Integration of K-Means Clustering and K-Nearest Neighbor for Mapping and Predicting Stunting Risk Among Toddlers in Indonesia

by Melinda Novianti

Submission date: 30-Oct-2026 03:23PM (UTC+0700)
Submission ID: 3075044887
File name: 42_Artikel_JTIK_V12N2_Melinda_3752.pdf (338.14K)
Word count: 5382
Character count: 32341

Jurnal Teknlogi Informatika dan Komputer MH. Thamrin p-ISSN 2656-9957; e-ISSN 2622-8475
Volume 12 No 2; September 2026 Hal 1721-1736

Integration of K-Means Clustering and K-Nearest Neighbor for Mapping and Predicting Stunting Risk Among Toddlers in Indonesia

Melinda Novianti[1)*)], Naeli Umniati[2)]

[1)2)]Program Studi Magister Manajemen Sistem Informasi, Program Magister Teknologi dan Rekayasa, Universitas Gunadarma
[*)]**Correspondence author**: melindanovianti37@gmail.com, Jakarta, Indonesia
DOI: https://doi.org/10.37012/jtik.v12i2.3752

Abstract

Stunting prevalence among children under five in Indonesia keeps falling, yet remains unevenly spread, so setting priority areas needs a verifiable quantitative basis. Previous studies generally use a single algorithm, so the resulting map yields no estimate for the next period. This study combines both in one workflow: K-Means Clustering groups 514 districts and cities by risk level, and the cluster labels become reference classes for K-Nearest Neighbor to predict the 2025 risk category. The data are under-five stunting prevalence from the Indonesian Nutritional Status Survey (SSGI) 2021–2024. Its 128 empty values were filled by mean computation and rescaled with Min-Max normalisation, distance used Euclidean Distance, cluster quality the Silhouette Coefficient, and classification 23 nearest neighbours from a 1,542-row training set. Iteration stopped at the fifteenth round, giving 60 districts in High Risk, 232 in Moderate, and 222 in Low Risk, with a mean Silhouette Score of 0.33. The 2025 prediction places 25 districts in High Risk, 194 in Moderate, and 295 in Low Risk, agreeing with the cluster labels on 378 of 514 districts, or 73.54%. Re-running it in Orange Data Mining differed by at most four districts, confirming the workflow and its map for setting intervention priorities.

Keywords: *Integration of Clustering And Classification, K-Means Clustering, K-Nearest Neighbor, Under-Five Stunting Risk Mapping, Indonesian Nutritional Status Survey*

Abstrak

Prevalensi *stunting* balita di Indonesia terus menurun, namun sebarannya belum merata sehingga penetapan daerah prioritas memerlukan dasar kuantitatif yang dapat diperiksa ulang. Penelitian terdahulu umumnya memakai satu algoritma sehingga peta yang dihasilkan tidak berlanjut menjadi dugaan bagi periode sesudahnya. Penelitian ini merangkai dua algoritma dalam satu alur kerja: *K-Means Clustering* memilah 514 kabupaten/kota menurut derajat risiko, lalu label kelompok tersebut dipakai sebagai kelas rujukan bagi *K-Nearest Neighbor* untuk menduga kategori risiko tahun 2025. Bahan analisisnya prevalensi *stunting* balita dari Survei Status Gizi Indonesia (SSGI) 2021–2024. Sebanyak 128 nilai kosong diisi dengan komputasi *mean* lalu dinormalisasi dengan *Min-Max*, jarak memakai *Euclidean Distance*, mutu kelompok ditakar melalui *Silhouette Coefficient*, dan penggolongan memakai 23 tetangga terdekat dari 1.542 baris data latih. Iterasi terhenti di putaran kelima belas dan menyisakan 60 wilayah pada Risiko Tinggi, 232 pada Risiko Sedang, serta 222 pada Risiko Rendah, dengan rerata *Silhouette Score* 0,33. Dugaan untuk 2025 menempatkan 25 wilayah pada Risiko Tinggi, 194 pada Risiko Sedang, dan 295 pada Risiko Rendah, sejalan dengan label kelompok pada 378 dari 514 wilayah atau 73,54%. Pengerjaan ulang pada Orange Data Mining berbeda paling jauh empat wilayah, sehingga rangkaian yang diusulkan terkonfirmasi dan petanya layak dipakai menyusun urutan prioritas.

Kata Kunci: Integrasi Clustering dan Klasifikasi, K-Means Clustering, K-Nearest Neighbor, Pemetaan Risiko Stunting Balita, Survei Status Gizi Indonesia

PENDAHULUAN

Stunting adalah gangguan pertumbuhan pada balita akibat kebutuhan gizi yang tidak tercukupi dalam waktu lama, ditandai tinggi badan yang jauh lebih pendek daripada anak lain yang seusia menurut standar pertumbuhan yang dipakai secara internasional (Kementerian Kesehatan Republik Indonesia, 2020). Dampaknya tidak berhenti pada tubuh yang lebih pendek, tetapi juga memperlambat kemampuan berpikir, menyulitkan anak mengikuti pelajaran di sekolah, dan memperkecil peluang memperoleh penghasilan yang layak setelah dewasa (Yohanis Ndapa Deda et al., 2024). Di Indonesia prevalensinya terus turun, berturut-turut 24,4% (2021), 21,6% (2022), 21,5% (2023), dan 19,8% (2024) (Kementerian Perencanaan Pembangunan Nasional, 2025), namun masih di atas sasaran Rencana Pembangunan Jangka Menengah Nasional 2025–2029, yaitu 18,8% untuk 2025 dan 14,2% untuk 2029, sementara waktu dan anggaran yang tersedia terbatas.

