Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Oct 2026, IX (5): 8939 – 8945 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

PENERAPAN ALGORITMA K-MEANS UNTUK KLASTERISASI PROVINSI DI INDONESIA BERDASARKAN POLA ANGKA KELAHIRAN MENURUT KELOMPOK UMUR

Yanti[1], Silvi Joya[2], Mahzuroh[3]
Universitas Pembangunan Panca Budi Medan, Sumatera Utara
e-mail: [1]choy.yanti@gmail.com, [2]joyasilvi2001@gmail.com, [3]mahzuros@gmail.com

Abstract: *Differences in fertility patterns across Indonesian provinces can be explored through age-specific fertility rates. This study clusters 38 provinces using seven Age-Specific Fertility Rate (ASFR) indicators for women aged 15–19 to 45–49, based on the 2025 Intercensal Population Survey (SUPAS) table published by Statistics Indonesia (BPS). The indicators represent births per 1,000 women in the corresponding age group. The workflow includes data checks, z-score standardization, K-Means clustering with 50 initializations for candidate values of k=2–7, and evaluation using inertia, silhouette coefficient, Davies–Bouldin index, and Calinski–Harabasz index. The k=2 solution achieved the highest silhouette score (0.295), resulting in clusters of 27 and 11 provinces. Centroid profiles indicate differences in ASFR distributions across age groups. However, the lowest Davies–Bouldin index occurred at k=6; therefore, k=2 is treated as a parsimonious primary solution rather than the only plausible partition. Findings are exploratory and do not establish causal relationships or rank provinces normatively. Further work should test cluster stability, compare alternative algorithms, and incorporate socio-demographic covariates.*

Keyword: *ASFR, K-Means, provincial clustering, fertility, SUPAS 2025.*

Abstrak: Perbedaan pola fertilitas antarprovinsi dapat dipelajari melalui distribusi angka kelahiran menurut kelompok umur ibu. Penelitian ini bertujuan mengelompokkan 38 provinsi di Indonesia berdasarkan tujuh indikator Age-Specific Fertility Rate (ASFR) kelompok umur 15–19 hingga 45–49 tahun dari hasil Survei Penduduk Antar Sensus (SUPAS) 2025. Data bersumber dari tabel statistik Badan Pusat Statistik (BPS) dan dinyatakan sebagai jumlah kelahiran per 1.000 perempuan pada kelompok umur yang sama. Tahapan analisis meliputi pemeriksaan data, standardisasi z-score, penerapan K-Means dengan 50 inisialisasi untuk setiap kandidat jumlah klaster (k=2–7), serta evaluasi menggunakan inertia, silhouette coefficient, Davies–Bouldin index, dan Calinski–Harabasz index. Kandidat k=2 menghasilkan silhouette tertinggi sebesar 0.295 dan membentuk klaster beranggotakan 27 serta 11 provinsi. Profil centroid menunjukkan perbedaan pola ASFR lintas kelompok umur; klaster dengan nilai rata-rata lebih tinggi menonjol pada beberapa kelompok umur, terutama kelompok umur lebih tua. Namun, Davies–Bouldin index terendah diperoleh pada k=6, sehingga pemilihan k=2 diposisikan sebagai solusi utama yang lebih sederhana dan didukung silhouette, bukan sebagai satu-satunya struktur yang mungkin. Hasil ini bersifat eksploratif dan tidak menyiratkan hubungan sebab-akibat atau penilaian normatif terhadap provinsi. Penelitian lanjutan disarankan menguji kestabilan klaster, membandingkan algoritma alternatif, dan menggabungkan kovariat sosial-demografi.

Kata kunci: ASFR, K-Means, klasterisasi provinsi, fertilitas, SUPAS 2025.

PENDAHULUAN

Fertilitas merupakan komponen utama dinamika penduduk bersama mortalitas dan migrasi. Untuk memahami fertilitas secara lebih rinci, angka

kelahiran menurut kelompok umur atau Age-Specific Fertility Rate (ASFR) menunjukkan banyaknya kelahiran hidup pada suatu tahun per 1.000 perempuan pada kelompok umur yang sama. Profil ASFR membantu menunjukkan kapan fertilitas terkonsentrasi sepanjang usia reproduksi dan bagaimana profil tersebut berbeda antarwilayah. BPS memublikasikan indikator ASFR provinsi berdasarkan hasil SUPAS 2025, yang mencakup 38 provinsi dan tujuh kelompok umur Badan Pusat Statistik (2026).

