Journal of Big Data Analytic and Artificial Intelligence
Volume 8(1), Juni 2025, pp. 16-21

Optimalisasi Pengelompokkan Konsumen dengan Multi Internal Metric Validation dan Boxplot Analysis

Rachmad Fitriyanto[1], Nurindah[2]

[1*,2]Program Studi Sistem Informasi, STMIK PPKIA Tarakanita Rahmawati, Tarakan, Kalimantan Utara
Email: [1*]fitriyanto7477@gmail.com, [2]nur291030@gmail.com

Abstrak

Penggunaan metrik-metrik validasi internal secara bersamaan untuk menentukan jumlah klaster optimal pada algoritma K-Means Clustering berpotensi memberikan nilai K yang berbeda-beda dan membingungkan bagi pengguna data untuk mengekstraksi informasi seperti untuk identifikasi karakteristik pelanggan. Penelitian ini bertujuan mengembangkan framework evaluasi untuk mengatasi ambiguitas nilai K dari beberapa metrik validasi internal. Framework evaluasi nilai K terdiri dari 2 tahap, dengan tahap pertama menggunakan lima metrik validasi, DBI, Silhouette Score, Elbow Method, Dunn Index dan Calinski-Harabasz Index sebagai filter untuk menghasilkan maksimal 5 nilai K terbaik. Evaluasi tahap kedua menggunakan analisis boxplot, inter quartile range dan elbow untuk mengeksplorasi tingkat kohesifitas dan stabilitas klaster yang terbentuk. Hasil yang diperoleh dari evaluasi tahap 1 terdapat 4 opsi jumlah klaster, K=2, 5, 7 dan 10. Evaluasi tahap kedua menunjukkan dari grafik elbow nilai rata-rata inter quartile range, K=5 menjadi jumlah klaster yang optimal dibanding 3 nilai K lainnya. Hasil ini menunjukkan bahwa semakin banyak metrik validasi internal yang digunakan, berpotensi menghasilkan banyak nilai K. Semakin banyak jumlah klaster tidak menjamin semakin baik kualitasnya. Implikasi dari penelitian ini menunjukkan pentingnya pendekatan evaluasi berlapis dalam menentukan jumlah klaster optimal, terutama saat menggunakan banyak metrik validasi internal secara bersamaan. Framework yang dikembangkan dapat membantu praktisi data dalam membuat keputusan yang lebih terinformasi dan mengurangi ambiguitas dalam proses klasterisasi. Ke depan, framework ini dapat dikembangkan lebih lanjut dengan mengintegrasikan metrik eksternal atau diadaptasi untuk algoritma klasterisasi lainnya.

Kata Kunci: Boxplot, Inter Quartile Range, K-Means Clustering, Segmentasi konsumen, validasi internal.

Optimizing Customer Segmentation Using Multi-Metric Internal Validation and Boxplot Analysis

Abstract

The simultaneous use of multiple internal validation metrics to determine the optimal number of clusters in K-Means Clustering often results in differing K values, which can confuse data practitioners when extracting insights, such as identifying customer characteristics. This study aims to develop an evaluation framework to address the ambiguity arising from varying K values produced by different internal validation metrics. The proposed K evaluation framework consists of two stages. In the first stage, five internal validation metrics—Davies-Bouldin Index (DBI), Silhouette Score, Elbow Method, Dunn Index, and Calinski-Harabasz Index—are used as filters to generate up to five top K candidates. The second stage involves boxplot analysis, interquartile range (IQR), and elbow visualization to explore the cohesiveness and stability of the resulting clusters. The first-stage evaluation yielded four potential cluster counts: K = 2, 5, 7, and 10. In the second stage, based on the elbow graph of the average interquartile range, K = 5 was identified as the most optimal number of clusters compared to the other candidates. These results indicate that using a larger number of internal validation metrics may increase the likelihood of producing multiple K values. However, a higher number of clusters does not necessarily guarantee better quality. The implications of this research highlight the importance of a layered evaluation approach in determining the optimal number of clusters, especially when employing multiple internal validation metrics. The proposed framework can assist data practitioners in making more informed decisions and reducing ambiguity in the clustering process. In the future, this framework can be extended by incorporating external validation metrics or adapted to other clustering algorithms.