Angka nasional tersebut menyimpan perbedaan yang lebar antardaerah. Pada 2021 prevalensi *stunting* di 514 kabupaten/kota berkisar antara 5,1% dan 55,4%, dan rentang selebar itu masih terlihat sampai 2024. Karena itu dibutuhkan cara pengelompokan wilayah yang dasar perhitungannya jelas serta dapat diperiksa ulang, sejalan dengan Peraturan Presiden Republik Indonesia No 72 Tahun 2021 tentang Percepatan Penurunan *Stunting* yang meminta penanganan dipusatkan pada wilayah prioritas. Upaya ke arah itu sudah dirintis (Wulandari et al., 2025) yang menduga prevalensi tingkat kecamatan di Banten dengan model Bayes spasial, tetapi pendekatan tersebut menuntut data pendukung yang rinci, sedangkan pengelompokan berdasarkan kemiripan pola prevalensi dapat dijalankan dengan data yang sudah tersedia untuk seluruh kabupaten/kota.

Pendekatan *data mining* menjawab kebutuhan tersebut, khususnya *clustering* yang mengelompokkan objek berdasarkan kemiripan nilainya, dan *K-Means* paling sering dipakai untuk data tingkat wilayah karena perhitungannya ringan serta hasilnya mudah dibaca. Penelitian (Chrisamudra & Abdurrahman, 2025) memakainya untuk memetakan beban kasus pneumonia balita pada wilayah padat penduduk, sedangkan Khusaeri (2024) menguji dua sampai sepuluh kelompok pada sebaran demam berdarah dengue. Pada kasus *stunting*, (Dayla May Cytry et al., 2023) mengelompokkan 71 kelurahan/desa dan mendapati hanya 3 wilayah (4,23%) berisiko tinggi, 36 wilayah (50,70%) berisiko sedang, dan 32 wilayah (45,07%) berisiko rendah, sehingga masalahnya menumpuk pada sebagian kecil wilayah.

Gambaran sejalan diperoleh (Ramadhani, 2025) di Sumatera Utara, sementara (Kadafi & Finandhita, 2024) dan (Putra et al., 2025) memanfaatkannya untuk menentukan materi penyuluhan kesehatan dan menata perbaikan pelayanan pasien rumah sakit.

Pada sisi klasifikasi, (Gina Purnama Insany et al., 2023) menemukan *K-Nearest Neighbor* dan jaringan saraf tiruan sama-sama bekerja baik untuk menggolongkan status gizi balita, dengan akurasi meningkat setelah skala data diseragamkan. Telaah (Syriopoulos et al., 2025) menyimpulkan *K-Nearest Neighbor* tetap banyak dipakai karena cara kerjanya sederhana, tidak menuntut asumsi bentuk sebaran, dan tidak mudah terganggu data yang menyimpang, sedangkan (Zhang, 2022) mencatat penentuan jumlah tetangga belum memiliki jawaban tunggal. Mutu pengelompokan umumnya dinilai memakai *Silhouette Coefficient* karena menimbang kerapatan anggota di dalam kelompok sekaligus jaraknya terhadap kelompok lain (Hasan, 2024; Rousseeuw, 1987). (Schubert, 2023) mengingatkan bahwa penentuan jumlah kelompok tidak cukup bersandar pada *Elbow Method*, dan (Hartono & Lusiana, 2026) menegaskan *Elbow* serta *Silhouette* lebih baik dibaca bersama-sama.

Dari kajian tersebut tampak tiga celah. Pertama, analisis yang ada umumnya berhenti pada satu kabupaten atau satu provinsi sehingga peta risiko *stunting* untuk seluruh Indonesia belum tersedia. Kedua, analisis berhenti begitu kelompok terbentuk dan tidak dilanjutkan menjadi dugaan bagi tahun berikutnya, padahal label *K-Means* dapat dipakai sebagai kelas acuan bagi algoritma klasifikasi sebagaimana dibuktikan (Hartanto et al., 2025) pada bidang lain, dan label seperti itu justru dibutuhkan karena SSGI tidak menyertakan kategori risiko. Ketiga, hampir semua penelitian mengandalkan satu program saja sehingga pembaca sulit menelusuri bagaimana angka akhir diperoleh, padahal pada penelitian yang menunjang kebijakan publik kemampuan menelusuri perhitungan merupakan syarat. Karena itu penelitian ini bertujuan menggabungkan *K-Means Clustering* dan *K-Nearest Neighbor* dalam satu rangkaian kerja untuk memetakan risiko *stunting* balita di 514 kabupaten/kota Indonesia dengan data SSGI 2021 sampai 2024, menilai mutu pengelompokan dengan *Silhouette Coefficient*, serta memprediksi kategori risiko tahun 2025, dengan setiap tahap dikerjakan dua kali pada Microsoft Excel dan Orange Data Mining agar hasil keduanya saling memeriksa.

