keunggulan dalam mengidentifikasi struktur *cluster* yang kompleks. Untuk memperoleh hasil pengelompokan yang optimal, penggunaan GMM juga memerlukan penentuan jumlah *cluster* terbaik melalui proses validasi yang tepat.

Hal yang perlu diperhatikan dalam analisis *cluster* adalah proses validasi terhadap hasil pengelompokan yang terbentuk. Validasi ini merupakan hal yang krusial guna memastikan bahwa jumlah *cluster* yang dipilih sudah sesuai dengan struktur data (Saidah et al., 2022). Validitas *cluster* internal merupakan faktor utama yang mampu mempengaruhi validasi dalam memilih metode *clustering* terbaik maupun jumlah *cluster* optimal (Aselnino & Wijayanto, 2023). Pengukuran validasi internal pada metode *cluster* umumnya didasarkan pada dua aspek utama, yaitu tingkat kerapatan objek dalam satu *cluster* (*compactness*) dan tingkat perbedaan antar *cluster* (*separation*) (Gustriansyah et al., 2022).

Penelitian sebelumnya yang dilakukan oleh (Sibuea et al., 2025) mengenai pengklasteran kualitas air tambak dengan metode *gaussian mixture model* telah menunjukkan pentingnya proses validasi tersebut. Pada penelitian tersebut menggunakan validasi indeks *silhouette score* dan *davies-bouldin index*, sehingga diperoleh tiga *cluster* terbaik. Berdasarkan pendekatan serupa, penelitian ini akan melakukan pengelompokan wilayah administratif di Indonesia berdasar pada indikator status gizi balita menggunakan metode *Gaussian Mixture Model* dengan

E-ISSN : 2962-9713
P-ISSN : 2963-928X

penentuan jumlah *cluster* terbaik melalui dua validasi indeks, yaitu *Davies-Bouldin Index* (DBI) dan *Silhouette Score*, guna memperoleh struktur *cluster* yang paling representatif terhadap data. Hasil *cluster* yang diperoleh diharapkan dapat memberikan gambaran tentang pola persebaran masalah gizi balita antarprovinsi di Indonesia.

METODE

Penelitian ini menggunakan data sekunder yang diperoleh dari publikasi Survei Status Gizi Indonesia (SSGI) 2024. Populasi dalam penelitian ini adalah seluruh provinsi di Indonesia dengan indikator determinan status gizi balita pada tahun 2024. Sampel penelitian terdiri dari 36 provinsi, karena provinsi Papua Tengah dan Papua Pegunungan tidak disertakan akibat jumlah sampel survei yang belum mencukupi untuk ditampilkan dalam publikasi SSGI 2024 (Kementerian Kesehatan Republik Indonesia, 2024). Variabel yang digunakan dalam penelitian adalah Ibu Balita Yang Mendapatkan Tablet Tambah Darah (X_1), Balita Usia 0–5 Bulan Yang Mendapatkan ASI Eksklusif (X_2), Balita Yang Mendapatkan Pemantauan Pertumbuhan Sesuai Standar (X_3), Imunisasi Dasar Lengkap Pada Anak Usia 12–23 Bulan (X_4), Pengetahuan Dan Sikap Yang Tepat Tentang Dampak Stunting (X_5), Ibu Balita Yang Menggunakan KB Pasca Salin (X_6), Program Keluarga Harapan (PKH) Yang Diterima Di Rumah Tangga (X_7). Rumah Tangga Dengan Akses Sanitasi Layak Sendiri (X_8). Teknik penarikan sampel yang digunakan pada penelitian ini adalah teknik *Sampling* jenuh. Perhitungan analisis menggunakan perangkat lunak R *Studio*.

Metode utama yang digunakan dalam penelitian ini adalah *Gaussian Mixture Model* (GMM). GMM adalah suatu metode pengelompokkan berbasis pendekatan probabilistik, yaitu mengelompokkan data berdasarkan kesamaan pola distribusinya. Model ini berasumsi bahwa data yang digunakan mengikuti karakteristik dari beberapa distribusi *gaussian* (normal) dengan parameter yang tidak diketahui. Dalam GMM masing – masing cluster memiliki tiga parameter, yaitu mean (μ), kovarians (Σ), dan bobot (π) (Novkaniza et al., 2023). Berikut ini adalah tahapan-tahapan analisis yang dilakukan dalam penelitian ini.

