Indonesian Journal of Multidisciplinary on Social and Technology
Homepage: https://journal.ilmudata.co.id/index.php/ijmst
Vol. 4 No. 3 (2026) pp: 3389-3395
P-ISSN: 2986-6790, e-ISSN: 2986-6782

Optimasi Random Search Pada K-Means Dan Agglomerative Clustering Untuk Pengelompokan Pangan Berdasarkan Profil Gizi

Hardiansyah[1], Muhammad Iqbal[2], Ifan Prihandil[3]
[1] Informatika, Ilmu Komputer, Universitas Indonesia Membangun
[2] Sistem Informasi, Ilmu Komputer, Universitas Indonesia Membangun
[3] PT. Atlasfizl Meraki Inovasi
hardiansyah@inaba.ac.id , claudluciffer@gmail.com , iprihandi@gmail.com

Abstrak
Basis data komposisi pangan memiliki dimensi gizi yang heterogen dan dapat mengandung nilai hilang, sehingga pengelompokan yang hanya mengandalkan parameter baku berisiko menghasilkan struktur klaster yang tidak stabil atau tidak substantif. Penelitian ini membandingkan K-Means dan Agglomerative Clustering dengan Ward linkage untuk mengelompokkan pangan berdasarkan profil gizi, serta mengoptimalkan konfigurasi keduanya melalui randomized hyperparameter search. Dataset awal terdiri atas 1.164 observasi pangan dan 24 atribut. Setelah penghapusan lima duplikasi identik, analisis menggunakan 1.159 observasi dan delapan atribut dengan kelengkapan memadai, yaitu air, energi, protein, lemak, karbohidrat, kalsium, fosfor, dan besi. Nilai hilang diimputasi dengan median dan data distandarisasi. Pencarian acak mengevaluasi jumlah klaster, transformasi, metode penskalaan, serta parameter spesifik algoritma dengan kendala ukuran klaster minimum 2% agar solusi tidak didominasi klaster pencilan. Evaluasi menggunakan Silhouette, Davies-Bouldin, Calinski-Harabasz, dan stabilitas bootstrap berbasis Adjusted Rand Index. K-Means terbaik menghasilkan lima klaster dengan Silhouette 0,4650, Davies-Bouldin 1,1714, Calinski-Harabasz 434,8470, dan stabilitas bootstrap rata-rata 0,9545. Agglomerative-Ward menghasilkan empat klaster dengan nilai masing-masing 0,4213, 1,4073, 376,0538, dan 0,8390. K-Means membentuk lima profil utama: tinggi air-energi rendah, karbohidrat-energi tinggi, protein tinggi, lemak-energi sangat tinggi, serta mineral-protein sangat tinggi. Hasil menunjukkan bahwa K-Means lebih kompak dan stabil, sedangkan kesepakatan antar model tetap tinggi dengan Adjusted Rand Index 0,8681. Kerangka optimasi yang menggabungkan validitas internal, batas ukuran klaster, dan stabilitas memberikan dasar pemilihan model yang lebih dapat dipertanggungjawabkan.

Kata Kunci: Agglomerative Clustering; Komposisi Pangan; K-Means; Profil Gizi; Random Search.

1. Pendahuluan

Berikut Basis data komposisi pangan merupakan infrastruktur informasi yang penting bagi analisis gizi, penyusunan rekomendasi konsumsi, pengembangan produk, dan pemetaan substitusi pangan. Akan tetapi, klasifikasi pangan konvensional lebih banyak disusun berdasarkan asal bahan atau penggunaan kuliner, sedangkan kemiripan komposisi gizinya belum tentu sejalan dengan kategori tersebut. Pendekatan berbasis data diperlukan untuk mengungkap kelompok pangan yang secara kuantitatif memiliki profil nutrien serupa. Penerapan clustering pada pangan maupun pola konsumsi telah digunakan untuk meringkas keragaman komposisi dan pola makan (O'Hara et al.,2022; Devlin et al., 2012). Penelitian terkini menunjukkan bahwa pengelompokan berbasis profil nutrien dapat melengkapi sistem kategori pangan dan meningkatkan kegunaan analisis basis data komposisi pangan (Shin & Seo, 2026).