METODE PENELITIAN

Penelitian ini merupakan penelitian kuantitatif deskriptif yang menggunakan data sekunder. Objek yang diteliti adalah seluruh 514 kabupaten/kota di Indonesia, sehingga tidak ada penarikan sampel dan semua wilayah ikut dianalisis. Pengamatan mencakup empat tahun, yaitu 2021 sampai 2024, sehingga datanya bersifat longitudinal dan memungkinkan pola perubahan antartahun untuk dibaca.

Data bersumber dari laporan SSGI yang diterbitkan Kementerian Kesehatan Republik Indonesia untuk keempat tahun tersebut (Kementerian Kesehatan Republik Indonesia, 2021, 2022, 2023, 2024). Dari setiap laporan diambil nama provinsi, nama kabupaten/kota, dan nilai prevalensi *stunting* balita. Dua atribut pertama dipakai hanya untuk menandai wilayah, jadi tidak masuk perhitungan, sedangkan empat nilai prevalensi menjadi variabel numerik yang diolah. Rentang nilai keempat atribut sebelum diseragamkan skalanya disajikan pada Tabel 1.

Tabel 1.

Rentang Nilai Atribut Prevalensi *Stunting* Sebelum Penskalaan (%)

Atribut	Nilai Terendah	Nilai Tertinggi	Lebar Rentang
Tahun_2021	5,10	55,40	50,30
Tahun_2022	4,80	54,50	49,70
Tahun_2023	1,60	52,15	50,55
Tahun_2024	5,20	56,80	51,60

Sumber: hasil pengolahan data SSGI 2021–2024

Tabel 1 menunjukkan bahwa lebar rentang setiap tahun tidak sama, paling lebar pada 2024 dan paling sempit pada 2022. Perbedaan ini perlu dihilangkan sebelum jarak dihitung karena ukuran jarak akan cenderung mengikuti atribut yang rentang nilainya paling lebar.

Praproses Data

Praproses dijalankan dalam empat tahap. Seleksi atribut menetapkan enam kolom yang dibawa ke tahap analisis, yaitu Provinsi, Kab/Kota, Prevalensi Stunting 2021, Prevalensi Stunting 2022, Prevalensi Stunting 2023 dan Prevalensi Stunting 2024. Pada saat proses pemeriksaan data, ditemukan 128 data kosong yang tersebar pada beberapa tahun pengamatan, sementara pengecekan terhadap 514 baris tidak menemukan wilayah yang

tercatat dua kali. Sel yang kosong tersebut diisi menggunakan nilai rata-rata sehingga tidak ada wilayah yang harus dikeluarkan dan jumlah objek tetap utuh. Tahap terakhir menyeragamkan skala keempat atribut ke rentang nol sampai satu menggunakan *Min-Max Normalization* sesuai Persamaan 1 (Han et al., 2012).

$$x' = (x - x_{min}) / (x_{max} - x_{min}) \quad (1)$$

Notasi x adalah nilai asal sebuah atribut, x_{min} dan x_{max} adalah nilai terendah dan tertinggi atribut tersebut pada seluruh wilayah, sedangkan x′ adalah nilai hasil penyeragaman skala. Sebagai contoh, prevalensi Kabupaten Simeulue tahun 2021 sebesar 25,9 berubah menjadi 0,41 setelah diseragamkan terhadap batas 5,1 dan 55,4.

Pengelompokan dan Penilaian Mutu Cluster

Pengelompokan menggunakan *K-Means*, algoritma yang pertama kali dirumuskan Macqueen (1967) dan bekerja dengan menarik setiap objek ke titik pusat terdekat lalu memperbarui titik pusat itu secara berulang. Jumlah kelompok ditetapkan tiga, mengikuti kebutuhan pengambil kebijakan yang membedakan daerah menjadi berisiko tinggi, sedang, dan rendah. Nama kategori tidak melekat pada nomor *cluster*, melainkan diberikan setelah iterasi berhenti dengan membandingkan besar *centroid* akhir masing-masing kelompok. *Centroid* awal ditetapkan berbeda untuk setiap kelompok: kelompok Risiko Tinggi berangkat dari nilai penyeragaman skala tertinggi pada setiap tahun, sedangkan kelompok Risiko Sedang dan Risiko Rendah diambil secara acak dari data hasil penyeragaman skala, masing-masing pada rentang 0,40–0,75 dan 0,1–0,39, agar ketiga titik awal tersebar dan tidak berimpit. Jarak setiap wilayah terhadap ketiga *centroid* diukur menggunakan *Euclidean Distance* pada Persamaan 2, lalu wilayah dimasukkan ke kelompok dengan jarak terkecil. Ukuran jarak ini dipilih karena keempat atribut bersifat numerik dan skalanya sudah setara, kondisi yang paling sesuai bagi jarak *Euclidean* menurut telaah (Abu Alfeilat et al., 2019) mengenai pengaruh pilihan ukuran jarak terhadap kinerja pengklasifikasi.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

Setelah semua wilayah menempati kelompoknya, *centroid* dihitung ulang sebagai rata-rata setiap atribut dari seluruh anggota kelompok tersebut. Ketiga langkah tersebut, yaitu

menghitung jarak, menempatkan wilayah, dan memperbarui *centroid*, dijalankan berulang hingga keanggotaan setiap kelompok tidak berubah lagi.

