Penentuan Cluster Optimal Menggunakan K-Harmonic Means Berdasarkan SSE dan DBI

Muhamad Raihan Ghafurun Halim[1], Muhammad Oktanto[2], R. Novarifqi Wicaksono[3], Rinci Kembang Hapsari[4], Iman Sapuguh[5]*

[1,2,3,4]Teknik Informatika, Institut Teknologi Adhi Tama Surabaya

[5]Universitas 45 Surabaya

**Penulis Korespondensi : rincikembang@itats.ac.id*

ABSTRACT

Customer segmentation is an important strategy for supporting marketing decision-making, especially for understanding consumer characteristics and behavior. One method widely used for customer segmentation is clustering. This study aims to determine the optimal number of clusters in the Mall Customers dataset using the K-Harmonic Means (KHM) algorithm and to evaluate the quality of the clustering results using the Sum of Squared Errors (SSE) and the Davies-Bouldin Index (DBI). The dataset used is secondary data on Mall Customers obtained from Kaggle, with a total of 200 data points and four attributes: Gender, Age, Annual Income, and Spending Score. The research stages include labeling categorical data, data normalization, clustering using K-Harmonic Means, and evaluating the number of clusters at K values of 3, 5, 7, 9, and 11. The results show that increasing the number of clusters decreases SSE and DBI values. The best value was obtained at K = 11, with an SSE of 15.10 and a DBI of 0.01, indicating the best level of cluster cohesion and separation among the cluster numbers. These results indicate that combining the K-Harmonic Means algorithm with SSE and DBI evaluation can determine a more optimal number of clusters for mall customer data. The findings of this study can serve as a basis for customer segmentation, supporting the development of more effective marketing strategies and data-driven decision-making.

Article History

Received : 15-07-2026
Revised : 29-07-2026
Accepted : 30-07-2026

Keywords

Davies-Bouldin Index.
Klastersasi
K-Harmonic Means
Segmentasi Pelanggan
Sum f Square Error

ABSTRAK

Segmentasi pelanggan merupakan salah satu strategi penting dalam mendukung pengambilan keputusan di bidang pemasaran, khususnya dalam memahami karakteristik dan perilaku konsumen. Salah satu metode yang banyak digunakan untuk segmentasi pelanggan adalah clustering. Penelitian ini bertujuan untuk menentukan jumlah cluster yang optimal pada dataset Mall Customers menggunakan algoritma K-Harmonic Means (KHM) serta mengevaluasi kualitas hasil clustering dengan Sum of Square Error (SSE) dan Davies-Bouldin Index (DBI). Dataset yang digunakan merupakan data sekunder Mall Customers yang diperoleh dari Kaggle dengan jumlah 200 data dan empat atribut, yaitu Gender, Age, Annual Income, dan Spending Score. Tahapan penelitian meliputi pelabelan data kategorik, normalisasi data, proses clustering menggunakan K-Harmonic Means, serta evaluasi jumlah cluster pada beberapa nilai K, yaitu 3, 5, 7, 9, dan 11. Hasil penelitian menunjukkan bahwa peningkatan jumlah cluster menghasilkan penurunan nilai SSE dan DBI. Nilai terbaik diperoleh pada K = 11 dengan SSE sebesar 15,10 dan DBI sebesar 0,01, yang menunjukkan tingkat kohesi dan separasi cluster yang paling baik dibandingkan jumlah cluster lainnya. Hasil tersebut menunjukkan bahwa kombinasi algoritma K-Harmonic Means dengan evaluasi menggunakan SSE dan DBI mampu menentukan jumlah cluster yang lebih optimal pada data pelanggan mall. Temuan penelitian ini dapat dimanfaatkan sebagai dasar dalam proses segmentasi pelanggan untuk mendukung penyusunan strategi pemasaran yang lebih efektif dan pengambilan keputusan berbasis data.