Cluster analysis relevan untuk kebutuhan tersebut karena mampu mempartisi objek tanpa label awal dengan memaksimalkan homogenitas intra-klaster dan heterogenitas antara klaster. Dalam penelitian pangan dan pola diet, standarisasi variabel diperlukan agar nutrien dengan skala besar tidak mendominasi fungsi jarak (Zhao et al., 2021). K-Means merupakan metode partitional yang efisien, tetapi mengharuskan penentuan jumlah klaster dan sensitif terhadap inisialisasi. Sebaliknya, Agglomerative Clustering membangun hierarki penggabungan objek dan dapat merepresentasikan struktur bertingkat; Ward linkage secara khusus meminimalkan kenaikan ragam dalam klaster pada setiap tahap penggabungan (MacQueen, 1967; Ward, 1963).

Sebagian penelitian komparatif memilih algoritma dan jumlah klaster hanya berdasarkan satu indeks, terutama Silhouette. Strategi tersebut belum memadai karena indeks validitas internal dapat memberikan rekomendasi yang berbeda dan solusi dengan nilai Silhouette tinggi dapat terbentuk karena satu klaster sangat kecil yang mengisolasi pencilan. Literatur menekankan bahwa pemilihan solusi harus mempertimbangkan keterbacaan, ukuran kelompok, stabilitas, dan reproduksibilitas, bukan sekadar optimum numerik (Zhao et al., 2021;Marini Govigli et al., 2025). Dengan demikian, proses tuning perlu mengintegrasikan kendala substantif agar hasil klaster dapat digunakan untuk interpretasi profil pangan.

Random search menyediakan cara yang efisien untuk mengeksplorasi ruang hiperparameter dibandingkan pencarian grid ketika hanya sebagian parameter yang sangat menentukan kualitas model (Bergstra & Bengio,

2012; Lloyd, 1982). Namun, penggunaannya pada clustering berbeda dari klasifikasi terawasi karena tidak tersedia label target dan pembagian validasi tidak selalu memungkinkan, khususnya pada Agglomerative Clustering yang tidak mempunyai fungsi prediksi alami untuk data baru. Dalam penelitian ini, istilah randomized hyperparameter search digunakan untuk pencarian konfigurasi secara acak pada keseluruhan data dengan evaluasi indeks internal, kemudian solusi terpilih diuji kembali melalui bootstrap stability. Pendekatan ini menjaga konsistensi dengan karakter pembelajaran tanpa pengawasan.

Penelitian ini bertujuan: (1) memperoleh konfigurasi K-Means dan Agglomerative Clustering yang optimal melalui randomized hyperparameter search dengan kendala ukuran klaster; (2) membandingkan kualitas kedua model berdasarkan Silhouette, Davies-Bouldin, Calinski-Harabasz, dan stabilitas bootstrap; serta (3) menginterpretasikan karakteristik profil gizi dari klaster terbaik. Kontribusi penelitian terletak pada integrasi tiga lapis evaluasi, yaitu kualitas internal, kelayakan ukuran klaster, dan reproduksibilitas hasil, untuk mengurangi pemilihan model yang unggul secara angka tetapi lemah secara substantif

2. Metode Penelitian

2.1. Desain Penelitian dan Sumber Data

Penelitian menggunakan desain kuantitatif komputasional dengan pendekatan eksperimen data mining. Unit analisis adalah bahan pangan, sedangkan objek yang dibandingkan adalah struktur klaster yang dihasilkan K-Means dan Agglomerative Clustering. Dataset sekunder bersumber dari berkas Dataset_Gizi.csv yang diturunkan dari Tabel Komposisi Pangan Indonesia (TKPI) 2017, diterbitkan oleh Kementerian Kesehatan Republik Indonesia pada 2018 (Direktorat Jenderal Kesehatan Masyarakat, Kementerian Kesehatan Republik Indonesia, 2018). Berkas memuat kode bahan pangan, nama bahan, rujukan internal sumber komposisi, dan nilai zat gizi per 100 gram bagian yang dapat dimakan. Identifikasi sumber dilakukan berdasarkan kesesuaian sistem kode, nama bahan, struktur kelompok pangan, serta metadata publikasi TKPI; peneliti tetap perlu memverifikasi bahwa berkas yang digunakan merupakan salinan sah dan tidak mengalami perubahan substantif.

Data awal berjumlah 1.164 baris dan 24 atribut zat gizi. Audit menemukan lima baris duplikat identik pada kode dan delapan atribut analisis, sehingga data akhir berjumlah 1.159 pangan. Dataset juga memuat kategori bahan berdasarkan awalan kode, yaitu serealia, umbi, kacang-kacangan, sayuran, buah, daging dan unggas, ikan, telur, susu, lemak/minyak, gula/konfeksioneri, serta bumbu. Kategori tersebut tidak digunakan untuk membentuk klaster, tetapi dimanfaatkan sebagai deskriptor eksternal dalam interpretasi.