1. Menginput data status gizi
2. Melakukan *pre-processing* data, tahap ini bertujuan untuk memastikan kualitas data yang akan digunakan telah sesuai (Pratiwi et al., 2026). Pada penelitian ini, akan dilakukan dua tahapan pre-processing yaitu:
   i. Melakukan deteksi *outlier* untuk mengidentifikasi nilai pengamatan yang menyimpang dari pola umum data, dengan menggunakan batas *interquartile range* (IQR), yaitu menentukan data yang berada di batas bawah $Q_1 - 1.5(\text{IQR})$ dan di batas atas $Q_3 + 1.5(\text{IQR})$ dengan visualisasi boxplot (Yolandari et al., 2025).
   ii. Melakukan penanganan *outlier* menggunakan *winsorizing* dengan teknik menyesuaikan nilai ekstrem agar tetap berada dalam rentang batas (IQR) (Putri et al., 2026).
3. Melakukan uji normalitas multivariat untuk melihat apakah sekelompok data dengan banyak variabel berdistribusi normal secara simultan (Ibrahim, 2024). Uji normal multivariat dilakukan menggunakan jarak mahalanobis dengan visualisasi *Q-Q plot*. Jika titik - titik pada plot mengikuti garis lurus dan lebih dari 50% nilai $d_i^2 \leq \chi_p^2(\alpha)$ maka data mengikuti distribusi normal multivariat, berlaku sebaliknya. Untuk menguji normalitas multivariat dapat ditunjukkan melalui persamaan berikut (Johnson & Wichern, 2007):

E-ISSN : 2962-9713
P-ISSN : 2963-928X

$$d_i^2 = (x_i - \mu)^T \Sigma^{-1} (x - \mu) \quad (1)$$

4. Melakukan analisis *cluster* menggunakan *gaussian mixture model* dengan algoritma *Expectation-Maximization* (EM) untuk k = 2, 3,..., dan 10, dengan tahapan berikut (Bishop, 2006):
   i. Inisialisasi parameter awal yaitu *mean* (μ), kovarians (Σ), dan bobot (π) secara acak untuk *j= 1, 2, ..., k*.
   ii. *Expectation step*, langkah ini dilakukan untuk menghitung probabilitas setiap titik data x_i menjadi anggota *cluster k* berdasarkan parameter saat ini, dengan rumus berikut:

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{k} \pi_j N(x_n|\mu_j,\Sigma_j)} \quad (2)$$

   Density distribusi *gaussian* dapat dihitung dengan menggunakan persamaan berikut:

$$N(x_n \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right) \quad (3)$$

   iii. *Maximization step*, langkah ini akan memperbarui parameter *mean* (μ), kovarians (Σ), dan bobot (π) masing - masing distribusi *gaussian* untuk memaksimalkan *log-likelihood* data berdasarkan probabilitas keanggotaan yang dihitung pada langkah *expectation*. Parameternya dihitung dengan menggunakan rumus berikut:

$$\mu_k^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \quad (4)$$

$$\Sigma_k^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k^{\text{new}}) (x_n - \mu_k^{\text{new}})^T}{\sum_{n=1}^{N} \gamma(z_{nk})} \quad (5)$$

$$\pi_k^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N} \quad (6)$$

   iv. Evaluasi *log-likelihood*, langkah ini dilakukan untuk evaluasi pada log-likelihood, dihitung menggunakan rumus berikut:

$$\ln p\,(X \mid \mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln\left\{\sum_{j=1}^{K} \pi_k \; N(X_n \mid \mu_k, \Sigma_k)\right\} \quad (7)$$

   v. Menghitung nilai *Bayes Information Criteria* untuk tiap jumlah *cluster* menggunakan rumus berikut (Tarumingkeng, 2025):

E-ISSN : 2962-9713
P-ISSN : 2963-928X

$$BIC = -2\ln(\hat{L}) + k\ln(n) \quad (8)$$

5. Validasi hasil *cluster* menggunakan validasi indeks:
   i. Menghitung nilai *davies-bouldin index* (DBI) dan *silhouette score* untuk tiap jumlah *cluster*. DBI dihitung menggunakan rumus (Jain & Dubes, 1988):

$$\mathrm{DB}(K) = \left(\frac{1}{K}\right)\sum_{k=1}^{K} R_k \qquad \text{for } K > 1 \quad (10)$$

$$R_k = \max_{j \neq k}\{R_{j,k}\} \quad (11)$$

$$R_{j,k} = \frac{e_j + e_k}{m_{j,k}} \quad (12)$$

silhouette score dihitung menggunakan rumus (Kaufman & Rousseeuw, 2005):