PENDAHULUAN

Perkembangan teknologi informasi telah mendorong organisasi untuk memanfaatkan data sebagai dasar dalam pengambilan keputusan. Pada sektor ritel, khususnya pusat perbelanjaan (mall), data pelanggan menjadi aset yang sangat berharga karena dapat digunakan untuk memahami karakteristik, preferensi, dan pola perilaku konsumen. Informasi tersebut sangat penting dalam

menyusun strategi pemasaran, meningkatkan kualitas layanan, serta membangun loyalitas pelanggan melalui pendekatan yang lebih terarah dan personal.

Salah satu teknik yang banyak digunakan untuk menganalisis karakteristik pelanggan adalah clustering, yaitu proses mengelompokkan data berdasarkan tingkat kemiripan karakteristik tanpa menggunakan label kelas. Melalui proses clustering, pelanggan dapat dikelompokkan ke dalam beberapa segmen yang memiliki karakteristik serupa, sehingga perusahaan dapat merancang strategi promosi, penawaran produk, maupun program loyalitas yang sesuai dengan kebutuhan masing-masing kelompok pelanggan.

Algoritma K-Means merupakan metode clustering yang paling banyak digunakan karena sederhana dan memiliki waktu komputasi yang relatif cepat. Namun demikian, K-Means memiliki kelemahan, yaitu sangat sensitif terhadap pemilihan centroid awal sehingga hasil pengelompokan dapat berbeda pada setiap proses inisialisasi. Selain itu, algoritma ini rentan terjebak pada solusi optimum lokal sehingga kualitas cluster yang dihasilkan belum tentu optimal, terutama pada data yang memiliki distribusi tidak merata.

Sebagai alternatif, K-Harmonic Means (KHM) dikembangkan untuk mengurangi sensitivitas terhadap pemilihan centroid awal dengan memanfaatkan fungsi rata-rata harmonik (harmonic average) dalam proses pembaruan centroid. Pendekatan ini memungkinkan setiap data memberikan kontribusi terhadap seluruh centroid selama proses iterasi sehingga menghasilkan proses konvergensi yang lebih stabil dibandingkan metode K-Means. Oleh karena itu, KHM menjadi salah satu metode clustering yang berpotensi menghasilkan segmentasi pelanggan yang lebih baik pada berbagai karakteristik data.

Meskipun demikian, penerapan algoritma clustering tidak hanya bergantung pada proses pembentukan cluster, tetapi juga pada kemampuan dalam menentukan jumlah cluster yang optimal. Penentuan jumlah cluster merupakan salah satu permasalahan utama dalam analisis clustering karena tidak adanya informasi awal mengenai jumlah kelompok yang sebenarnya terdapat pada data. Pemilihan jumlah cluster yang kurang tepat dapat menyebabkan hasil segmentasi menjadi kurang representatif dan sulit diinterpretasikan.

Berbagai metode validasi cluster telah dikembangkan untuk mengevaluasi kualitas hasil clustering. Dua metode yang paling banyak digunakan adalah Sum of Square Error (SSE) dan Davies-Bouldin Index (DBI). SSE mengukur tingkat homogenitas anggota dalam setiap cluster melalui jumlah kuadrat jarak antara data dengan centroid, sehingga nilai SSE yang lebih kecil menunjukkan cluster yang semakin kompak. Sementara itu, DBI mengevaluasi keseimbangan antara kohesi dalam cluster dan separasi antarcluster, di mana nilai DBI yang semakin kecil menunjukkan kualitas cluster yang semakin baik. Beberapa penelitian sebelumnya telah menerapkan K-Harmonic Means pada berbagai bidang, seperti pengelompokan wilayah, data kesehatan, maupun segmentasi pelanggan. Namun, sebagian besar penelitian hanya berfokus pada penerapan algoritma clustering tanpa melakukan analisis yang mendalam terhadap pemilihan jumlah cluster optimal menggunakan lebih dari satu indeks validasi. Selain itu, pembahasan mengenai konsistensi hasil evaluasi antara SSE dan DBI pada dataset pelanggan masih relatif terbatas. Kondisi tersebut menunjukkan adanya kebutuhan untuk melakukan evaluasi kualitas clustering menggunakan beberapa indeks validasi sehingga jumlah cluster yang diperoleh memiliki dasar yang lebih kuat.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk menerapkan algoritma K-Harmonic Means pada dataset Mall Customers serta mengevaluasi kualitas hasil clustering menggunakan Sum of Square Error (SSE) dan Davies-Bouldin Index (DBI) pada beberapa variasi jumlah cluster. Hasil penelitian diharapkan dapat memberikan gambaran mengenai jumlah cluster yang paling representatif dalam menggambarkan karakteristik pelanggan serta menjadi referensi

ISSN: 2774-4345, DOI: 10.31284/j.kernel.2026.v7i1.9260

dalam penerapan metode clustering untuk mendukung pengambilan keputusan berbasis data pada sektor ritel.