2.2. Seleksi Atribut dan Pra-pemrosesan

Pemilihan atribut didasarkan pada kelengkapan data dan relevansi representasi makronutrien-mineral. Delapan atribut dipertahankan karena persentase nilai hilangnya tidak melebihi sekitar 6%, yaitu air, energi, protein, lemak, karbohidrat, kalsium, fosfor, dan besi. Enam belas atribut lain tidak digunakan dalam model utama karena sebagian memiliki kehilangan data lebih dari 60%; memasukkannya akan membuat struktur klaster lebih banyak ditentukan oleh imputasi daripada observasi empiris. Keputusan ini merupakan pengendalian bias pengukuran, bukan klaim bahwa mikronutrien tersebut tidak penting.

Tabel 1. Atribut yang digunakan dalam pemodelan

Atribut	Satuan	Nilai hilang (%)	Perlakuan
AIR	g	0.3	Imputasi median
ENERGI	kkal	0.3	Imputasi median
PROTEIN	g	0.5	Imputasi median
LEMAK	g	0.7	Imputasi median
KH	g	0.3	Imputasi median
KALSIUM	mg	4.6	Imputasi median
FOSFOR	mg	4.5	Imputasi median
BESI	mg	5.2	Imputasi median

Nilai simbolik seperti tanda hubung, karakter kosong, dan simbol non numerik dikonversi menjadi nilai hilang. Angka yang berada di luar batas fisik yang layak per 100 gram dikonversi menjadi nilai hilang. Setelah itu, setiap atribut diimputasi menggunakan median. Median dipilih karena distribusi beberapa mineral bersifat menceng ke kanan dan mengandung pangan dengan konsentrasi ekstrem. Seluruh atribut kemudian distandarisasi menjadi z-score agar jarak Euclidean tidak didominasi oleh energi, kalsium, atau fosfor yang berskala lebih besar. Implementasi komputasional menggunakan Python dan pustaka scikit-learn(Pedregosa et al., 2011).

DOI: https://doi.org/10.69693/ijmst.v4i2.13024

2.3. Algoritma Clustering

K-Means meminimalkan within-cluster sum of squares dengan mengalokasikan setiap pangan ke centroid terdekat (Lloyd, 1982). Model akhir menggunakan inisialisasi k-means++, 30 pengulangan inisialisasi, maksimum 300 iterasi, toleransi 10^{-4}, dan algoritma Lloyd. Agglomerative Clustering dimulai dari klaster singleton, kemudian menggabungkan pasangan klaster secara bertahap. Ward linkage dipilih pada model terbaik karena meminimalkan kenaikan jumlah kuadrat galat akibat penggabungan (Ward, 1963, Murtagh & Legendre, 2014). Kedua algoritma menggunakan ruang fitur yang sama untuk menjamin perbandingan yang setara

2.4. Randomized Hyperparameter Search

Pencarian konfigurasi dilakukan secara acak menggunakan ParameterSampler. Karena model bersifat tanpa pengawasan, istilah random search dalam penelitian ini tidak merujuk pada RandomizedSearchCV dengan label target, melainkan pengambilan sampel konfigurasi dari ruang hiperparameter dan evaluasi menggunakan indeks validitas internal. Sebanyak 80 konfigurasi K-Means dan 60 konfigurasi Agglomerative diuji. Solusi dianggap layak apabila setiap klaster memuat sekurang-kurangnya 2% dari seluruh observasi atau sekitar 23 pangan. Kendala tersebut menghindari solusi degeneratif yang memperoleh Silhouette tinggi hanya dengan memisahkan satu atau dua pencilan.

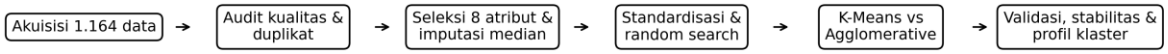
Tabel 2. Ruang hiperparameter randomized search

Algoritma	**Hiperparameter**	**Ruang pencarian**
K-Means	Jumlah klaster	2-10
K-Means	n_init	10, 20, 30, 50
K-Means	max_iter	300, 500, 1.000
K-Means	tol	10^{-3}, 10^{-4}, 10^{-5}
K-Means	algoritma	Lloyd, Elkan
Agglomerative	Jumlah klaster	2-10
Agglomerative	linkage	Ward, complete, average, single
Keduanya	penskalaan	StandardScaler, RobustScaler
Keduanya	transformasi	tanpa transformasi, log1p