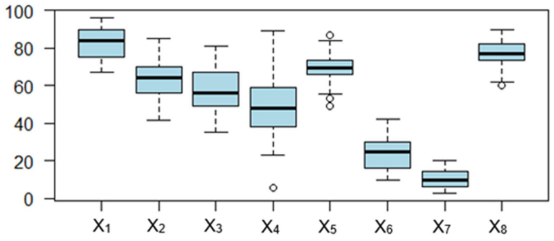
$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (13)$$

   ii. Membandingkan nilai DBI dan *Silhouette Score* untuk tiap jumlah *cluster*. Jumlah *cluster* dengan nilai DBI terendah dan nilai terbesar *Silhouette Score* menunjukan hasil *cluster* terbaik.
6. Melakukan interprestasi karakteristik *cluster* berdasrakan jumlah *cluster* yang terbaik.

HASIL DAN PEMBAHASAN

Pre-processing Data

Pre-processing data dilakukan guna memastikan kualitas data yang digunakan telah sesuai dan siap di analisis. Pada penelitian ini, tahapan *pre-processing* meliputi deteksi *outlier*, serta penanganan *outlier* menggunakan winsorizing. Tahapan awal dilakukan deteksi *outlier* menggunakan *interquartile range* (IQR), yaitu menentukan data yang berada di batas bawah $Q_1 - 1.5(\mathrm{IQR})$ dan di batas atas $Q_3 + 1.5(\mathrm{IQR})$. Data yang berada di luar kedua batas tersebut dikategorikan sebagai *outlier* dengan menggunakan visualisasi *boxplot*. Hasil deteksi *outlier* disajikan pada **Gambar 1**.



Gambar 1. Boxplot Data Status Gizi

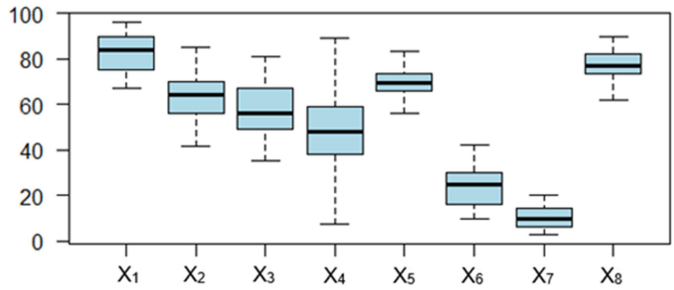
E-ISSN : 2962-9713
P-ISSN : 2963-928X

Berdasarkan **Gambar 1**, outlier teridentifikasi pada variabel X_4, X_5 dan X_8. Pada variabel X_4 terdapat satu outlier pada batas bawah. Variabel X_5 memiliki outlier pada batas bawah maupun batas atas. Variabel X_8 menunjukkan satu outlier pada batas bawah. Adanya outlier pada beberapa variabel dapat memengaruhi hasil analisis akhir. Oleh karena itu, diperlukan penanganan outlier agar distribusi data menjadi lebih stabil. Pada penelitian ini, penanganan outlier dilakukan menggunakan winsorizing yaitu dengan menyesuaikan nilai ekstrem berdasarkan batas atas dan batas bawah interquartile range (IQR) agar tetap berada dalam rentang tersebut. Hasil penanganan outlier disajikan pada **Tabel 3**.

Tabel 3. Hasil *Winsorizing*

Variabel	Batas Bawah	Batas Atas	Observasi	Sebelum *Winsorizing*	Sesudah *Winsorizing*
X_4	7,2	89,8	1	5,4	7,2
X_5	55,7	88,9	13	86,8	83,5
			14	84,2	83,5
			32	53,3	55,9
			33	55,3	55,9
			36	49,1	55,9
X_8	61,7	94,7	11	60	61,7

Berdasarkan **Tabel 3**, diperoleh hasil *winsorizing* pada variabel yang teridientifikasi *outlier*. Pada variabel X_4 nilai observasi ke-1 yang semula berada di bawah batas bawah sebesar 5,4 disesuaikan menjadi 7,2. Variabel X_5 nilai observasi ke-13 dan ke-14 yang melebihi batas atas disesuaikan menjadi 83,5, sedangkan observasi ke-32, ke-33, dan ke-36 yang berada di bawah batas bawah disesuaikan menjadi 55,9. Selain itu, pada variabel X_8 nilai observasi ke-11 yang berada di bawah batas IQR sebesar 60 disesuaikan menjadi 61,7. Adapun hasil visualisasi boxplot setelah *winsorizing* disajikan pada **Gambar 2**.