Analisis deskriptif satu indikator pada satu waktu dapat mengungkap nilai tinggi atau rendah, tetapi belum tentu menangkap kemiripan profil secara serentak pada seluruh kelompok umur. Misalnya, dua provinsi dapat memiliki ASFR pada usia 25–29 yang serupa tetapi berbeda pada usia 15–19 atau 40–44. Karena itu, pendekatan multivariat dibutuhkan untuk menyusun kelompok provinsi berdasarkan keseluruhan profil, bukan hanya satu nilai.

Klasterisasi merupakan teknik pembelajaran tanpa pengawasan yang mengelompokkan objek berdasarkan kemiripan fitur. K-Means banyak digunakan karena relatif sederhana, efisien, dan centroid-nya mudah ditafsirkan MacQueen (1967), Hartigan & Wong (1979). Meski demikian, hasil K-Means dipengaruhi skala fitur, inisialisasi, outlier, dan jumlah klaster yang dipilih. Evaluasi sebaiknya tidak hanya bergantung pada inertia, melainkan mempertimbangkan metrik validasi internal yang berbeda Rousseeuw (1987)–Caliński & Harabasz (1974).

Celah praktis yang ditangani penelitian ini adalah kebutuhan pemetaan awal pola ASFR lintas 38 provinsi dengan memanfaatkan seluruh tujuh kelompok umur dalam satu proses analitik dan melaporkan dasar pemilihan jumlah klaster secara transparan. Kontribusi penelitian bukan mengklaim algoritma baru, melainkan (i) menyajikan alur reprodusibel untuk mengelompokkan profil ASFR SUPAS 2025, (ii) membandingkan beberapa nilai k dengan empat metrik evaluasi, dan (iii) menginterpretasikan centroid serta keanggotaan klaster dalam konteks demografi secara hati-hati.

Pertanyaan penelitian dirumuskan sebagai berikut: (1) bagaimana provinsi Indonesia dapat dikelompokkan berdasarkan profil ASFR tujuh kelompok umur pada SUPAS 2025; (2) berapa jumlah klaster yang paling layak menurut evaluasi internal dan pertimbangan interpretabilitas; dan (3) bagaimana karakteristik rata-rata serta sebaran anggota dari klaster yang terbentuk?.

TINJAUAN PUSTAKA

Age-Specific Fertility Rate

ASFR adalah ukuran fertilitas spesifik umur yang menghitung banyaknya kelahiran hidup per 1.000 perempuan pada kelompok umur tertentu. Berbeda dari Total Fertility Rate (TFR), ASFR mempertahankan informasi tentang distribusi fertilitas pada rentang usia 15–49 tahun. BPS menjelaskan bahwa tabel SUPAS 2025 memuat kelompok umur 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, dan 45–49 Badan Pusat Statistik (2026). Dalam penelitian ini, ketujuh kelompok umur menjadi fitur numerik, sedangkan provinsi menjadi unit observasi.

K-Means

K-Means membagi n observasi menjadi K klaster dengan meminimalkan jumlah kuadrat jarak observasi terhadap centroid klaster masing-masing. Untuk data terstandardisasi x_i dan centroid μ_k, fungsi objektifnya dapat ditulis sebagai:

$$J = \Sigma_{k=1}^{K} \Sigma_{xi \in C_k} ||x_i - \mu_k||^2$$

Algoritma secara iteratif (a) mengalokasikan setiap observasi ke centroid terdekat dan (b) memperbarui centroid sebagai rata-rata observasi anggota klaster hingga konvergen. Metode ini berakar pada penelitian MacQueen MacQueen (1967) dan implementasi algoritmik yang dibahas

Hartigan dan Wong Hartigan & Wong (1979). Karena fungsi objektif mengandalkan jarak Euclidean, standardisasi menjadi penting ketika fitur memiliki skala atau variasi berbeda.

Standardisasi dan Validasi Internal

Standardisasi z-score dihitung dengan $z_{ij}=(x_{ij}-\mu_j)/\sigma_j$, dengan μ_j dan σ_j masing-masing merupakan rata-rata dan simpangan baku fitur j. Proses ini menempatkan fitur pada skala yang sebanding. Silhouette mengukur seberapa dekat observasi dengan klasternya sendiri dibandingkan klaster terdekat lain; nilai yang lebih tinggi umumnya lebih baik Rousseeuw (1987). Davies–Bouldin membandingkan kekompakan dan pemisahan klaster, dengan nilai lebih rendah umumnya lebih baik Davies & Bouldin (1979). Calinski–Harabasz membandingkan dispersi antar-klaster terhadap dispersi dalam klaster; nilai lebih tinggi umumnya lebih baik Caliński & Harabasz (1974). Inertia adalah jumlah kuadrat jarak ke centroid dan biasanya menurun saat k bertambah, sehingga harus dibaca bersama metrik lain.