Mutu pengelompokan dinilai menggunakan *Silhouette Coefficient* (Rousseeuw, 1987). Untuk setiap wilayah dihitung *a(i),* yaitu jarak rata-rata wilayah itu ke anggota lain di kelompoknya sendiri, dan *b(i),* yaitu rata-rata jarak terkecil menuju anggota kelompok lain yang paling dekat. Kedua nilai tersebut digabungkan melalui Persamaan 3, dan rata-rata *s(i)* dari seluruh objek dipakai sebagai ukuran mutu keseluruhan.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}} \tag{3}$$

Integrasi *K-Means* dan KNN untuk Prediksi Risiko

Tahap inilah yang menggabungkan kedua algoritma. Label kelompok hasil pengelompokan dipakai sebagai kelas acuan bagi KNN. Data latih disusun memanjang per tahun: nilai hasil penyeragaman skala tahun 2021, 2022, dan 2023 pada setiap wilayah masing-masing dijadikan satu baris tersendiri beserta label kelompok hasil *clustering K-Means*, sehingga 514 wilayah menghasilkan 1.542 baris data latih. Data uji berupa nilai hasil penyeragaman skala tahun 2024 sebanyak 514 baris tanpa label, dan karena atribut yang dibandingkan hanya satu, jarak Euclidean pada tahap ini menyederhana menjadi selisih mutlak antara dua nilai penyeragaman skala. Susunan inilah yang membuat model membaca kedekatan pola prevalensi tahun 2024 terhadap pola tahun-tahun sebelumnya, sehingga hasilnya merupakan dugaan tingkat risiko setiap wilayah untuk 2025.

Data latih yang berjumlah 1.542 baris tersebut tetap mewakili 514 wilayah berlabel, sehingga jumlah tetangga yang dilibatkan mengikuti aturan akar kuadrat dari jumlah wilayah tersebut (Lall & Sharma, 1996), yaitu $\sqrt{514} = 22{,}67$ yang dibulatkan menjadi 23. Bilangan ganjil dipilih agar pemungutan suara mayoritas tidak berpeluang menghasilkan jumlah suara yang sama antarkategori. Aturan ini diperlakukan sebagai titik awal yang objektif, bukan jaminan nilai terbaik, karena penentuan k memang belum memiliki jawaban tunggal (Zhang, 2022). Kelas setiap wilayah uji diambil dari kategori terbanyak di antara ke-23 tetangga terdekat tersebut sesuai Persamaan 4 (Cover & Hart, 1967).

$$\hat{y} = \arg\max_c \sum_{j \in N_k(x)} I(y_j = c) \tag{4}$$

Lambang $\hat{y}$ adalah kelas hasil dugaan bagi wilayah uji x, C adalah himpunan ketiga kategori risiko, $Nk(x)$ adalah himpunan 23 tetangga terdekat x, y_j adalah label kategori tetangga ke-j, dan I adalah fungsi penanda yang bernilai satu apabila kondisinya terpenuhi dan nol apabila tidak.

Prosedur Validasi Implementasi

Seluruh tahapan di atas dikerjakan dua kali. Pengerjaan pertama dilakukan sel per sel pada Microsoft Excel sehingga setiap rumus penyeragaman skala, perhitungan jarak, pembaruan *centroid*, dan pemungutan suara tercatat dan dapat diperiksa kembali. Pengerjaan kedua mengulang tahapan yang sama pada Orange Data Mining (Demšar et al., 2013) melalui rangkaian *widget* File, Preprocess dengan pilihan *normalize to interval* 0-1, *k-Means*, *Silhouette Plot*, kNN, dan *Predictions*. Parameter kedua pengerjaan dibuat sama, yaitu tiga kelompok, *Euclidean Distance* sebagai ukuran jarak, k sebesar 23, dan bobot tetangga yang seragam. Kesesuaian hasil dinilai dari dua hal, yaitu selisih jumlah anggota setiap kategori risiko dan selisih rata-rata *Silhouette Score*.

HASIL DAN PEMBAHASAN

Kondisi Data Setelah Praproses

Dari 2.056 nilai prevalensi yang tersedia, yaitu 514 wilayah dikalikan empat tahun, sebanyak 128 nilai atau sekitar 6,2% ditemukan kosong. Pengisian dengan nilai rata-rata dipilih agar jumlah objek tidak berkurang, sebab apabila baris yang kosong dihapus maka cakupan wilayah akan menyusut dan perbandingan antarprovinsi menjadi tidak setara. Kabupaten Aceh Barat Daya, misalnya, semula tidak memiliki catatan tahun 2024 dan terisi menjadi 32,1 setelah pengisian dilakukan. Penyeragaman skala berikutnya memindahkan keempat atribut ke rentang nol sampai satu, sehingga tidak ada satu tahun pun yang pengaruhnya lebih besar hanya karena rentang angkanya lebih lebar.

Pembentukan Cluster dengan K-Means

Pengelompokan terhadap 514 wilayah mencapai kondisi stabil pada iterasi kelima belas. Perubahan jumlah anggota setiap kelompok sepanjang iterasi disajikan pada Tabel 2.

Tabel 2.