Gambar 2. Boxplot Setelah *Winsorizing*

Berdasarkan **Gambar 2**, diperoleh bahwa setelah proses winsorizing, tidak terlihat lagi adanya outlier pada variabel yang sebelumnya teridentifikasi memiliki outlier, sehiingga proses analisis dapat dilanjutkan.

Normal Multivariat

Uji normal multivariat dilakukan untuk melihat apakah data yang digunakan dalam penelitian ini mengikuti distribusi normal secara simultan pada seluruh variabel. Hasil uji normal multivariat dari data penelitian disajikan pada **Tabel 4**.

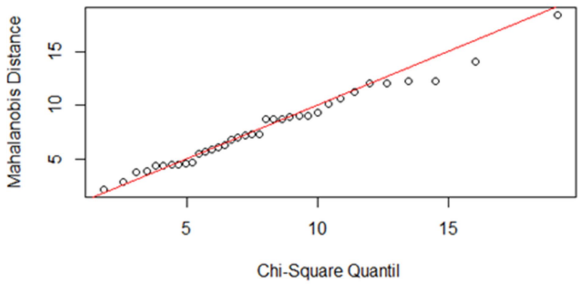
Tabel 4. Hasil Uji Normal Multivariat

Observasi	Jarak Mahalanobis	*Chi-Square*	Keputusan

E-ISSN : 2962-9713
P-ISSN : 2963-928X

1	2,147671	1,817243	Tolak H_0
2	2,908096	2,571091	Tolak H_0
3	3,754383	3,059696	Tolak H_0
4	3,832462	3,453377	Tolak H_0
5	4,375823	3,796515	Tolak H_0
6	4,384844	4,108302	Tolak H_0
7	4,452006	4,399064	Tolak H_0
8	4,472598	4,675133	Terima H_0
⋮	⋮	⋮	⋮
36	18,416639	19,187408	Terima H_0

Adapun hasil visualisasi pola sebaran data pada Q-Q plot antara jarak mahalanobis (*square distance*) dengan *chi-square* disajikan pada **Gambar 3**



Gambar 3. Q-Q Plot Normal Multivariat

Berdasarkan hasil dari uji normal multivariat pada **Gambar 3**, diperoleh bahwa titik - titik data cenderung mengikuti garis lurus dan berdasarkan **Tabel 4** diperoleh banyaknya nilai $d_i^2 \leq \chi_{36}^2(0{,}05)$ maka H_0 diterima, sehingga dapat disimpulkan bahwa data berdistribusi normal multivariat. Karena telah memenuhi asumsi dari metode gmm, maka semua variabel tersebut dapat digunakan pada penelitian ini.

Gaussian Mixture Model

Analisis *cluster* dilakukan menggunakan metode *Gaussian Mixture Model* (GMM) dengan algoritma *Expectation-Maximization* (EM), pada penelitian ini analisis dilakukan untuk jumlah *cluster* K = 2, 3,..., 10. Tahap awal dalam analisis GMM adalah menginisialisasi parameter awal yaitu *mean* (μ), kovarians (Σ), dan bobot (π) untuk setiap *cluster* secara acak. Kemudian parameter ini diestimasi menggunakan algoritma EM, proses iterasi antara E-step dan M-step dilakukan berulang kali sampai mencapai kondisi konvergen yaitu ketika perubahan sangat kecil. Setelah parameter model mencapai kondisi konvergen, diperoleh nilai probabilitas masing-masing provinsi terhadap setiap *cluster*. Suatu provinsi kemudian dimasukkan ke dalam *cluster* dengan nilai probabilitas terbesar. Selanjutnya untuk menentukan jumlah *cluster* yang paling optimal, dilakukan evaluasi model menggunakan kriteria BIC. Adapun hasil estimasi parameter GMM dengan EM untuk *cluster* K = 2 sebagai berikut.