Posisi dan Kebaruan Penelitian

Kajian klasterisasi yang baik perlu menjelaskan objek, fitur, preprocessing, parameter, pemilihan k, dan keterbatasan interpretasi. Penelitian ini menerapkan prinsip tersebut pada profil ASFR tujuh dimensi di tingkat provinsi. Kebaruan yang diklaim bersifat kontekstual dan metodologis: analisis semua kelompok umur sekaligus pada tabel SUPAS 2025, pelaporan empat metrik validasi untuk rentang k yang sama, serta penekanan pada ketidakpastian pemilihan k ketika metrik tidak sepenuhnya sejalan. Klaim ini tidak berarti metode K-Means baru secara algoritmik.

METODE

Desain dan Sumber Data

Penelitian ini merupakan penelitian kuantitatif eksploratif menggunakan data sekunder. Sumber data adalah tabel statistik BPS berjudul "Angka Kelahiran Menurut Kelompok Umur dan Provinsi, Hasil SUPAS, 2025", yang diperbarui pada 28 Agustus 2026 Badan Pusat Statistik (2026). Data yang diberikan dalam bentuk Excel dicocokkan dengan struktur tabel tersebut. Setiap baris mewakili provinsi, dan tujuh kolom fitur merupakan ASFR kelompok umur. Satuan indikator adalah kelahiran per 1.000 perempuan pada kelompok umur yang sama Badan Pusat Statistik (2026).

Tabel sumber mencantumkan 38 provinsi serta baris agregat "INDONESIA". Baris agregat nasional dikeluarkan karena bukan unit provinsi dan dapat mendistorsi pembentukan klaster. Dengan demikian, sampel analisis terdiri atas 38 observasi. Seluruh tujuh fitur memiliki nilai numerik yang tersedia sehingga tidak diperlukan imputasi pada data yang dianalisis.

Variabel Penelitian

Tabel 1 Variabel Penelitian

Variabel	Definisi operasional	Peran
Provinsi	Nama 38 provinsi di Indonesia	ID/label observasi
ASFR 15–19	Kelahiran per 1.000 perempuan usia 15–19 tahun	Fitur
ASFR 20–24	Kelahiran per 1.000 perempuan usia 20–24 tahun	Fitur
ASFR 25–29	Kelahiran per 1.000 perempuan usia 25–29 tahun	Fitur
ASFR 30–34	Kelahiran per 1.000 perempuan usia 30–34 tahun	Fitur
ASFR 35–39	Kelahiran per 1.000 perempuan usia 35–39 tahun	Fitur
ASFR 40–44	Kelahiran per 1.000 perempuan usia 40–44 tahun	Fitur
ASFR 45–49	Kelahiran per 1.000 perempuan usia 45–49 tahun	Fitur

Pra-pemrosesan

Nama provinsi dibersihkan dari spasi di awal dan akhir; nilai tujuh fitur dikonversi ke numerik; baris dengan nilai fitur yang tidak lengkap dikeluarkan; dan baris agregat Indonesia dikecualikan. Setelah itu, setiap fitur distandardisasi menggunakan StandardScaler. Standardisasi digunakan untuk mencegah variabel dengan variasi absolut besar—misalnya ASFR usia 25–29—mendominasi jarak hanya karena rentang nilainya lebih besar. Nilai centroid untuk interpretasi dikembalikan ke skala asli melalui penghitungan rata-rata data asli pada anggota klaster.

Implementasi K-Means dan Penentuan k

K-Means dijalankan pada data terstandardisasi untuk kandidat k=2, 3, 4, 5, 6, dan 7. Parameter yang digunakan adalah random_state=42 dan n_init=50. Penggunaan banyak inisialisasi membantu memilih solusi dengan inertia rendah di antara beberapa titik awal dan mendukung reprodusibilitas pada versi pustaka yang sebanding. Jarak Euclidean digunakan oleh implementasi K-Means.