Keywords: *Boxplot,Customer Segmentation, Internal Validation, Inter Quartile Range, K-Means Clustering.*

I. PENDAHULUAN

Perkembangan teknologi informasi membuat data menjadi sumber daya penting bagi perkembangan usaha, mulai usaha kecil dan menengah (UKM) sampai skala internasional. Data konsumen menjadi acuan bagi pemilik dan pengelola usaha untuk menentukan strategi bisnis agar mampu mengambil Keputusan yang dapat memajukan usahanya. Knowledge Discovery in Database (KDD) menjadi konsep utama yang digunakan untuk mengekstraksi informasi dari kumpulan data yang berjumlah banyak. KDD memberikan kerangka kerja bagi penggunanya untuk mempersiapkan data, mengolah dan mengekstraksi informasi. Data mining, menjadi salah satu tahapan dalam KDD yang digunakan untuk memproses data dengan teknik-teknik yang dipilih disesuaikan dengan bentuk data dan tujuan pengolahan data.

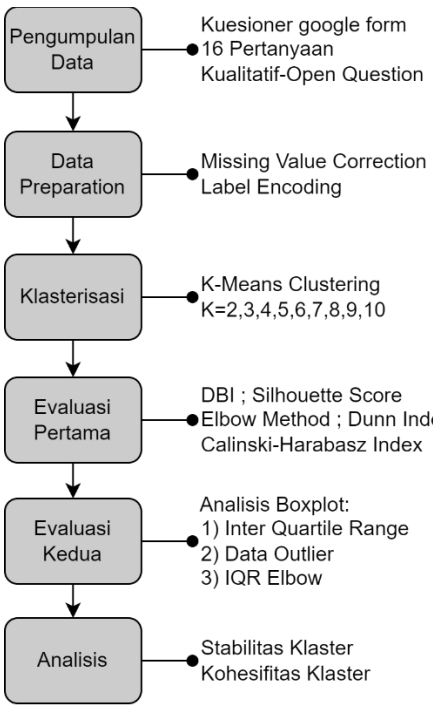
Clustering merupakan teknik data mining untuk mengolah data yang belum memiliki label atau kelas yang mendeskripsikan data. Penerapan teknik clustering untuk pengambilan Keputusan dalam dunia usaha adalah untuk segmentasi pelanggan. Data mining dan teknik clustering menawarkan banyak keuntungan untuk bisnis UKM (Usaha Kecil dan Menengah). Beberapa keuntungan tersebut adalah sebagai dasar pengambilan keputusan bisnis yang lebih objektif berdasarkan pola dan tren data aktual. Selain itu dapat juga untuk mengurangi risiko keputusan yang hanya berdasarkan intuisi atau pengalaman terbatas. Dari sisi pengembangan usaha, dapat memberikan manfaat dengan cara mengidentifikasi tren pasar yang mungkin tidak terlihat secara kasat mata dan dengan cara mmahami perubahan perilaku konsumen secara lebih akurat [1], [2].

Algoritma clustering bekerja dengan cara mengelompokkan data berdasarkan kemiripan karakteristik dari atribut-atribut atau data feature. Beberapa jenis metode clustering yang dapat diimplementasikan, antara lain hierarchical clustering, density-base clustering, dan partition clustering. K-Means Clustering adalah algoritma clustering kategori partition clustering, yang terkenal dengan kesederhanaan cara kerja dan tingkat kompleksitas algoritma yang rendah [3], [4]. Algoritma ini bekerja dengan menentukan K pusat cluster secara acak, kemudian mengalokasikan setiap objek data ke cluster terdekat berdasarkan jarak Euclidean, dan menghitung ulang pusat cluster sebagai rata-rata dari semua objek dalam cluster tersebut. Proses ini diulang hingga pusat cluster tidak berubah secara signifikan atau kriteria penghentian lainnya terpenuhi.