Bobot (π):

Cluster 1	*Cluster 2*
$\mathbf{(0.217)}$	(0.782)

Mean (μ):

E-ISSN : 2962-9713
P-ISSN : 2963-928X

Cluster 1	*Cluster 2*
89,51099	80,852357
69,99110	62,515233
72,49268	53,345488
71,83575	42,157055
78,57230	66,817691
26,85823	22,626598
11,42958	9,819064
70,25223	78,503452

Kovarians (Σ):

Cluster 1	*Cluster 2*
57,6851	57,6851
107,3337	107,3337
60,51898	60,51898
152,1982	152,1982
26,37042	26,37042
66,67319	66,67319
21,85504	21,85504
46,73574	46,73574

Setelah parameter model mencapai kondisi konvergen, diperoleh nilai probabilitas masing-masing provinsi terhadap setiap *cluster*. Nilai probabilitas ini menunjukkan tingkat kedekatan suatu wilayah terhadap *cluster* tertentu dan digunakan sebagai dasar dalam penentuan hasil pengelompokan akhir dengan cara memilih nilai peluang yang terbesar. Hasil probabilitas setiap provinsi pada setiap *cluster* disajikan pada **Tabel 5**.

Tabel 5. Hasil Probabilitas Provinsi Pada Setiap *Cluster*

Provinsi	*Cluster 1*	*Cluster 2*	Probabilitas Terbesar	*Cluster*
Aceh	0,0000007	0,9999993	0,9999993	2
Sumatera Utara	0,0000003	0,9999997	0,9999997	2
Sumatera Barat	0,0030436	0,9969564	0,9969564	2
Riau	0,0000002	0,9999998	0,9999998	2
Jambi	0,0000002	0,9999998	0,9999998	2
Sumatera Selatan	0,0000038	0,9999961	0,9999961	2
Bengkulu	0,0000154	0,9999845	0,9999845	2
Lampung	0,0004622	0,9995377	0,9995377	2
Bangka Belitung	0,0000047	0,9999952	0,9999952	2
Kepulauan Riau	0,0000130	0,9999870	0,9999870	2
⋮	⋮	⋮	⋮	⋮
Papua Selatan	0,0000006	0,9999994	0,9999994	2

Berdasarkan **Tabel 5**, setiap provinsi memiliki nilai probabilitas terhadap masing-masing *cluster*. Sebagian besar provinsi memiliki nilai probabilitas yang sangat tinggi pada salah satu *cluster* (mendekati 1) dan sangat kecil pada *cluster* lainnya (mendekati 0). Hal ini menunjukkan bahwa struktur *cluster* yang terbentuk cukup jelas, sehingga sebagian besar provinsi memiliki karakteristik yang sangat

E-ISSN : 2962-9713
P-ISSN : 2963-928X

sesuai dengan satu *cluster* tertentu. Sebagai contoh, Aceh memiliki probabilitas terbesar pada *cluster* 2 (0,9999993) dan pada *cluster* 1 (0,0000007). Hal ini menunjukkan bahwa hampir seluruh karakteristik Aceh mengikuti pola *cluster* 2, sementara tingkat kesesuaiannya terhadap *cluster* 1 sangat kecil, yang berarti karakteristik Aceh hampir tidak merepresentasikan distribusi *cluster* tersebut. Tahapan yang sama selanjutnya diterapkan untuk jumlah *cluster* K = 3, 4,..., 10. Selanjutnya, nilai BIC dari masing-masing jumlah *cluster* dibandingkan untuk menentukan model terbaik. Model dengan nilai BIC tertinggi dipilih sebagai model yang paling optimal dalam merepresentasikan struktur data. Hasil dari BIC disajikan pada **Tabel 6**.

Tabel 6. Hasil BIC

Jumlah *Cluster*	Model	BIC
$k = 2$	EEI	-2104,791
$k = 3$	EEI	-2105.169
$k = 4$	EEI	-2207,326
⋮	⋮	⋮
$k = 10$	EEI	-2200,234

Berdasarkan **Tabel 6**, diperoleh nilai *Bayesian Information Criterion* (BIC) untuk jumlah *cluster* K = 2, 3,..., 10 dengan model terbaik adalah EEI. Penentuan jumlah *cluster* terbaik ditunjukkan pada nilai BIC yang lebih besar. Nilai BIC terbesar diperoleh pada K = 2 yaitu sebesar (-2104,791), maka jumlah *cluster* optimal adalah jumlah *cluster* K = 2.