Pemilihan jumlah klaster dilakukan dengan mempertimbangkan empat ukuran: (i) inertia untuk melihat perubahan jumlah kuadrat jarak dalam klaster; (ii) silhouette coefficient yang diharapkan lebih tinggi; (iii) Davies–Bouldin index yang diharapkan lebih rendah; dan (iv) Calinski–Harabasz index yang diharapkan lebih tinggi. Karena metrik dapat memberikan rekomendasi berbeda, pemilihan akhir mempertimbangkan kombinasi bukti dan kemudahan interpretasi, bukan satu angka tunggal.

Alur Penelitian

1. Pengambilan data sekunder BPS SUPAS 2025 dan pemeriksaan struktur tabel.
2. Pembersihan data, pengeluaran agregat nasional, serta penetapan 38 provinsi sebagai unit observasi.
3. Standardisasi tujuh fitur ASFR menggunakan z-score.
4. Eksperimen K-Means untuk k=2 sampai k=7 dengan 50 inisialisasi.
5. Evaluasi menggunakan inertia, silhouette, Davies–Bouldin, dan Calinski–Harabasz.
6. Interpretasi centroid pada skala ASFR asli dan pemeriksaan daftar anggota klaster.

Lingkungan Analisis dan Reprodusibilitas

Analisis dilakukan dengan Python, pandas untuk pengolahan data, scikit-learn untuk standardisasi, K-Means dan metrik validasi, serta Matplotlib untuk visualisasi profil centroid. Parameter random_state=42 dan n_init=50 dilaporkan agar analisis dapat direplikasi. Untuk reproduksi lintas lingkungan, peneliti perlu mencatat versi Python, pandas, scikit-learn, dan NumPy saat menjalankan ulang analisis.

HASIL DAN PEMBAHASAN

Karakteristik Data

Setelah baris agregat Indonesia dikeluarkan, data analisis terdiri atas 38 provinsi dan tujuh fitur ASFR. Puncak ASFR pada tingkat nasional berada pada kelompok umur 25–29 tahun, dengan nilai 124,95 kelahiran per 1.000 perempuan pada kelompok umur tersebut Badan Pusat Statistik (2026). Nilai nasional ini digunakan hanya sebagai konteks deskriptif dan tidak dimasukkan sebagai observasi klaster. Pengelompokan pada penelitian ini didasarkan pada pola relatif tujuh fitur di setiap provinsi setelah standardisasi.

Evaluasi Kandidat Jumlah Klaster

Tabel 2 Hasil Evaluasi Internal

k	Inertia	Silhouette	Davies–Bouldin	Calinski–Harabasz
2	183.69	0.295	1.303	16.13
3	149.22	0.227	1.381	13.70
4	122.1	0.234	1.271	13.34

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Oct 2026, IX (5): 8939 – 8945 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

k	Inertia	Silhouette	Davies–Bouldin	Calinski–Harabasz
	6			
5	99.75	0.246	1.151	13.75
6	88.12	0.234	1.025	12.92
7	77.12	0.229	1.093	12.65

Tabel 2. Hasil evaluasi internal untuk kandidat k=2–7.

Silhouette tertinggi diperoleh pada k=2 (0.295), disusul k=5 (0.246). Nilai Davies–Bouldin terendah diperoleh pada k=6 (1.025), sedangkan Calinski–Harabasz tertinggi diperoleh pada k=2 (16.13). Inertia menurun seiring bertambahnya k, sesuai karakteristik fungsi objektif K-Means.

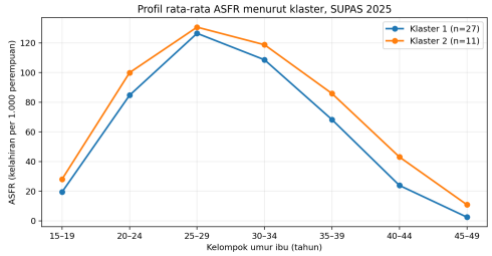
Karena silhouette dan Calinski–Harabasz mendukung k=2, serta dua kelompok lebih mudah dijelaskan sebagai pemetaan awal, k=2 digunakan sebagai konfigurasi utama. Akan tetapi, perbedaan metrik menunjukkan bahwa k=2 tidak dapat disebut sebagai jawaban mutlak; solusi k=5 atau k=6 patut diuji pada penelitian lanjutan, terutama bila tujuan praktis memerlukan segmentasi yang lebih rinci.