METODE

1. Pengumpulan Data

   Penggunaan jenis data mall dipilih berdasarkan penelitian yang menggunakan data mall customers dalam proses clustering K-Harmonic Means. Penelitian ini menggunakan variabel serupa namun ada modifikasi dalam penentuan value standar.

2. Pre-Processing Data

   2.1 Label Data

   Memberikan label pada data berdasarkan karakteristik tertentu. Misalnya, dalam data mall customers, diberikan label "genre" berdasarkan nilai numerik.

   2.2 Normalisasi

   Tahapan prepocessing dalam data mining, sangat diperlukan untuk mendapatkan hasil yang optimal. Dengan adanya prepocessing menjadikan penerapan algoritma menjadi lebih efisien. Salah satu proses preprosesing data adalah normalisasi. Tujuan dari normalisasi data dalam dataset adalah untuk membentuk data dalam posisi nilai dengan rentang yang sama.

3. Proses Perhitungan K-Harmonic Means

   K-Harmonic Means (KHM) pertama kali diperkenalkan oleh Zhang, Hsu, dan Dayal (1999) dari HP Laboratories Palo Alto yang kemudian dikembangkan oleh Hammerly dan Elkan pada tahun 2002. Tujuan pengembangan metode KHM adalah untuk menangani masalah utama dalam K-Means yang hasil clusteringnya sangat sensitif dengan inisialisasi data yang dijadikan sebagai centroid awal. Hasil yang sering berbeda (lokal optima) dari proses clusteringnya (pada set data yang sama) disebabkan oleh inisialisasi centroid yang berbeda.

   Algoritma clustering dengan K-Harmonic Means sebagai berikut :

   1. Tentukan Nilai K sebagai jumlah kelompok / cluster.
   2. Inisialisasi posisi centroid awal dimana C = $\{c_j \mid j = 1, ..., K\}$ sebanyak K centroid secara acak dari data yang ada.
   3. Hitung Jarak data terhadap masing-masing centroid. Misalnya menggunakan rumus jarak euclidean seperti persamaan berikut :

      $$d_{i,j} = |x_i - c_j|_2 = \sqrt[2]{(x_i - c_j)^2}$$

      Dimana X = $\{ x_i \mid i = 1 ..... N \}$, N adalah jumlah data yang akan diklaster dengan metode KHM.
   4. Cari jarak terdekat $d_{i,min}$ dan masukkan X kedalam cluster sesuai dengan kelompok/centroid tersebut.
   5. Cari centroid baru sebanyak K dengan persamaan KHM seperti berikut :

$$m_{i,k} = \frac{\sum_{i=1}^{N} \frac{1}{d_{i,k}^{3}\left(\sum_{j=1}^{K} \frac{1}{d_{i,j}^{2}}\right)^{2}} \cdot x_i}{\sum_{i=1}^{N} \frac{1}{d_{i,k}^{3}\left(\sum_{j=1}^{K} \frac{1}{d_{i,j}^{2}}\right)^{2}}}$$

Dimana : $m_{i,k}$ = centroid baru metode KHM
N = data yang dicluster
$d_{i,k}$ = jarak antara i ke k
$d_{i,j}$ = jarak antara i ke j
x_i = data ke-i

Catatan : $d_{i,min} = 0$ maka vektor m_k diset menjadi 0.