Validasi Indeks

Validasi indeks dilakukan untuk mengevaluasi jumlah cluster dan menentukan jumlah cluster terbaik yang dihasilkan oleh metode clustering yang digunakan, sehingga pengelompokan yang diperoleh sesuai dengan karakteristik data penelitian. Penelitian ini menggunakan validasi indeks yaitu Davies-Bouldin Index (DBI) dan Silhouette Score. Jumlah cluster terbaik ditentukan berdasarkan nilai DBI terkecil dan silhouette score yang lebih besar. Hasil beberapa validasi indeks pada jumlah cluster K = 2, 3,..., 10 disajikan pada **Tabel 7**.

Tabel 7. Hasil Validasi Indeks

Validasi Indeks	Jumlah *Cluster*			
	2	**3**	⋯	**10**
Davies-Bouldin Index	1,065	1,511	⋯	1,294
Silhouette Score	0,327	0,209	⋯	0, 148

Berdasarkan **Tabel 7,** diperoleh bahwa nilai DBI terkecil pada K = 2 yaitu (1,065) dan *silhouette score* yang lebih besar pada K = 2 yaitu (0,327). Sehingga, K = 2 dipilih sebagai jumlah *cluster* terbaik karena didukung oleh dua validasi indeks. Selain itu, hasil ini juga merujuk pada **Tabel 6** Hasil BIC yang menunjukkan bahwa K = 2 merupakan jumlah *cluster* optimal, sehingga K = 2 ditetapkan sebagai jumlah *cluster* optimal yang paling merepresentasikan struktur data pada penelitian ini.
Berdasarkan **Tabel 7** dapat diketahui bahwa jumlah *cluster* terbaik adalah K = 2 karena didukung oleh dua validasi indeks, serta merujuk pada **Tabel 6** Hasil BIC yang menunjukkan bahwa K = 2 merupakan jumlah *cluster* optimal. Profilisasi hasil pengelompokan pada jumlah cluster terbaik ditampilkan pada **Tabel 8**.

Tabel 8. Hasil *Cluster*

Cluster	Jumlah Anggota	Provinsi

E-ISSN : 2962-9713
P-ISSN : 2963-928X

1	8	DKI Jakarta, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, Bali, Nusa Tenggara Barat, dan Nusa Tenggara Timur.
2	28	Aceh, Sumatera Utara, Sumatera Barat, Sumatera Selatan, Riau, Jambi, Bengkulu, Lampung, Bangka Belitung, Kepulauan Riau, Banten, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Papua Barat Daya, Papua, dan Papua Selatan.

Setiap *cluster* merepresentasikan kelompok wilayah dengan karakteristik yang berbeda, sehingga diperlukan identifikasi lebih lanjut untuk menggambarkan ciri spesifik dari setiap *cluster*. Untuk mengidentifikasi karakteristik setiap *cluster* yang terbentuk, nilai rata-rata setiap variabel pada masing-masing *cluster* dianalisis untuk melihat kecenderungan dan ciri khas setiap kelompok, yang disajikan pada **Tabel 9**.

Tabel 9. Rata – Rata Setiap Variabel

Variabel	*Cluster 1*	*Cluster 2*
X_1	**89,51099**	80,852357
X_2	**69,99110**	62,515233
X_3	**72,49268**	53,345488
X_4	**71,83575**	42,157055
X_5	**78,57230**	66,817691
X_6	**26,85823**	22,626598
X_7	**11,42958**	9,819064
X_8	70,25223	**78,503452**

Berdasarkan **Tabel 9** diperoleh rata - rata dari setiap variabel pada masing - masing *cluster*, sehingga diberikan interpretasi sebagai berikut.

1. *Cluster* 1 dicirikan oleh nilai rata-rata yang relatif tinggi pada sebagian besar indikator. Kondisi ini menunjukkan bahwa wilayah pada *cluster* 1 memiliki kondisi yang relatif baik pada sebagian besar indikator pendukung status gizi balita. Namun demikian, pada indikator sanitasi layak sendiri, *cluster* 1 mencakup nilai rata-rata yang relatif lebih rendah daripada *cluster* 2. Maka, *cluster* 1 merepresentasikan wilayah dengan kondisi pendukung status gizi balita yang relatif baik, namun masih perlu peningkatan pada aspek sanitasi.
2. *Cluster* 2 dicirikan oleh nilai rata-rata yang rendah pada hampir seluruh indikator. Kondisi ini menunjukkan bahwa wilayah pada *cluster* 2 memiliki kondisi yang relatif rendah pada sebagian besar indikator pendukung status gizi balita. Namun demikian, pada indikator sanitasi layak sendiri, *cluster* 2 mencakup nilai rata-rata yang relatif lebih tinggi daripada *cluster* 1. Dengan demikian, *cluster* 2 merepresentasikan wilayah dengan kondisi sanitasi yang relatif baik, namun masih memiliki keterbatasan pada sebagian besar indikator pendukung status gizi balita.