Perubahan Komposisi Anggota *Cluster* pada Setiap Iterasi *K-Means*

Iterasi	Risiko Tinggi	Risiko Sedang	Risiko Rendah	Keterangan
1	14	290	210	Belum konvergen
2	33	248	233	Belum konvergen
3	42	233	239	Belum konvergen
4	50	225	239	Belum konvergen
5	53	228	233	Belum konvergen
6	56	227	231	Belum konvergen
7	58	225	231	Belum konvergen
8	58	228	228	Belum konvergen
9	58	229	227	Belum konvergen
10	59	228	227	Belum konvergen
11	59	229	226	Belum konvergen
12	59	231	224	Belum konvergen
13	60	231	223	Belum konvergen
14	60	232	222	Belum konvergen
15	60	232	222	Konvergen

Sumber: hasil perhitungan pada Microsoft Excel

Tabel 2 memperlihatkan perpindahan wilayah yang banyak terjadi pada iterasi awal lalu mengecil secara bertahap. Kelompok Risiko Tinggi bertambah dari 14 menjadi 60 wilayah, sedangkan Risiko Sedang menyusut dari 290 menjadi 232 wilayah. Sejak iterasi keempat belas jumlah anggotanya tidak berubah lagi, sehingga hasil iterasi kelima belas ditetapkan sebagai hasil akhir. Titik pusat akhir dan jumlah anggota setiap kelompok disajikan pada Tabel 3.

Tabel 3.

Centroid* Akhir dan Komposisi *Cluster* Hasil *K-Means

Cluster	Kategori Risiko	Th. 2021	Th. 2022	Th. 2023	Th. 2024	Jumlah	Persentase
C3	Risiko Tinggi	0,66	0,69	0,71	0,62	60	11,67%
C2	Risiko Sedang	0,46	0,45	0,49	0,39	232	45,14%
C1	Risiko Rendah	0,31	0,26	0,29	0,22	222	43,19%
	Total	-	-	-	-	**514**	**100,00%**

Sumber: hasil perhitungan pada Microsoft Excel

Urutan besar titik pusat pada Tabel 3 menjadi dasar penamaan ketiga kategori. Kelompok C3 memiliki nilai tertinggi pada keempat tahun sehingga dinamai Risiko Tinggi, sedangkan C1 dengan nilai terendah dinamai Risiko Rendah. Yang menarik, jarak antarkategori tetap

terjaga pada seluruh tahun pengamatan dan bukan hanya pada satu tahun tertentu. Artinya, wilayah yang masuk kategori Risiko Tinggi bukan wilayah yang kebetulan mengalami lonjakan pada satu periode, melainkan wilayah yang prevalensinya tetap tinggi selama empat tahun berturut-turut. Sifat inilah yang membuat pemetaan ini berguna untuk menyusun prioritas, karena yang tersaring adalah masalah yang menetap.

Sebaran anggotanya terpusat pada kategori menengah. Hampir separuh wilayah, yaitu 45,14%, berada pada Risiko Sedang, disusul 43,19% pada Risiko Rendah, sementara Risiko Tinggi hanya berisi 11,67%. Bentuk sebaran seperti ini sejalan dengan temuan (Dayla May Cytry et al., 2023) pada tingkat desa, yang juga mendapati bahwa masalah *stunting* terpusat pada sekelompok wilayah tertentu dan tidak tersebar merata. Bagi perencana program, 60 wilayah pada kategori Risiko Tinggi merupakan jumlah yang masih terjangkau untuk ditangani secara intensif.

Penilaian Mutu Cluster

Perhitungan *Silhouette Coefficient* terhadap seluruh 514 wilayah menghasilkan rata-rata 0,33. Kriteria yang dipakai untuk menafsirkan nilai tersebut disajikan pada Tabel 4.

Tabel 4.

Kriteria Interpretasi Rata-Rata *Silhouette Score*

Selang Nilai	Penafsiran Struktur *Cluster*
0,71 – 1,00	Sangat kuat
0,51 – 0,70	Baik
0,26 – 0,50	Cukup baik
0,00 – 0,25	Lemah
< 0,00	Kemungkinan salah pengelompokan

Sumber: Rousseeuw (1987); Hasan (2024)

Merujuk Tabel 4, nilai 0,33 masuk kelompok 0,26–0,50, yang berarti struktur kelompoknya tergolong cukup baik. Pengelompokan sudah mampu memisahkan karakteristik wilayah, meskipun batas antarkelompok belum sepenuhnya tegas. Hasil seperti ini wajar terjadi pada data prevalensi wilayah yang nilainya berurutan tanpa jarak pemisah yang alami, sebab sebagian wilayah memang berada di daerah peralihan antara dua kategori. Nilai tersebut masih cukup untuk menjadikan label kelompok sebagai kelas acuan pada tahap klasifikasi.

Sebagai pembanding, penelusuran nilai k dari dua sampai delapan pada Orange Data Mining memberikan gambaran seperti pada Tabel 5.

Tabel 5.

Rata-Rata *Silhouette Score* pada Berbagai Nilai *k*

Jumlah *Cluster* (k)	2	3	4	5	6	7	8
Rata-rata *Silhouette Score*	0,395	0,342	0,266	0,211	0,232	0,226	0,243

Sumber: penelusuran pada Orange Data Mining

Tabel 5 memperlihatkan bahwa dua kelompok justru memberikan rata-rata tertinggi, yaitu 0,395, dan nilainya menurun ketika jumlah kelompok ditambah. Apabila hanya berpatokan pada angka tersebut, *k* sebesar dua adalah pilihan terbaik. Penelitian ini tetap menetapkan tiga kelompok karena dua alasan. Pertama, selisihnya terhadap *k* sebesar tiga hanya 0,053 dan keduanya berada pada kategori penafsiran yang sama. Kedua, pembagian menjadi dua kategori hanya menghasilkan pemisahan tinggi dan rendah, sehingga tidak ada tempat bagi kelompok peralihan padahal justru di situlah sebagian besar wilayah berada.