Meskipun K-Means Clustering menawarkan berbagai keunggulan, algoritma ini memiliki keterbatasan yang signifikan, yaitu saat menentukan jumlah cluster (nilai k) di awal proses yang dilakukan secara acak atau sesuai keinginan pengguna. Penentuan nilai K yang tidak optimal dapat menghasilkan kelompok data yang kurang representatif terhadap struktur data yang sebenarnya, sehingga mengurangi kualitas segmentasi dan berpotensi menghasilkan keputusan bisnis yang kurang tepat [5]. Untuk mengatasi permasalahan ini, berbagai metode validasi internal seperti Silhouette Score dan Davies-Bouldin Index telah banyak digunakan untuk mengevaluasi kualitas clustering dan menentukan jumlah cluster optimal. Silhouette Score mengukur seberapa baik objek dikelompokkan dalam clusternya dibandingkan dengan cluster lainnya, sementara Davies-Bouldin Index mengevaluasi rasio antara jarak intra-cluster dan inter-cluster. Namun, penelitian sebelumnya menunjukkan bahwa kedua metrik tersebut terkadang memberikan nilai k optimal yang berbeda, menghasilkan ambiguitas dalam pengambilan Keputusan [5]. Penelitian tersebut memunculkan pertanyaan baru tentang hasil evaluasi internal dengan teknik lain seperti Dunn Index, Elbow Method dan Calinski-Harabasz Index. Apakah teknik-teknik validasi tersebut juga menimbulkan inkonsistensi jumlah klaster saat dibandingkan dengan teknik lainnya. Oleh karena itu, penelitian ini bertujuan untuk mengeksplorasi dan membandingkan ketiga teknik validasi dengan dua teknik validasi dari penelitian sebelumnya, dalam konteks segmentasi pelanggan minuman boba, serta mengusulkan pendekatan integratif untuk mengatasi potensi ketidakkonsistenan dalam penentuan jumlah klaster optimal. Penelitian ini diharapkan untuk memberikan kontribusi signifikan dalam beberapa aspek. Pertama, pengembangan metodologi yang sistematis untuk mengatasi ketidakkonsistenan antara metrik validasi internal. Kedua, penyediaan framework pengambilan keputusan yang lebih objektif dan robust untuk penentuan jumlah cluster optimal yang dapat diadopsi oleh praktisi bisnis dengan pengetahuan statistik terbatas

II. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode K-Means Clustering untuk segmentasi pelanggan minuman boba yang dikembangkan dari metodologi penelitian sebelumnya [nurindah, techno vol.1]. Gambar 1 mengilustrasikan tahapan penelitian yang dimaksud.



Gambar 1. Metodologi Penelitian

Tahap pertama penelitian adalah perancangan kuesioner yang komprehensif untuk mengumpulkan data karakteristik

Journal of Big Data Analytic and Artificial Intelligence
Volume 8(1), Juni 2025, pp. 16-21

pelanggan yang terdiri dari 16 pertanyaan seperti dirumuskan dalam Tabel 1.

Tabel 1. Pertanyaan

No.	Fitur / Pertanyaan	Pilihan Jawaban
1	Usia Pelanggan	17 tahun; 17-24 Tahun; 25-34 tahun; 35-44 tahun
2	Jam Pembelian	10 sd 12; 12 sd 15; 15 sd 17; 17 sd 20; 20 sd 22
3	Jumlah Pembelian	1 Cup; 2 Cup; 3 sd 5 Cup
4	Ukuran Cup	Medium; Jumbo
5	Persepsi harga jual	Murah; Standar; Mahal
6	Pengalaman promo	Pernah; Tidak Pernah
7	Kualitas kemasan	Kurang Memadai; Memadai
8	Frekuensi pembelian dalam 1 pekan	1 Kali; 2-3 Kali; >3 Kali
9	Pekerjaan	Pelajar; Mahasiswa; Karyawan Swasta; Wirausaha; PNS
10	Jarak tempat tinggal ke counter	< 1 Km; 1-5 Km; 5-10Km; >50Km
11	Menu favorit	Boba; Non Boba
12	Kualitas desain cup	Biasa; Menarik
13	Kesesuaikan harga dengan rasa	Sesuai; Tidak sesuai
14	Fasilitas yang kurang baik	Tenda; Tempat Duduk; Tempat Parkir
15	Sumber informasi tentang produk	Teman; Instagram; Facebook
16	Tingkat kepuasan pelayanan	Puas; Tidak Puas

Kuesioner dibagikan kepada pelanggan dan diisi saat pelanggan menunggu pesanan mereka dibuat. Pelanggan mengisi langsung pada perangkat smartphone dan tablet yang telah disediakan. Target responden yang diharapkan sebanyak minimal 100 responden dari beragam umur dan pekerjaan.