SIMPULAN

Berdasarkan hasil dan pembahasan, diperoleh jumlah *cluster* optimal dalam pengelompokan provinsi di Indonesia berdasarkan indikator status gizi balita

E-ISSN : 2962-9713
P-ISSN : 2963-928X

menggunakan *Gaussian Mixture Model* (GMM) adalah sebanyak dua *cluster* (K=2). Penentuan ini didasarkan pada hasil validasi indeks, yaitu *davies-bouldin index* dan *silhouette score*, yang menunjukkan hasil terbaik pada K=2 dengan nilai DBI (1,065) dan *silhouette score* (0,327) serta didukung oleh nilai *bayesian information criterion* (BIC) tertinggi pada jumlah *cluster* K=2 yaitu sebesar (-2104,791), maka K=2 ditetapkan sebagai jumlah *cluster* yang paling optimal. Hasil pengelompokan memperoleh dua kelompok provinsi dengan karakteristik yang berbeda, yaitu *cluster* 1 yang terdiri atas 8 provinsi dan *cluster* 2 yang terdiri atas 28 provinsi.

Cluster 1 dicirikan oleh nilai rata-rata yang relatif tinggi pada sebagian besar indikator, sehingga merepresentasikan wilayah dengan kondisi pendukung status gizi balita yang relatif baik, meskipun masih memerlukan peningkatan pada aspek sanitasi. Sebaliknya, *cluster* 2 memiliki nilai rata-rata yang relatif rendah pada sebagian besar indikator, sehingga merepresentasikan wilayah dengan kondisi sanitasi yang relatif baik, namun masih memiliki keterbatasan dalam aspek pendukung status gizi balita. Hasil penelitian ini menunjukkan adanya perbedaan karakteristik antarprovinsi dalam faktor-faktor yang mempengaruhi status gizi balita, sehingga dapat digunakan sebagai landasan dalam penyusunan kebijakan dan intervensi yang lebih tepat sasaran sesuai dengan karakteristik masing-masing kelompok wilayah.

DAFTAR PUSTAKA

Aselnino, P., & Wijayanto, A. W. (2023). Analisis Perbandingan Metode Hierarchical dan Non-Hierarchical dalam Pembentukan Cluster Provinsi di Indonesia Berdasarkan Indikator Women Empowerment. *Indonesian Journal of Applied Statistics*, *6*(1), 57–68.

Bishop, C. M. (2006). *Pattern Recognition And Machine Learning* (M. Jordan, J. Kleinberg, & B. Scho (eds.)). Springer. https://link.springer.com/book/9780387310732

Gustriansyah, R., Alie, J., & Suhandi, N. (2022). Hierarchical clustering for crime rate mapping in Indonesia. *ILKOM Jurnal Ilmiah*, *14*(3), 275–283.

Ibrahim, N. S. (2024). *Analisis Diskriminan Linear Robust dengan Penduga Minimum Covariance Determinant (Studi Kasus : Indeks Kerentanan Pangan Menurut Kabupaten / Kota di Indonesia Tahun 2023)*. *2*(2), 264–279.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc. https://books.google.co.id/books/about/Algorithms_for_Clustering_Data.html?id=7eBQAAAAMAAJ&redir_esc=y

Jannah, R., & Rosha, M. (2024). Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Status Gizi Balita dengan Analisis Cluster Ensemble . *Jurnal Pendidikan Tambusai*, *8*(2), 22741–22751. https://jptam.org/index.php/jptam/article/view/16982

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis (Sixth Edition)*. Prentice hall Upper Saddle River.

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data: an introduction to cluster analysis*. John Wiley \& Sons. https://www.perlego.com/book/2760052/finding-groups-in-data-an-introduction-to-cluster-analysis-pdf

Kementerian Kesehatan Republik Indonesia. (2020). *Peraturan Menteri Kesehatan Republik Indonesia Nomor 2 Tahun 2020 tentang Standar Antropometri Anak*.