Profil Centroid Klaster

Tabel 3 Rata-rata ASFR

Klaster	n	15–19	20–24	25–29	30–34	35–39	40–44	45–49
Klaster 1	27	19.46	84.69	126.47	108.62	68.28	23.86	2.49
Klaster 2	11	28.10	99.94	130.64	118.76	85.87	43.05	10.86

Tabel 3. Rata-rata ASFR pada skala asli menurut klaster (kelahiran per 1.000 perempuan).



Gambar 1 Profil centroid ASFR

Gambar 1. Profil centroid ASFR untuk solusi K-Means k=2. Klaster 1 mencakup 27 provinsi dan Klaster 2 mencakup 11 provinsi. Berdasarkan centroid pada skala asli, Klaster 2 memiliki rata-rata lebih tinggi pada kelompok umur 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, dan 45–49 dibandingkan Klaster 1. Selisih rata-rata paling mencolok secara absolut tampak pada kelompok umur 40–44 dan 45–49. Pola tersebut menunjukkan bahwa Klaster 2 bukan semata kelompok dengan ASFR lebih tinggi pada usia puncak 25–29, melainkan juga mempertahankan nilai relatif lebih tinggi pada usia 40–49. Interpretasi ini bersifat deskriptif dan tidak menjelaskan faktor penyebabnya.

Anggota Klaster

Klaster 1 (27 provinsi)

Aceh, Bali, Banten, Bengkulu, Di Yogyakarta, Dki Jakarta, Gorontalo, Jambi, Jawa Barat, Jawa Tengah, Jawa Timur, Kalimantan Barat, Kalimantan Selatan, Kalimantan Tengah, Kalimantan Timur, Kalimantan Utara, Kep. Bangka Belitung, Kep. Riau, Lampung, Maluku Utara, Riau, Sulawesi Selatan, Sulawesi Tengah, Sulawesi Utara, Sumatera Barat, Sumatera Selatan, Sumatera Utara.

Klaster 2 (11 provinsi)

Maluku, Nusa Tenggara Barat, Nusa Tenggara Timur, Papua, Papua Barat, Papua Barat Daya, Papua Pegunungan, Papua Selatan, Papua Tengah, Sulawesi Barat, Sulawesi Tenggara.

Nomor klaster adalah label nominal yang dapat berubah jika inisialisasi atau data berubah. Oleh karena itu, interpretasi harus merujuk pada centroid dan bukan pada angka label. Pengelompokan ini tidak dimaksudkan sebagai peringkat provinsi, label kualitas, atau evaluasi kinerja kebijakan.

Pembahasan dalam Konteks Demografi

Hasil menunjukkan adanya variasi

profil ASFR antarprovinsi yang dapat diringkas menjadi dua kelompok utama pada solusi yang dipilih. Kelompok dengan ASFR rata-rata lebih tinggi pada berbagai kelompok umur, terutama usia 40–49, mengindikasikan bahwa perbedaan profil tidak hanya berasal dari fertilitas usia muda atau usia puncak. Namun, karena klaster dibentuk hanya menggunakan ASFR satu tahun, hasil tidak dapat digunakan untuk menyimpulkan tren waktu, perubahan perilaku reproduksi, atau penyebab sosial ekonomi.

Kekuatan pendekatan ini adalah penggunaan tujuh kelompok umur secara simultan, sehingga kemiripan provinsi ditentukan oleh profil multivariat. Standardisasi mengurangi dominasi fitur yang memiliki variasi absolut lebih besar, sedangkan pelaporan empat metrik memberi gambaran lebih lengkap dibandingkan penggunaan inertia saja. Keterbatasannya adalah jumlah observasi relatif kecil (38 provinsi), sensitifitas K-Means terhadap outlier dan inisialisasi, serta ketidakselarasan antar-metrik. Silhouette sebesar 0.295 menunjukkan pemisahan yang tidak sangat kuat; beberapa provinsi dapat berada dekat batas antarklaster.

Penelitian lanjutan dapat menguji stabilitas melalui variasi random seed atau bootstrap, membandingkan K-Means dengan hierarchical clustering dan Gaussian mixture models, serta menilai pemisahan klaster dengan analisis komponen utama sebagai visualisasi pendukung. Penambahan variabel seperti urbanisasi, pendidikan perempuan, akses kontrasepsi, kondisi ekonomi, dan indikator kesehatan reproduksi dapat membantu menguji hubungan kontekstual, tetapi variabel tersebut perlu dikumpulkan dari sumber resmi yang kompatibel dan dianalisis dengan desain yang sesuai.