Tahap data preparation untuk memastikan tidak adanya missing-value dari dataset yang digunakan dan dilanjutkan dengan proses label-encoding untuk mengubah nilai kualitatif menjadi numerik. Klasterisasi pada tahap ketiga menggunakan K-Means Clustering dengan variasi nilai K mulai dari K=2 sampai K=10. Aplikasi bantu untuk klasterisasi menggunakan jupyter notebook dengan Bahasa pemrograman python. Formula penghitungan jarak antar data menggunakan Euclidien Distance seperti ditunjukkan pada formula (1)[6]

$$d(i,j) = \sqrt{(x_{1i} - x_{1j})^2 + \cdots + (x_{ki} - x_{kj})^2} \quad (1)$$

Tahap keempat penelitian adalah evaluasi pertama menggunakan 5 metrik internal validasi. Metrik pertama adalah Davies-Bouldin Index yang dihitung menggunakan formula (2)[5], [6].

$$DBI = \frac{1}{k} * \sum_{i=1}^{k} max_{j \neq 1} \frac{S_i + S_j}{d(c_i, c_j)} \quad (2)$$

Metrik kedua adalah silhouette score dihitung menggunakan formula (3)[5].

$$S = \frac{1}{n} * \sum_{i=1}^{n} S_i \quad (3)$$

Keterangan:
S : Nilai Silhouette Score
n : Jumlah Klaster
Si : Nilai Silhouette Score pada klaster ke i

Metrik ketiga adalah elbow method, dihitung menggunakan formula (4)[7], [8].

$$SSE = \sum_{i=1}^{k} \sum_{x_i \in C_k} (x_i - \varphi_k)^2 \quad (4)$$

Keterangan:
SSE : Sum of Square Error (nilai elbow)
K : jumlah klaster
x_i : data ke x pada fitur ke i
φ_k : rata-rata K clusterpada nilai k (k=1,2,3,...,K)

Metrik keempat adalah Dunn Index, dihitung menggunakan formula (5)[9], [10].

$$D_{nc} = min_{i=1,..,nc} \left\{ min_{j=i+1,..,nc} \left(\frac{d(c_i, c_j)}{max_{k=1,..,nc} * Diameter(c_k)} \right) \right\} \quad (5)$$

Metrik kelima adalah Calinski-Harabasz Index (CHI) dihitung menggunakan formula (6)(7) dan (8)[11].

$$s = \frac{tr(B_k)}{tr(W_k)} * \frac{n-k}{n-1} \quad (6)$$

$t_r(B_k)$ adalah trace dari matriks disperse antar kluster dan $t_r(W_k)$ adalah trace matriks disperse dalam kluster yang dihitung menggunakan formula (7) dan (8)

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_x)(c_q - c_x)^T \quad (7)$$

$$B_k = \sum_{q=1}^{k} (x - c_q)(x - c_x)^T \quad (8)$$

C_q adalah kumpulan titik dalam kluster q, c_q adalah pusat kluster q, C_x adalah pusat data klaster x, dan n_q adalah jumlah titik dalam kluster q.

Evaluasi tahap 2 menggunakan analisis boxplot untuk memilih nilai K optimal dari masing-masing metrik validasi yang digunakan. Tujuan penggunaan boxplot untuk mengevaluasi kelima hasil klasterisasi dari evaluasi tahap 1 dari perspektif kohesifitas klaster yang terbentuk. Pada tahap ini menggunakan 2 bentuk nilai inter quartile range (IQR), yaitu rata-rata IQR setiap klaster di masing-masing nilai K seperti ditunjukkan pada formula (9) dan rata-rata IQR dari masing-masing nilai K pada formula (10).