2.5. Evaluasi Model

Silhouette mengukur kedekatan objek dengan klasternya dibandingkan dengan klaster terdekat; nilai yang lebih tinggi menunjukkan kohesi dan separasi yang lebih baik (Rousseeuw, 1987). Davies-Bouldin membandingkan dispersi intra-klaster dengan jarak antar klaster, sehingga nilai lebih rendah lebih baik (Davies & Bouldin, 1979). Calinski-Harabasz membandingkan ragam antar klaster dengan ragam intra-klaster; nilai lebih tinggi menunjukkan struktur yang lebih tegas (Caliński & Harabasz, 1974). Silhouette dijadikan kriteria utama, sedangkan Davies-Bouldin dan Calinski-Harabasz digunakan sebagai verifikasi silang sesuai prinsip evaluasi multi-indeks pada validasi clustering(Halkidi et al., 2001;Jain, 2010).

Stabilitas diuji melalui 50 pengambilan sub sampel tanpa pengembalian sebesar 80% dari data. Model dibangun ulang pada setiap subsampel dan hasilnya dibandingkan dengan label model penuh menggunakan Adjusted Rand Index (ARI). Nilai ARI mendekati satu menunjukkan reprodusibilitas tinggi. Kesepakatan struktur antara K-Means dan Agglomerative juga dihitung dengan ARI. PCA dua komponen hanya digunakan untuk visualisasi, bukan sebagai masukan utama model(Jolliffe & Cadima, 2016).



Gambar 1. Tahapan Penelitian

3. Hasil dan Pembahasan

3.1. Kualitas Data dan Implikasi Pra-pemrosesan

Audit kualitas menunjukkan bahwa delapan atribut terpilih memiliki kehilangan data rendah, berkisar 0,3%-5,2%. Sebaliknya, atribut serat, natrium, kalium, tembaga, seng, retinol, beta-karoten, karoten total, dan beberapa vitamin memiliki kehilangan data yang jauh lebih besar. Penggunaan seluruh atribut memang akan memperluas cakupan

DOI: https://doi.org/10.69693/ijmst.v4i2.13024

profil gizi, tetapi menghasilkan ketergantungan yang berlebihan pada imputasi. Oleh sebab itu, model utama dibatasi pada delapan atribut. Pilihan ini meningkatkan validitas internal, tetapi konsekuensinya klaster belum dapat ditafsirkan sebagai klasifikasi mikronutrien lengkap.

Standardisasi terbukti diperlukan karena rentang kalsium dan fosfor jauh lebih besar daripada protein, lemak, dan besi. Tanpa standardisasi, fungsi jarak akan memberikan bobot implisit lebih besar pada mineral. Random search juga menunjukkan bahwa konfigurasi terbaik kedua algoritma menggunakan StandardScaler tanpa transformasi log. Temuan ini mengindikasikan bahwa setelah penanganan nilai hilang dan pembatasan nilai tidak layak, variasi asli data masih informatif untuk memisahkan profil pangan.

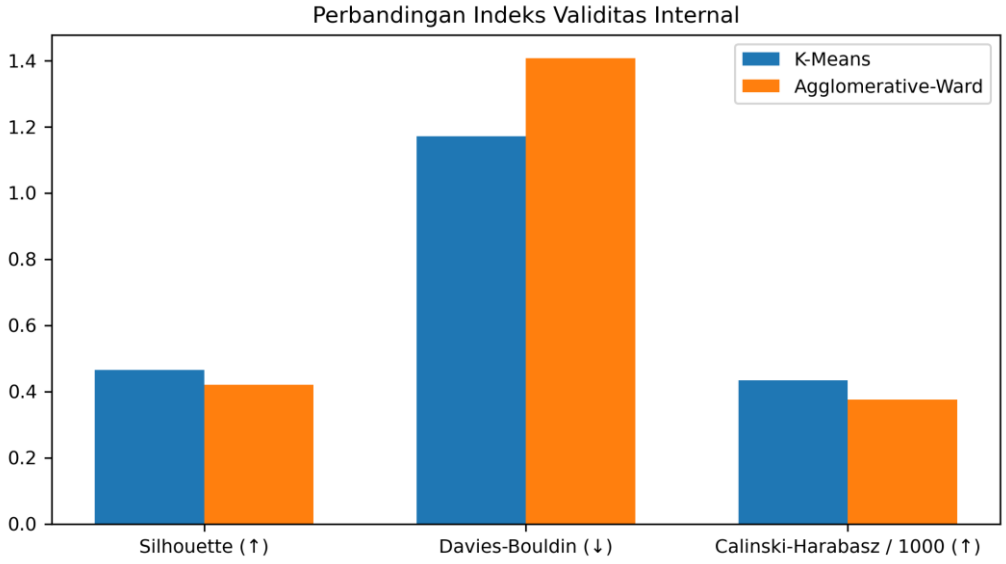
3.2. Hasil Randomized Search dan Perbandingan Algoritma