Prediksi Kategori Risiko Tahun 2025

Klasifikasi KNN dengan *k* sebesar 23 menghasilkan dugaan kategori risiko tahun 2025 untuk seluruh wilayah. Pemungutan suara dilakukan atas 23 baris data latih terdekat menurut kemiripan nilai penyeragaman skalanya. Pada data uji Kabupaten Simeulue, misalnya, jarak ke-23 tetangga terdekatnya berkisar antara 0,0003 dan 0,0037, dan tetangga tersebut berasal dari berbagai kabupaten/kota serta tahun pengamatan yang berbeda. Jarak yang sangat kecil itu muncul karena perbandingan hanya memakai satu atribut, sehingga kedekatan yang dibaca model adalah kedekatan pola angka pada tahun tertentu, bukan kedekatan letak geografis.

Hasil dugaan tersebut menggeser jumlah anggota setiap kategori dibandingkan hasil pengelompokan. Kategori Risiko Tinggi menyusut tajam dari 60 menjadi 25 wilayah, Risiko Sedang menyusut dari 232 menjadi 194 wilayah, sedangkan Risiko Rendah bertambah dari 222 menjadi 295 wilayah. Arah pergeserannya sejalan dengan penurunan prevalensi nasional yang terekam pada empat tahun pengamatan, sehingga dugaan model konsisten dengan kecenderungan yang ada pada datanya. Dicocokkan kembali dengan label hasil *K-Means*, dugaan KNN sejalan pada 378 dari 514 wilayah, atau 73,54%, sehingga sebagian besar pergeseran kategori tetap searah dengan pola clustering yang sama. Meski demikian, masih tersisa 25 wilayah yang diperkirakan tetap berada pada kategori Risiko Tinggi. Wilayah

itulah yang memerlukan penanganan berkelanjutan, sebab pola prevalensinya tidak menunjukkan tanda mengikuti perbaikan nasional.

Validasi Rangkaian pada Dua Perangkat

Perbandingan jumlah anggota setiap kategori risiko hasil pengelompokan pada kedua perangkat disajikan pada Tabel 6.

Tabel 6.

Komposisi Kategori Risiko Hasil *K-Means* pada Kedua Perangkat

Kategori Risiko	Microsoft Excel	Orange Data Mining	Selisih
Risiko Tinggi (C3)	60	56	4
Risiko Sedang (C2)	232	233	1
Risiko Rendah (C1)	222	225	3
Total	**514**	**514**	–

Sumber: hasil pengolahan data penelitian

Selisih terbesar pada Tabel 6 tercatat empat wilayah, yaitu pada kategori Risiko Tinggi, atau kurang dari satu persen dari keseluruhan objek. Sumber selisih ini dapat ditelusuri pada titik pusat awal: pada Microsoft Excel titik Risiko Tinggi ditetapkan dari nilai penyeragaman skala tertinggi, sedangkan titik Risiko Sedang dan Risiko Rendah diambil acak dari rentang nilai menengah dan rendah; pada Orange Data Mining ketiga titik itu ditentukan sendiri secara acak. Karena *K-Means* berhenti pada minimum lokal yang bergantung pada titik awalnya, kedua pengerjaan berakhir pada titik pusat dan jumlah anggota yang sedikit berbeda. Selisih sekecil itu tidak mengubah urutan besar ketiga kategori maupun bentuk sebarannya.

Rata-rata *Silhouette Score* memperkuat gambaran tersebut. Microsoft Excel mencatat 0,33 dan Orange Data Mining 0,342, dengan selisih 0,012. Keduanya berada pada rentang penafsiran yang sama, sehingga mutu pengelompokan yang dihasilkan kedua perangkat sebanding. Selisih yang kecil itu berasal dari nilai *a(i)* dan *b(i)* yang dihitung atas keanggotaan yang tidak sepenuhnya sama, ditambah pembulatan bilangan desimal pada pengerjaan manual.

Pada tahap klasifikasi kesesuaiannya lebih dekat lagi, sebagaimana terlihat pada Tabel 7.

Tabel 7.

Komposisi Prediksi Risiko Tahun 2025 Hasil KNN pada Kedua Perangkat

Kategori Risiko	Microsoft Excel	Orange Data Mining	Selisih
Risiko Tinggi (C3)	25	23	2
Risiko Sedang (C2)	194	195	1
Risiko Rendah (C1)	295	296	1
Total	**514**	**514**	–

Sumber: hasil pengolahan data penelitian

Tabel 7 mencatat selisih paling banyak dua wilayah, yaitu pada kategori Risiko Tinggi, dan hanya berbeda satu wilayah pada dua kategori lainnya. Selisih itu tidak bersumber dari unsur acak seperti penetapan titik pusat awal pada *K-Means*: KNN tidak memuat unsur acak semacam itu, sehingga sumber selisihnya adalah keadaan seri, yaitu banyak baris data latih yang berjarak sama persis terhadap satu data uji karena jarak hanya dihitung atas satu atribut. Pemilihan 23 tetangga terdekat pun bergantung pada cara masing-masing perangkat memutus keadaan seri tersebut: Microsoft Excel memutusnya berdasarkan nomor urut data latih, sedangkan Orange Data Mining memakai mekanisme internalnya sendiri, dan perbedaan itulah yang menggeser sebagian wilayah yang berada di perbatasan antar-kategori.