$$IQR_{kc} = \frac{\sum_{i=1}^{n} Q3_i - Q1_i}{n} \quad (9)$$

Keterangan:
IQR_{kc} : nilai rata-rata IQR pada cluster ke c dengan jumlah klaster K
$Q3_i$: nilai Q3 pada fitur ke i
$Q1_i$: nilai Q1 pada fitur ke i
n : jumlah fitur data

$$IQR_m = \frac{\sum_{c=1}^{m} IQR_c}{m} \quad (10)$$

Keterangan:
IQR_m : nilai rata-rata IQR pada dengan jumlah klaster m
IQR_c : nilai rata-rata IQR pada klaster c
m : jumlah klaster

III. HASIL DAN PEMBAHASAN

Proses klasterisasi menggunakan tool Jupyter Notebook dan library dari Python untuk algoritma klasterisasi serta evaluasi jumlah klaster menggunakan 5 metrik validasi internal. Tabel 2 menunjukkan nilai masing-masing metrik validasi internal untuk 9 variasi nilai K.

Tabel 2. Hasil Metrik Validasi Internal

K	DBI	SC	Elbow	Dunn	CHI
2	2.6631	0.0942	611.20	0.1336	11.1814
3	2.9844	0.0681	586.28	0.2209	7.8304
4	2.6828	0.0440	552.44	0.2209	7.4432
5	2.6893	0.0486	525.26	0.1562	7.0394
6	2.5338	0.0412	498.44	0.2209	6.8835
7	2.4109	0.0415	479.19	1.4142	6.5258
8	2.4109	0.0380	471.57	0.1768	5.8353
9	2.3748	0.0348	462.80	0.1768	5.3615
10	2.3211	0.0267	450.50	0.1768	5.1152
Optimal K	10	2	5	7	2

Nilai K optimal berdasarkan DBI terdapat pada K=10 yang tidak ditemui di empat metrik evaluasi lainnya. Dari evaluasi nilai silhouette score ditemukan K optimal pada K=2 yang memiliki kesamaan dengan Calinski-Harabasz Index. Pada nilai WCSS dari elbow method ditemukan nilai K optimal pada K=5 yang juga seperti DBI tidak ditemui pada metrik lainnya. Sedangkat untuk nilai Dunn-Index menunjukkan nilai K optimal pada penelitian ini terdapat di K=7 yang sekali lagi tidak ditemukan di empat metrik lainnya.

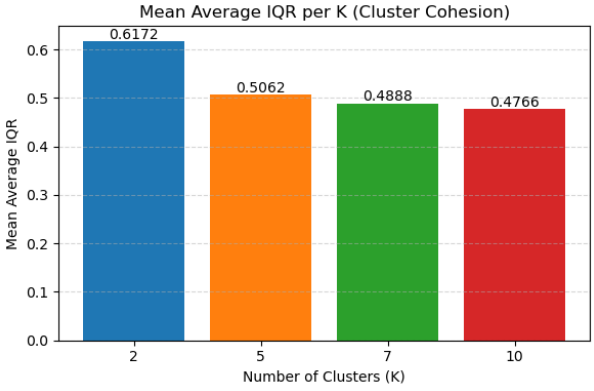
Hasil evaluasi dari 5 metrik validasi internal klaster menunjukkan 4 nilai K optimal yang berbeda-beda, yaitu K=2, 5, 7 dan 10. Hasil tersebut sesuai dengan penelitian sebelumnya yang menunjukkan nilai K yang berbeda dengan dataset yang berbeda dan algoritma klasterisasi yang berbeda pula [techno:Rachmad-Ardi].Perbedaan inilah menguatkan alasan penggunaan satu metrik validasi internal klaster tidak mungkin dilakukan untuk menentukan jumlah klaster terbaik.

Hasil evaluasi tahap pertama menunjukkan 4 nilai K yaitu K=2, 5, 7 dan 10. Hasil klasterisasi dengan keempat nilai K tersebut dievaluasi di tahap kedua diawali menghitung nilai rata-rata IQR untuk masing-masing fitur data di setiap nilai K dan dilanjutkan menghitung nilai rata-rata IQR masing-masing nilai K seperti dirumuskan pada Tabel 3.