Konfigurasi terbaik K-Means menghasilkan lima klaster dengan Silhouette 0.4650, Davies-Bouldin 1.1714, dan Calinski-Harabasz 434.8470. Agglomerative-Ward menghasilkan empat klaster dengan Silhouette 0.4213, Davies-Bouldin 1.4073, dan Calinski-Harabasz 376.0538. Ketiga indeks konsisten mengunggulkan K-Means: Silhouette dan Calinski-Harabasz lebih tinggi, sedangkan Davies-Bouldin lebih rendah. Keunggulan tersebut bukan hanya akibat jumlah klaster yang berbeda karena seluruh konfigurasi k=2-10 telah dipertimbangkan dalam ruang pencarian.

Tabel 3. Kinerja konfigurasi terbaik

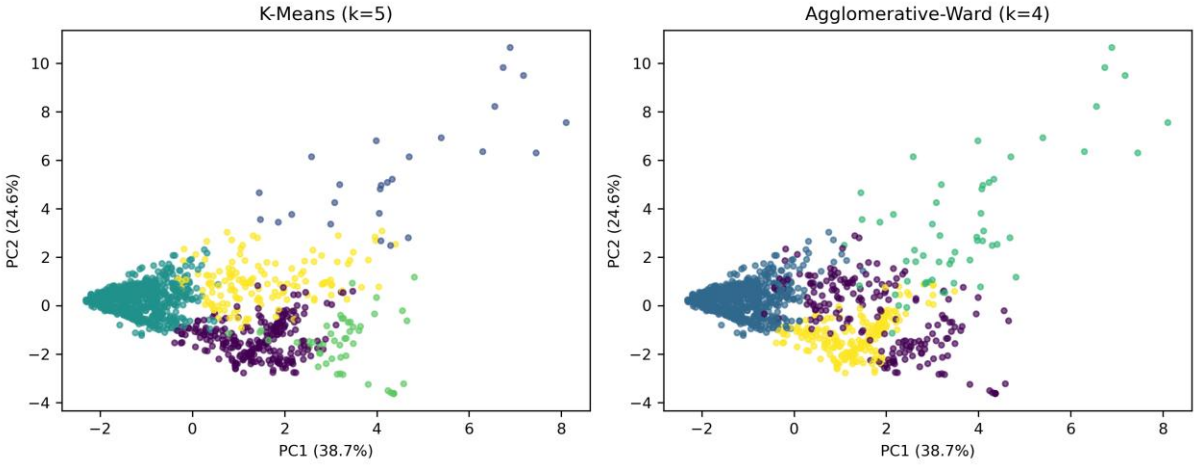
Algoritma	k	Silhouette ↑	DBI ↓	CH ↑	Min. n	Maks. n	Stabilitas ARI
K-Means	5	0.4650	1.1714	434.8470	26	695	0.9545
Agglomerative-Ward	4	0.4213	1.4073	376.0538	57	701	0.8390

K-Means juga lebih stabil. Rata-rata ARI bootstrap mencapai 0.9545 dengan simpangan baku 0.0343, sedangkan Agglomerative mencapai 0.8390 dengan simpangan baku 0.0876. Kesepakatan label kedua model tetap tinggi dengan ARI 0.8681, yang berarti struktur utama data relatif konsisten, tetapi K-Means memecah kelompok pangan padat protein-lemak-mineral menjadi profil yang lebih spesifik. Hasil ini sejalan dengan prinsip bahwa model klaster sebaiknya dipilih berdasarkan stabilitas dan reproduksibilitas, bukan satu indeks internal saja(Zhao et al., 2021;Hennig, 2007).