Berdasarkan ketiga perbandingan tersebut, hasil kerja manual di Microsoft Excel dapat dinyatakan sesuai dengan keluaran Orange Data Mining. Pengerjaan sel per sel pada lembar kerja membuat setiap tahap perhitungan terbuka untuk diperiksa, sehingga hasilnya dapat dipakai untuk menunjang keputusan publik, keterbukaan seperti ini memperkuat pertanggungjawaban hasil. Sebaliknya, perangkat *data mining* tetap diperlukan sebagai pemeriksa pembanding yang cepat.

KESIMPULAN DAN REKOMENDASI

Penggabungan *K-Means Clustering* dan *K-Nearest Neighbor* dalam satu rangkaian kerja terbukti dapat dijalankan atas data prevalensi *stunting* balita 514 kabupaten/kota dari SSGI 2021 sampai 2024. Tahap pengelompokan membentuk tiga kategori wilayah yang stabil pada iterasi kelima belas, terdiri atas 60 wilayah pada kategori Risiko Tinggi, 232 pada Risiko Sedang, dan 222 pada Risiko Rendah, dengan rata-rata *Silhouette Score* 0,33 yang tergolong cukup baik. Label ketiga kategori tersebut kemudian dipakai sebagai kelas acuan

pada tahap klasifikasi, dan dengan 23 tetangga terdekat dugaan untuk tahun 2025 bergeser menjadi 25 wilayah pada kategori Risiko Tinggi, 194 pada Risiko Sedang, serta 295 pada Risiko Rendah, sejalan dengan kecenderungan penurunan prevalensi nasional dan sejalan dengan label *K-Means* pada 378 dari 514 wilayah, atau 73,54%. Pengulangan seluruh rangkaian pada Orange Data Mining memberikan selisih paling besar empat wilayah pada tahap pengelompokan, 0,012 pada rata-rata *Silhouette Score*, dan dua wilayah pada tahap klasifikasi, sehingga perhitungan yang dikerjakan pada Microsoft Excel dinyatakan sesuai.

Bagi pemangku kebijakan, 25 wilayah yang diperkirakan tetap berada pada kategori Risiko Tinggi layak dijadikan sasaran utama percepatan penurunan *stunting*, sebab pola prevalensinya tidak mengikuti perbaikan yang terjadi pada tingkat nasional. Bagi penelitian selanjutnya, terbuka empat arah pengembangan. Pertama, menambah variabel dengan faktor sosial ekonomi dan lingkungan agar pemetaan tidak berhenti pada sebaran tetapi juga menerangkan penyebabnya. Kedua, membandingkan *K-Means* dengan metode pengelompokan lain seperti *Fuzzy C-Means* atau metode berbasis kerapatan yang lebih sesuai bagi data yang nilainya berurutan tanpa jarak pemisah alami. Ketiga, memakai ukuran mutu tambahan seperti *Davies-Bouldin Index* agar penentuan jumlah kelompok tidak bergantung pada satu ukuran saja. Keempat, menuangkan hasil pemetaan ke dalam *Dashboard* sehingga pembaruan data dapat dilakukan setiap kali laporan survei terbaru diterbitkan.

REFERENSI

1. Abu Alfeilat, H. A., Hassanat, A. B. A., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. B. S. (2019). Effects of distance measure choice on K-nearest neighbor classifier performance: A review. *Big Data, 7*(4), 221–248. https://doi.org/10.1089/big.2018.0175
2. Chrisamudra, R. W., & Abdurrahman, S. (2025). Pemanfaatan algoritma K-Means untuk klastering spasial beban kasus pneumonia pada kelompok balita di wilayah dengan kepadatan populasi tinggi. *TIN: Terapan Informatika Nusantara, 6*(6), 761–768. https://doi.org/10.47065/tin.v6i6.8598

3. Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

4. Cytry, D. M., Defit, S., & Nurcahyo, G. (2023). Penerapan metode K-Means dalam klasterisasi status desa terhadap keluarga beresiko stunting. *Jurnal KomtekInfo*, 122–127. https://doi.org/10.35134/komtekinfo.v10i3.423

5. Demšar, J., Curk, T., Erjavec, A., Hočevar, T., Milutinovič, M., Možina, M., Toplak, M., Umek, L., Zbontar, J., & Zupan, B. (2013). Orange: Data mining toolbox in Python. *Journal of Machine Learning Research, 14*.