Kementerian Kesehatan Republik Indonesia. (2024). *Survei Status Gizi Indonesia (SSGI) 2024 Dalam Angka*. Kementerian Kesehatan Republik Indonesia.

Nikmah, A., Nisa, C., & Riefky, M. (2025). Penerapan Algoritma K-Medoids untuk

E-ISSN : 2962-9713
P-ISSN : 2963-928X

Pengelompokkan Provinsi di Indonesia Berdasarkan Status Gizi Anak Balita. *Emerging Statistics and Data Science Journal*, *3*(1), 516–524.

Novkaniza, F., Nico, & Al Kafi, R. (2023). PEMODELAN JUMLAH KASUS BARU HARIAN COVID-19 DI INDONESIA MENGGUNAKAN GAUSSIAN MIXTURE MODEL. *Jurnal Riset Dan Aplikasi Matematika (JRAM)*, *07*(02), 116–127.

Prabiantissa, C. N., & Yuliastuti, G. E. (2021). Prediksi Pergerakan Ikan Di Pesisir Pulau Madura Menggunakan Metode Gaussian Mixture Model Dan K-Means Clustering. *Jurnal Teknologi Informasi Dan Terapan*, *8*(2), 121–128.

Pratiwi, M. P., Natalia, E. Y., & Fauzi, R. (2026). Desain Sistem Big Data untuk Deteksi Pola Perilaku Konsumen dengan Menggunakan Algoritma Machine Learning. *JURNAL DESAIN DAN ANALISIS TEKNOLOGI (JDDAT)*, *5*, 12–19.

Putri, M. S., Junaidi, S., & Mardiyah, A. (2026). Pemodelan dan Prediksi Tingkat Kemiskinan Provinsi Sumatera Barat Menggunakan Support Vector Machine. *Jurnal Computer Science and Information Technology (CoSciTech)*, *7*(1), 130–140.

Riyono, J., Puspa, S. D., & Pujiastuti, C. E. (2022). Simulasi Clustering Provinsi di Indonesia dalam Penyebaran Covid-19 Berdasarkan Indikator Kesehatan Masyarakat Menggunakan Algoritma Gaussian Mixture Model. *Majamath: Jurnal Matematika Dan Pendidikan Matematika*, *5*(1), 43–60.

Saidah, D. A., Santoso, R., & Widiharih, T. (2022). Pengelompokan provinsi di indonesia berdasarkan indikator kesehatan lingkungan menggunakan metode partitioning around medoids dengan validasi indeks internal. *Jurnal Gaussian*, *11*, 302–312.

Saskirana, M. A., Pardede, N. A., & Alfiansyah, R. (2025). Analisis Data Statistik dalam Pengembangan Aplikasi Mobile untuk Kesehatan Masyarakat. *Journal Computer Science and Information Technology (JCoInT)*, *6*(1), 394–403.

Setiawan, K. E., & Kurniawan, A. (2023). Pengelompokan Rumah Sakit di Jakarta Menggunakan Model DBSCAN, Gaussian Mixture, dan Hierarchical Clustering. *Jurnal Informatika Terpadu*, *9*(2), 149–156.

Sibuea, A. Y., Asrianda, & Aidilof, H. A. (2025). Pengklasteran Data Kualitas Air Tambak Menggunakan Metode Gaussian Mixture Model. *RABIT : Jurnal Teknologi Dan Sistem Informasi Univrab*, *10*(2), 957–966.

Tarumingkeng, R. C. (2025). *Bayes Information Criteria*. RUDYCT e-PRESS.

Vidyananta, I. P. A., & Dermawan, K. T. (2025). PERBANDINGAN ALGORITMA K-MEANS DAN K-MEDOIDS DALAM PENGELOMPOKAN PROVINSI DI INDONESIA BERDASARKAN INDIKATOR KEADAAN SEKOLAH DASAR. *Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(3S1).

Yolandari, N. A., Butarbutar, L. E., Rajagukguk, G. C. H., Zulfi, M. F., Arnita, & Ramadhani, F. (2025). ANALISIS PERBANDINGAN K-MEANS DAN DBSCAN DALAM PENGELOMPOKAN DATA TRAVEL REVIEW RATINGS MENGGUNAKAN EVALUASI SILHOUETTE INDEX DAN DAVIES-BOULDIN INDEX. *JITET (Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(3).