Gambar 2. Perbandingan indeks validitas internal

DOI: https://doi.org/10.69693/ijmst.v4i2.13024



Gambar 3. Visualisasi PCA hasil K-Means dan Agglomerative

Visualisasi PCA menunjukkan satu kelompok besar pangan berkadar air tinggi dan energi rendah, kelompok pangan kaya karbohidrat, serta beberapa kelompok lebih kecil yang padat protein, lemak, atau mineral. Tumpang tindih pada bidang dua dimensi tidak secara langsung menandakan kegagalan klaster karena PCA hanya merangkum sebagian ragam delapan atribut. Penilaian utama tetap menggunakan jarak pada ruang delapan dimensi yang telah distandardisasi.

3.3. Profil Klaster K-Means

Tabel 4. Rata –rata profil gizi klaster K-Means per 100gram

Klaster	n	Air	Energi	Protein	Lemak	KH	Ca	P	Fe
K1	695	77.29	93.65	6.09	2.37	12.93	92.61	99.18	2.12
K2	249	17.69	349.28	7.60	7.09	65.14	112.80	168.35	3.81
K3	137	41.87	294.69	32.09	14.86	12.02	219.77	305.37	7.07
K4	52	9.35	626.94	13.11	57.65	17.92	105.51	219.20	2.77
K5	26	35.81	278.15	42.44	7.43	16.11	1760.85	1193.77	15.33

K1 - Tinggi air, energi rendah. Kategori dominan: Sayur-sayuran dan hasil olahannya (29.1%); Ikan dan hasil olahannya (17.0%); Buah-buahan dan hasil olahannya (15.8%).
K2 - Karbohidrat dan energi tinggi. Kategori dominan: Serealia dan hasil olahannya (39.4%); Kacang-kacangan dan hasil olahannya (19.7%); Umbi berpati dan hasil olahannya (18.1%).
K3 - Protein tinggi. Kategori dominan: Daging, unggas, dan hasil olahannya (40.9%); Ikan dan hasil olahannya (27.0%); Kacang-kacangan dan hasil olahannya (17.5%).
K4 - Lemak dan energi sangat tinggi. Kategori dominan: Kacang-kacangan dan hasil olahannya (51.9%); Lemak dan minyak (25.0%); Daging, unggas, dan hasil olahannya (11.5%).
K5 - Mineral dan protein sangat tinggi. Kategori dominan: Ikan dan hasil olahannya (73.1%); Bumbu-bumbu (15.4%); Serealia dan hasil olahannya (3.8%).

Klaster K1 merupakan kelompok terbesar dengan 695 pangan. Rata-rata kadar airnya 77,23 g dan energinya 93,94 kkal, sehingga menggambarkan pangan tinggi air dan berenergi rendah. Sayuran, ikan, dan buah merupakan kategori dominan. Meskipun ikan biasanya dipersepsikan sebagai pangan protein, sebagian ikan segar memiliki kadar air tinggi sehingga secara multivariat lebih dekat dengan kelompok ini daripada pangan protein kering.

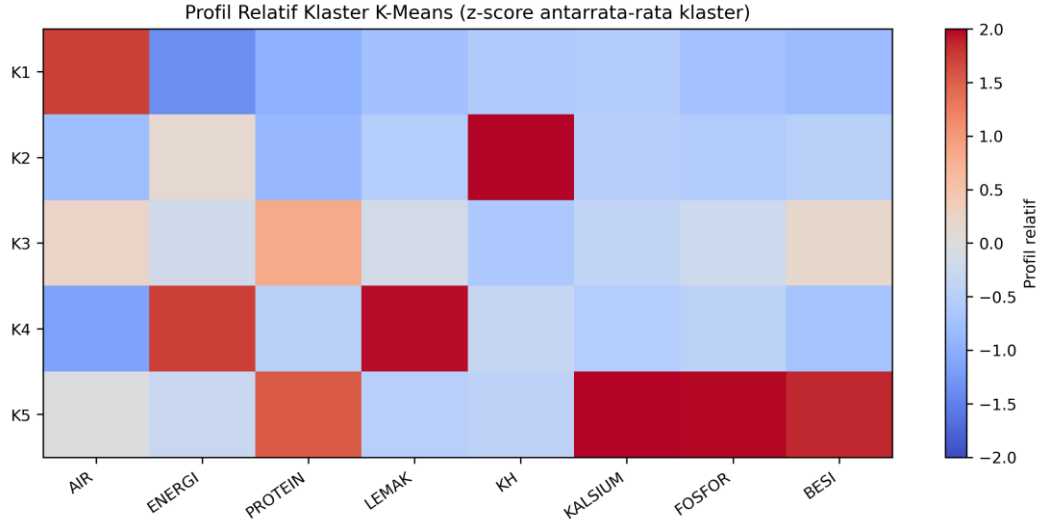
Klaster K2 terdiri atas 249 pangan dengan karbohidrat rata-rata 65,14 g dan energi 349,28 kkal. Komposisinya didominasi serealia, kacang-kacangan, dan umbi. Profil tersebut menunjukkan fungsi kelompok sebagai sumber energi-karbohidrat. Kehadiran kacang-kacangan dalam klaster ini menegaskan bahwa penempatan klaster ditentukan oleh gabungan delapan atribut, bukan satu zat gizi tunggal.