6. Insany, G. P., Yustiana, I., & Rahmawati, S. (2023). Penerapan KNN dan ANN pada klasifikasi status gizi balita berdasarkan indeks antropometri. *Jurnal CoSciTech (Computer Science and Information Technology), 4*(2), 385–393. https://doi.org/10.37859/coscitech.v4i2.5079

7. Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

8. Hartanto, U. I., Putu, G., Buditjahjanto, A., & Yustanti, W. (2025). Hybrid clustering and classification of at-risk customer segments in network marketing. *JIEET (Journal of Information Engineering and Educational Technology), 9*(1), 42–50. https://doi.org/10.26740/jieet.v9n1.p42-50

9. Hartono, B., & Lusiana, V. (2026). Analisis metode Elbow SSE, Silhouette Score, dan Jaccard Stability dalam pemilihan jumlah klaster data yang optimal. *TIN: Terapan Informatika Nusantara, 6*(8), 1521–1532. https://doi.org/10.47065/tin.v6i8.9271

10. Hasan, Y. (2024). Pengukuran Silhouette Score dan Davies-Bouldin Index pada hasil cluster K-Means dan DBSCAN, *6*(1).

11. Kadafi, M. N., & Finandhita, A. (2024). Penerapan data mining dalam pengelompokan penyakit untuk rekomendasi materi penyuluhan kesehatan di Klinik Keluarga Cianjur. *Komputa: Jurnal Ilmiah Komputer dan Informatika, 13*(1), 33–42. https://doi.org/10.34010/komputa.v13i1.11273

12. Kementerian Kesehatan Republik Indonesia. (2020). *Peraturan Menteri Kesehatan Republik Indonesia Nomor 2 Tahun 2020 tentang standar antropometri anak*. https://peraturan.bpk.go.id/Details/152505/permenkes-no-2-tahun-2020

13. Kementerian Kesehatan Republik Indonesia. (2021). *Hasil SSGI tahun 2021 tingkat kabupaten/kota*. https://www.badankebijakan.kemkes.go.id/buku-saku-hasil-studi-status-gizi-indonesia-ssgi-tahun-2021/

14. Kementerian Kesehatan Republik Indonesia. (2022). *Hasil survei status gizi Indonesia*. https://stunting.go.id/buku-saku-hasil-survei-status-gizi-indonesia-ssgi-2022/

15. Kementerian Kesehatan Republik Indonesia. (2023). *Hasil Survei Kesehatan Indonesia (SKI) 2023*. https://www.badankebijakan.kemkes.go.id/hasil-ski-2023

16. Kementerian Kesehatan Republik Indonesia. (2024). *Hasil Survei Status Gizi Indonesia (SSGI) 2024*. https://www.badankebijakan.kemkes.go.id/survei-status-gizi-indonesia-ssgi-2024/

17. Kementerian Perencanaan Pembangunan Nasional. (2025). *Ringkasan Rencana Pembangunan Jangka Menengah Nasional tahun 2025–2029*.

18. Khusaeri, A. (2024). Analisis algoritma K-Means clustering dalam pengelompokan daerah penyebaran penyakit demam berdarah dengue. *Journal of Information System, Informatics and Computing, 8*(2), 434. https://doi.org/10.52362/jisicom.v8i2.1795

19. Lall, U., & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research, 32*(3), 679–693. https://doi.org/10.1029/95WR02966

20. MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*.

21. Putra, S. H., Jamaludin, J., Simbolon, H., Ibrahim, I., & Rubianto, R. (2025). Data mining untuk pola perbaikan pelayanan pasien dengan metode klustering K-Means pada Rumah Sakit Mitra Sejati. *Remik, 9*(3), 1000–1012. https://doi.org/10.33395/remik.v9i3.15192

22. Ramadhani, F. (2025). Spatial clustering analysis of stunting in North Sumatra based on environmental factors using K-Means algorithm. *Data Science: Journal of Computing and Applied Informatics, 9*(2). https://doi.org/10.32734/jocai.v9.i2-17179

23. Republik Indonesia. (2021). *Peraturan Presiden Republik Indonesia Nomor 72 Tahun 2021 tentang percepatan penurunan stunting*.

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

25. Schubert, E. (2023). Stop using the elbow criterion for K-Means and how to choose the number of clusters instead. *ACM SIGKDD Explorations Newsletter, 25*(1), 36–42. https://doi.org/10.1145/3606274.3606278

26. Syriopoulos, P. K., Kalampalikis, N. G., Kotsiantis, S. B., & Vrahatis, M. N. (2025). kNN classification: A review. *Annals of Mathematics and Artificial Intelligence, 93*(1), 43–75. https://doi.org/10.1007/s10472-023-09882-x

27. Wulandari, I., Darmawan, E. S., & Zamzanah, S. (2025). Spatial Bayesian small area estimation of stunting prevalence at the subdistrict level. *InPrime: Indonesian Journal of Pure and Applied Mathematics, 7*(2), 128–143. https://doi.org/10.15408/inprime.v7i2.46491

28. Yohanis Ndapa Deda, Maria Paula Marla Nahak, & Aplonia Pala. (2024). *Tren global stunting dalam satu dekade terakhir: Analisis bibliometrik*.

29. Zhang, S. (2022). Challenges in KNN classification. *IEEE Transactions on Knowledge and Data Engineering, 34*(10), 4663–4675. https://doi.org/10.1109/TKDE.2021.3049250

Integration of K-Means Clustering and K-Nearest Neighbor for Mapping and Predicting Stunting Risk Among Toddlers in Indonesia

ORIGINALITY REPORT

8%	7%	5%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Perguruan Tinggi Pelita Bangsa Student Paper	6%
2	journal.universitaspahlawan.ac.id Internet Source	2%
3	repository.unissula.ac.id Internet Source	1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		