Tabel 3. Nilai Rata-Rata IQR per Cluster dan per Nilai K

K	Klaster	Avg. IQR per Cluster	Avg. IQR per K Value
2	C1	0.5938	0.6172
	C2	0.6406	
5	C1	0.4375	0.5063
	C2	0.4531	
	C3	0.6406	
	C4	0.4688	
	C5	0.5313	
7	C1	0.4688	0.4888
	C2	0.5625	
	C3	0.5000	
	C4	0.5156	
	C5	0.3438	
	C6	0.6250	
	C7	0.4063	
10	C1	0.4688	0.4766
	C2	0.4375	
	C3	0.4688	
	C4	0.4375	
	C5	0.7031	
	C6	0.5000	
	C7	0.4688	
	C8	0.2188	
	C9	0.6250	
	C10	0.4375	

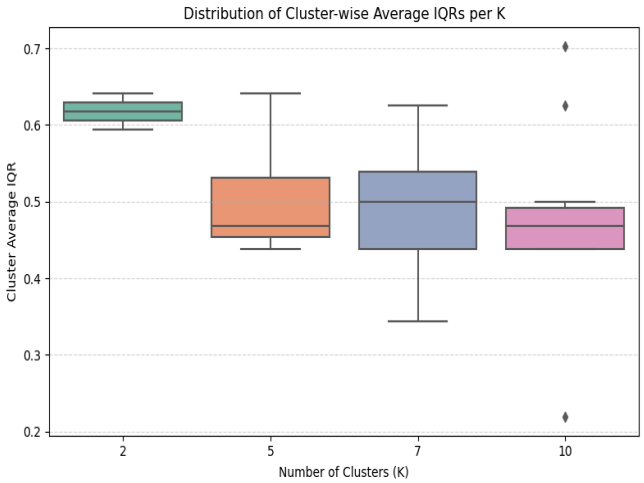
Tabel 3 terdiri dari 4 kolom, dengan kolom paling kiri menunjukkan nilai K hasil filter dari evaluasi tahap pertama yaitu, K=2, ,5, 7 dan K=10. Kolom kedua berisikan kode klaster di masing-masing nilai K. kolom ketiga berisikan nilai rata-rata IQR setiap klaster di sebuah nilai K dan kolom keempat berisi nilai rata-rata IQR untuk masing-masing nilai yang divisualisasikan seperti pada Gambar 2.



Gambar 2. Komparasi Nilai Rata-Rata IQR

Journal of Big Data Analytic and Artificial Intelligence
Volume 8(1), Juni 2025, pp. 16-21

Sesuai kriteria, nilai IQR yang menunjukkan kondisi kohesifitas internal hasil klasterisasi adalah yang bernilai paling kecil, yaitu pada K=10 dengan nilai 0.4766. Hasil ini dapat ditegaskan dengan membandingkan boxplot dari keempat nilai K seperti ditunjukkan pada Gambar 3.



Gambar 3. Komparasi Boxplot Antar Nilai K

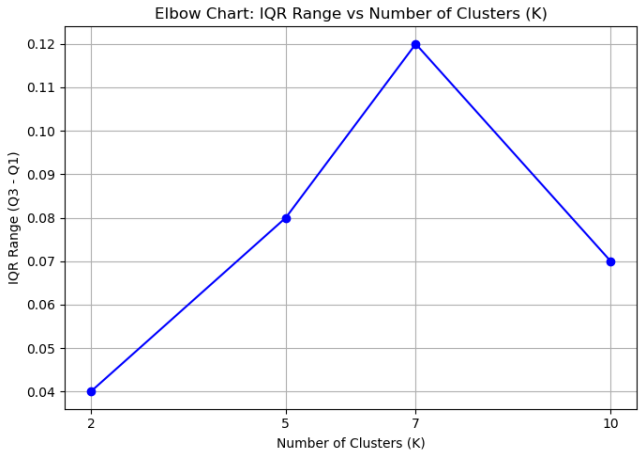
Terlihat dari boxplot masing-masing nilai K, terjadi penurunan tingkat kohesifitas seiring bertambahnya jumlah klaster yang dibentuk. Berdasarkan bentuk boxplot dihitung jangkauan nilai IQR untuk masing-masing nilai K seperti dirumuskan pada Tabel 4.