Klaster K3 memuat 137 pangan dengan protein rata-rata 32,09 g, fosfor 300,88 mg, dan lemak 14,86 g. Daging, ikan, dan kacang-kacangan menjadi kategori utama. Klaster ini merepresentasikan pangan kaya protein dengan kepadatan mineral sedang. Klaster K4 berjumlah 52 pangan dan mempunyai energi 626,94 kkal serta lemak 57,65 g; kacang-kacangan serta lemak/minyak mendominasi kelompok tersebut. Dengan demikian, K4 merupakan kelompok paling padat energi.

Klaster K5 merupakan kelompok terkecil yang masih memenuhi batas kelayakan, yaitu 26 pangan atau 2,24% data. Kelompok ini memiliki protein 42,44 g, kalsium 1.760,85 mg, fosfor 1.193,77 mg, dan besi 15,33 mg. Sebanyak 73,1% anggotanya berasal dari kategori ikan dan hasil olahan, terutama produk kering atau

DOI: https://doi.org/10.69693/ijmst.v4i2.13024

terkonsentrasi. Profil tersebut berbeda secara substantif dari K3 karena pembeda utamanya bukan hanya protein, melainkan kepadatan mineral yang sangat tinggi.



Gambar 4. Profil relatif lima klaster K-Means

3.4. Pembahasan Metodologis dan Implikasi

Keunggulan K-Means pada data ini dapat dijelaskan oleh kecenderungan profil pangan membentuk kelompok kompak di sekitar centroid setelah standardisasi. Ward linkage memiliki tujuan yang berdekatan dengan minimisasi ragam, sehingga kesepakatan antara model tinggi; namun proses penggabungan yang bersifat greedy tidak dapat membatalkan merger awal ketika struktur lokal kurang tepat. K-Means melakukan pembaruan centroid secara iteratif sehingga lebih fleksibel memperbaiki batas antarkelompok.

Kendala ukuran klaster minimum merupakan bagian penting dari prosedur tuning. Ketika kendala tersebut dihilangkan, pencarian dapat memilih konfigurasi dengan Silhouette sangat tinggi tetapi menghasilkan klaster berisi satu atau dua objek ekstrem. Solusi demikian tidak layak dijadikan dasar tipologi pangan karena lebih menyerupai deteksi pencilan. Oleh karena itu, optimum statistik harus dibaca bersama distribusi ukuran klaster dan makna profil. Prinsip kompromi antara indeks validitas dan kegunaan interpretatif juga ditekankan dalam studi segmentasi pangan mutakhir (Marini Govigli et al., 2025).

Secara praktis, hasil klaster dapat digunakan sebagai lapisan tambahan pada basis data pangan untuk pencarian substitusi berbasis kedekatan profil, penyederhanaan eksplorasi data, atau penyusunan rekomendasi awal. Akan tetapi, label klaster tidak boleh diterjemahkan langsung menjadi kategori 'sehat' atau 'tidak sehat'. Penilaian kesehatan pangan memerlukan ukuran porsi, tingkat pengolahan, natrium, gula, serat, vitamin, kebutuhan individu, dan konteks pola makan. Hasil penelitian ini hanya menggambarkan kemiripan komposisi delapan zat gizi per 100 gram.

Keterbatasan penelitian meliputi tiga aspek. Pertama, sumber bibliografis utama dataset harus dipastikan agar keterlacakan data memenuhi standar publikasi. Kedua, sejumlah mikronutrien tidak dimasukkan karena kehilangan data tinggi; akibatnya profil klaster belum mencakup seluruh dimensi kualitas gizi. Ketiga, validasi menggunakan indeks internal dan stabilitas komputasional, belum menggunakan penilaian ahli gizi sebagai validasi eksternal. Penelitian lanjutan dapat menerapkan multiple imputation, model-based clustering, k-medoids yang lebih robust terhadap pencilan, dan validasi oleh pakar domain