Implikasi, Keterbatasan, dan Validitas

Secara praktis, hasil klaster dapat menjadi langkah awal untuk menyusun profil wilayah dalam studi demografi dan mengidentifikasi provinsi yang memiliki pola ASFR serupa untuk dianalisis lebih lanjut. Klaster tidak boleh langsung dipakai untuk menentukan intervensi tanpa mempertimbangkan ukuran populasi, ketidakpastian estimasi, konteks sosial budaya, dan indikator lain.

Keterbatasan penelitian meliputi: (1) data hanya mencakup satu periode, yaitu SUPAS 2025; (2) hanya tujuh indikator ASFR yang digunakan; (3) jumlah unit analisis terbatas pada 38 provinsi; (4) pilihan k bersifat kompromi karena metrik tidak sepenuhnya sejalan; dan (5) penelitian tidak mengevaluasi ketidakpastian estimasi ASFR pada tingkat provinsi. Karena itu, hasil sebaiknya dipahami sebagai eksplorasi struktur data, bukan klasifikasi permanen atau bukti kausal.

SIMPULAN

Penelitian ini menerapkan algoritma K-Means untuk mengelompokkan 38 provinsi Indonesia berdasarkan tujuh indikator ASFR dari hasil SUPAS 2025. Data distandardisasi dengan z-score dan diuji pada k=2 hingga k=7. Solusi k=2 menghasilkan silhouette tertinggi sebesar 0.295 dan Calinski–Harabasz tertinggi sebesar 16.13, sedangkan Davies–Bouldin terendah diperoleh pada k=6 sebesar 1.025. Dengan mempertimbangkan kesederhanaan dan interpretabilitas, k=2 dipilih sebagai solusi utama, terdiri atas 27 provinsi pada Klaster 1 dan 11 provinsi pada Klaster 2.

Profil centroid menunjukkan bahwa Klaster 2 memiliki nilai rata-rata ASFR lebih tinggi pada seluruh kelompok umur dibandingkan Klaster 1, dengan perbedaan besar pada kelompok umur 40–49 tahun. Meski demikian, nilai silhouette mengindikasikan pemisahan yang moderat dan metrik validasi tidak sepenuhnya menyepakati satu nilai k. Dengan demikian, temuan harus dipandang sebagai hasil eksploratif. Rekomendasi untuk penelitian berikutnya

adalah menguji kestabilan solusi, membandingkan algoritma alternatif, dan menambahkan variabel sosial-demografi untuk analisis yang lebih komprehensif.

DAFTAR PUSTAKA

Statistik, diperbarui 28 Agustus 2026. [Daring]. Tersedia: https://www.bps.go.id/id/statistics-table/2/MjgwMyMy/angka-kelahiran-menurut-kelompok-umur-dan-provinsi-hasil-supas.html. [Diakses: 17-Okt-2026].

J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297.

J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 28, no. 1, pp. 100–108, 1979, doi: 10.2307/2346830.

P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

D. L. Davies and D. W. Bouldin, "A cluster separation measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.

T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," Communications in Statistics, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.

A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 2010, doi: 10.1016/j.patrec.2009.09.011.

D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007, pp. 1027–1035.

P. Berkhin, "A survey of clustering data mining techniques," in Grouping Multidimensional Data: Recent Advances in Clustering, J. Kogan, C. Nicholas, and M. Teboulle, Eds. Berlin: Springer, 2006, pp. 25–71, doi: 10.1007/3-540-28349-8_2.

C. C. Aggarwal and C. K. Reddy, Eds., Data Clustering: Algorithms and Applications. Boca Raton, FL: CRC Press, 2014.

L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley, 1990.

G. J. S. Ross, "Theoretical foundations of clustering," in Clustering and Classification, P. Arabie, L. J. Hubert, and G. De Soete, Eds. Singapore: World Scientific, 1996, pp. 1–25.

Scikit- learn- developers, klearn. cluster. KMeans," scikit-learn documentation. [Daring]. Tersedia: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html. [Diakses: 17-Okt-2026].

Scikit-learn developers, "Clustering performance evaluation," scikit-learn documentation. [Daring]. Tersedia: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation. [Diakses: 17-Okt-2026].

United Nations, Department of Economic and Social Affairs, Population Division, World Population Prospects 2024: Methodology of the United Nations Population Estimates and Projections. New York: United Nations, 2024.

J. R. Weeks, Population: An Introduction to Concepts and Issues, 13th ed. Boston, MA: Cengage Learning, 2016.