Tabel 4. Jangkauan IQR Masing-Masing K

K	Q1	Q3	Range
2	0.59	0.63	0.04
5	0.46	0.54	0.08
7	0.44	0.56	0.12
10	0.45	0.52	0.07

Nilai jangkauan IQR kemudian disusun untuk dikelompokkan menjadi 3 kelompok Small untuk IQR K=2, moderate untuk IQR K=5 dan K=10 dan Large untuk IQR K=7. Semakin kecil nilai rentang IQR maka semakin kuat kohesivitas pada jumlah klaster tersebut. Jumlah klaster K=2 memiliki kohesifitas paling baik dibandingkan ketiga nilai K lainnya, menunjukkan distribusi anggota klaster yang cukup rapat. Selain itu pada nilai K=2 tidak ditemukan outlier. Untuk K=5 memiliki kohesifitas tingkat menengah dengan distribusi data agak melebar. Kondisi serupa ditemukan pada K=7 dengan sedikit perbedaan pada tingkat variabilitas yang lebih tinggi dari K=5. Pada kedua nilai K ini tidak ditemukan data outlier. Pada nilai K=10, memiliki kohesifitas yang paling rendah dan ditemukan 3 data outlier.

Hasil identifikasi nilai IQR dan jumlah data outlier menunjukkan 3 nilai K berpotensi menjadi solusi akhir yaitu K=2, 5 dan 7. Untuk menentukan satu solusi nilai K sebagai solusi akhir, pada penelitian ini menggunakan konsep elbow method menggunakan nilai IQR seperti ditunjukkan pada Gambar 4.



Gambar 4. Grafik Elbow 4 Nilai IQR

Penggunaan elbow dengan nilai IQR untuk menemukan pola terkait stabilitas pembentukan klaster data antar nilai K. Nilai K terpilih akan bersifat konsisten dan reliabel. Dari 4 nilai K hasil evaluasi tahap 1, K=2 memiliki nilai IQR terkecil (0.04) menunjukkan kedua klaster yang terbentuk memiliki konsistensi dan tingkat variabilitas yang rendah. Artinya setiap data di masing-masing klaster memiliki kemiripan atau similaritas yang kuat. Perubahan nilai K menjadi 5 klaster membuat nilai IQR untuk K=5 menjadi 0.08 atau 2 kali lebih besar dari IQR K=2. Peningkatan nilai ini disebabkan oleh meningkatnya variabilitas antar data di lima klaster yang terbentuk. Titik "elbow" dari grafik IQR pada gambar terbentuk pada saat K=7 yang ditegaskan dengan menurunnya nilai IQR pada K=10. Peningkatan nilai IQR dari 2 ke 5 dan ke 7 menunjukkan klasterisasi semakin konsisten dengan sedikit penurunan tingkat stabilitas yang ditandai dengan nilai IQR tertinggi saat K=7. Penurunan IQR dari K=7 ke K=10 menunjukkan terjadinya perbaikan kualitas klaster yaitu kohesifitas makin menguat dari K=7 namun dari diagram boxplot diketahui terbentuk 3 data outlier. Sesuai dengan konsep penentuan nilai K dari elbow method, maka dari penelitian ini nilai K terbaik hasil 2 kali evaluasi terdapat pada K=5, dikarenakan setelah K=5 tidak terjadi perbaikan kohesifitas klaster. Pemilihan K=5 memiliki keuntungan dari K=2 dari sisi variasi (granularity) klaster yang akan diketahui jika kelima klaster yang terbentuk di proses lebih lanjut pada tahap interpretasi data.

IV. KESIMPULAN

Penelitian ini bertujuan untuk mengembangkan framework penentuan jumlah klaster yang tepat untuk mengatasi perbedaan hasil nilai K dari lima metrik validasi internal, Elbow, DBI, Silhouette Score, Dunn-Index dan Calinszki-Harabasz Index. Hasil penelitian menegaskan temuan dari penelitian sebelumnya bahwa terdapat perbedaan nilai K untuk masing-masing metrik validasi yang digunakan. Framework evaluasi yang diusulkan pada penelitian ini berhasil menyelesaikan permasalahan tersebut dalam 2 tahap. Tahap pertama memanfaatkan kelima metrik validasi internal sebagai filter untuk menentukan beberapa nilai K terbaik yaitu