4. Kesimpulan

Kesimpulan Randomized hyperparameter search dengan kendala ukuran klaster minimum berhasil menghasilkan perbandingan K-Means dan Agglomerative Clustering yang lebih dapat dipertanggungjawabkan pada 1.159 data pangan dan delapan atribut gizi. K-Means dengan lima klaster merupakan model terbaik karena memperoleh Silhouette 0,4650, Davies-Bouldin 1,1714, Calinski-Harabasz 434,8470, dan stabilitas bootstrap ARI 0,9545. Agglomerative-Ward dengan empat klaster menghasilkan nilai masing-masing 0,4213, 1,4073, 376,0538, dan 0,8390. Kesepakatan antar model tetap tinggi dengan ARI 0,8681, sehingga keduanya menangkap struktur utama yang serupa, tetapi K-Means memberikan pemisahan yang lebih rinci dan stabil.

Lima profil K-Means terdiri atas pangan tinggi air-energi rendah, karbohidrat-energi tinggi, protein tinggi, lemak-energi sangat tinggi, serta mineral-protein sangat tinggi. Secara metodologis, penelitian menegaskan bahwa pemilihan clustering tidak boleh hanya berdasarkan maksimum Silhouette; distribusi ukuran klaster dan stabilitas resampling perlu dimasukkan agar solusi pencilan tidak keliru dipilih sebagai model terbaik. Hasil dapat menjadi dasar pengembangan fitur pencarian dan substitusi pangan berbasis komposisi, tetapi belum dapat digunakan

DOI: https://doi.org/10.69693/ijmst.v4i2.13024

sebagai label kesehatan tanpa penambahan atribut gizi dan validasi ahli..

Reference

H. Shin and H. Seo, "Nutrient profile-based food categorization and group-wise missing data imputation for commercial food composition database," Journal of Food Composition and Analysis, vol. 150, art. 108828, 2026, doi: 10.1016/j.jfca.2025.108828.

J. Zhao et al., "A review of statistical methods for dietary pattern analysis," Nutrition Journal, vol. 20, art. 37, 2021, doi: 10.1186/s12937-021-00692-7.

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

J. H. Ward Jr., "Hierarchical grouping to optimize an objective function," Journal of the American Statistical Association, vol. 58, no. 301, pp. 236-244, 1963, doi: 10.1080/01621459.1963.10500845.

V. Marini Govigli et al., "From plate to policy: segmenting consumers' food behaviours to tailor nutritional recommendations in African cities," Agricultural and Food Economics, vol. 13, art. 8, 2025, doi: 10.1186/s40100-025-00349-7.

J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," Journal of Machine Learning Research, vol. 13, pp. 281-305, 2012.

S. Lloyd, "Least squares quantization in PCM," IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129-137, 1982, doi: 10.1109/TIT.1982.1056489.

F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion?" Journal of Classification, vol. 31, pp. 274-295, 2014, doi: 10.1007/s00357-014-9161-z.

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.

C. Hennig, "Cluster-wise assessment of cluster stability," Computational Statistics & Data Analysis, vol. 52, no. 1, pp. 258-271, 2007, doi: 10.1016/j.csda.2006.11.025.

F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," Philosophical Transactions of the Royal Society A, vol. 374, art. 20150202, 2016, doi: 10.1098/rsta.2015.0202.

M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," Journal of Intelligent Information Systems, vol. 17, pp. 107-145, 2001, doi: 10.1023/A:1012801612483.

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.

B. Bischl et al., "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," WIREs Data Mining and Knowledge Discovery, vol. 13, no. 2, e1484, 2023, doi: 10.1002/widm.1484.

C. O'Hara, A. O'Sullivan, and E. R. Gibney, "A clustering approach to meal-based analysis of dietary intakes applied to population and individual data," The Journal of Nutrition, vol. 152, no. 10, pp. 2297–2308, 2022, doi: 10.1093/jn/nxac151.

U. M. Devlin, B. A. McNulty, A. P. Nugent, and M. J. Gibney, "The use of cluster analysis to derive dietary patterns: methodological considerations, reproducibility, validity and the effect of energy mis-reporting," Proceedings of the Nutrition Society, vol. 71, no. 4, pp. 599-609, 2012, doi: 10.1017/S0029665112000729.

Direktorat Jenderal Kesehatan Masyarakat, Kementerian Kesehatan Republik Indonesia, Tabel Komposisi Pangan Indonesia 2017. Jakarta: Kementerian Kesehatan RI, 2018, ISBN 978-602-416-407-2.

DOI: https://doi.org/10.69693/ijmst.v4i2.13024