K=2, 5, 7 dan 10. Tahap kedua mengevaluasi keempat hasil klasterisasi dari 4 nilai K tersebut dari perspektif kohesifitas internal klaster menggunakan analisis boxplot, IQR dan elbow. Evaluasi tahap kedua menunjukkan dengan dataset yang digunakan jumlah klaster yang tepat adalah K=5. Hasil ini menegaskan bahwa semakin banyak jumlah klaster belum tentu semakin baik kualitasnya. Stabilitas dan kohesifitas menjadi tujuan yang harus dipenuhi untuk memilih jumlah klaster yang optimal. Penelitian ini dibatasi pada penggunaan 1 dataset dan pada aspek kohesifitas internal klaster. Penelitian ini dapat dikembangkan lebih lanjut dengan membandingkan framework evaluasi yang digunakan untuk 2 atau lebih dataset. Selain itu, penelitian lebih lanjut dapat mengeksplorasi metrik-metrik validasi internal lainnya atau mengeksplorasi pengaruh algoritma klasterisasi terhadap kualitas klaster yang terbentuk.

REFERENSI

[1] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," *Inf Syst E-Bus Manage*, vol. 21, no. 3, pp. 527–570, Sep. 2023, doi: 10.1007/s10257-023-00640-4.

[2] R. W. B. S. Berahmana, F. A. Mohammed, and K. Chairuang, "Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods," *LKJITI*, vol. 11, no. 1, p. 32, Apr. 2020, doi: 10.24843/LKJITI.2020.v11.i01.p04.

[3] N. A. S. Z. Abidin, R. D. Avila, A. Hermatyar, and R. Rismayani, "Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Daerah Produksi Kakao," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 2, Art. no. 2, Aug. 2022, doi: 10.28932/jutisi.v8i2.4897.

[4] A. A. Aldino, D. Darwis, A. T. Prastowo, and C. Sujana, "Implementation of K-Means Algorithm for Clustering Corn Planting Feasibility Area in South Lampung Regency," *J. Phys.: Conf. Ser.*, vol. 1751, no. 1, p. 012038, Jan. 2021, doi: 10.1088/1742-6596/1751/1/012038.

[5] R. Fitriyanto and M. Ardi, "Feature Selection Comparative Performance for Unsupervised Learning on Categorical Dataset," *Journal of Computing and Information Technology*, vol. 22, no. 1, pp. 61–69, 2025.

[6] R. Fitriyanto and U. Syafiqoh, "Multilevel Modal Value Analysis for Interpreting Categorical K-Medoids Clusters Data," *techno*, vol. 21, no. 2, pp. 134–143, Sep. 2024, doi: 10.33480/techno.v21i2.5796.

[7] N. A. Maori and E. Evanita, "Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering," *Simetris J. Teknik Mesin, Elektro dan Ilmu Komput.*, vol. 14, no. 2, pp. 277–288, Nov. 2023, doi: 10.24176/simet.v14i2.9630.

[8] A. Winarta and W. J. Kurniawan, "Optimasi Cluster K-Means Menggunakan Metode Elbow Pada Data Pengguna Narkoba Dengan Pemrograman Python," *JTIK*, vol. 5, no. 1, pp. 113–119, Jan. 2021, doi: 10.59697/jtik.v5i1.593.

[9] A. Azzahra and A. W. Wijayanto, "Comparison of Agglomerative Hierarchical and K-Means in Grouping Provinces Based on Maternal Health Services," *SISTEMASI*, vol. 11, no. 2, p. 481, May 2022, doi: 10.32520/stmsi.v11i2.1829.

[10] S. Monalisa, "Klusterisasi Customer Lifetime Value dengan Model LRFM menggunakan Algoritma K-Means," *JTIIK*, vol. 5, no. 2, pp. 247–252, May 2018, doi: 10.25126/jtiik.201852690.

[11] A. M. Sikana and A. W. Wijayanto, "Analisis Perbandingan Pengelompokan Indeks Pembangunan Manusia Indonesia Tahun 2019 dengan Metode Partitioning dan Hierarchical Clustering," *jik*, vol. 14, no. 2, p. 66, Sep. 2021, doi: 10.24843/JIK.2021.v14.i02.p01.