6. Lakukan langkah nomor 2 – 4 hingga posisi anggota cluster tidak berubah.

4. Evaluasi Cluster Menggunakan SSE dan DBI

4.1 Metode Elbow dan Sum of Square Error (SSE)

Metode Elbow bekerja membandingkan nilai atau persentase dari sejumlah k yang telah diujikan dan membentuk siku pada suatu titik. Nilai k pada kombinasi siku dengan K-Harmonic Means adalah grafik hubungan cluster dengan penurunan error. Peningkatan nilai k menyebabkan grafik menurun perlahan sampai hasil nilai k stabil. Jumlah cluster k yang dihasilkan dari pengujian dengan K-Harmonic Means, dievaluasi menggunakan teknik SSE. Semakin banyak jumlah k yang diujian, maka nilai SSE semakin kecil. SSE merupakan cara dalam melakukan validasi cluster melalui jumlah kuadrat setiap anggota cluster menuju pusatnya. Semakin jauh jarak yang membentuk titik siku, maka jumlah cluster tersebut menjadi yang paling optimal.

Rumus SSE adalah sebagai berikut :

$$SSE = \sum_{K=1}^{K} \sum_{x_i \in S_K} \| x_i - c_k \|_2^2$$

4.2 Davies Bouldin Index (DBI)

Davies Bouldin Index (DBI) adalah metric untuk mengevaluasi atau mempertimbangkan hasil algoritma clustering. Pertama kali diperkenalkan oleh David L. Davies dan Donald W. Bouldin pada tahun 1979. Dengan menggunakan DBI suatu cluster akan dianggap memiliki skema clustering yang optimal adalah yang memiliki DBI minimal.

Langkah – langkah perhitungan Davies Bouldin Index ialah sebagai berikut :

1. Sum Of Square Within-Cluster (SSW)

Untuk mengetahui kohesi dalam sebuah cluster ke-I salah satunya adalah dengan menghitung nilai dari Sum Of Square Within-Cluster (SSW). Dengan rumus sebagai berikut :

$$SSW_i = \frac{1}{mi} \sum_{j=i}^{mi} d(X_j, C_j)$$

Dimana : mi = jumlah data dalam cluster ke-i
ci = centroid cluster ke-i
$d(X_j, C_j)$ = jarak setiap data ke centroid i yang dihitung menggunakan jarak euclidean.

2. Sum Of Square Between-Cluster (SSB)

Perhitungan Sum Of Square Between-Cluster (SSB) bertujuan untuk mengetahui separasi atau jarak antar cluster. Dengan rumus perhitungan sebagai berikut :

$SSB_{ij} = d\ (X_i, X_j)$

Dimana : d (X_i, X_j) = jarak antara data ke i dengan data ke j di cluster lain

3. Ratio

Perhitungan ratio ($R_{i,j}$) ini bertujuan untuk mengetahui nilai perbandingan antara cluster ke-i dan cluster ke-j untuk menghitung nilai ratio yang dimiliki oleh masing-masing cluster. Untuk menentukan nilai ratio dengan rumus sebagai berikut:

$$R_{ij\ldots\ldots,n} = \frac{SSW_i + SSW_j + \cdots + SSW_n}{SSB_{i,j} + \cdots + SSB_{ni,nj}}$$

Dimana : SSW_i = Sum Of Square Within-Cluster pada centroid i
$SSB_{i,j}$ = Sum Of Square Between Cluster data ke i dengan j pada cluster yang berbeda

4. Davies Bouldin Index (DBI)

Nilai ratio yang diperoleh, digunakan untuk mencari nilai DBI dengan menggunakan perhitungan sebagai berikut :

$$DBI = \frac{1}{k} \sum_{i=1}^{k} max\ _{i \neq j}(Ri, j, \ldots k)$$

Dimana, $R_{i,j}$ merupakan ratio dari nilai SSW dan SSB melalui perhitungan rumusnya, maka dapat diketahui k adalah jumlah cluster. Dari perhitungan Davies Bouldin Index (DBI) dapat disimpulkan bahwa jika semakin kecil nilai Davies Bouldin Index (DBI) yang diperoleh (non negatif >= 0) maka cluster tersebut semakin baik.

HASIL DAN PEMBAHASAN

1. Pengumpulan Data

Dalam pengumpulan dataset yang kami pilih diperoleh dari data sekunder (pihak lain), yaitu bersumber pada data yang diambil dari kaggle.com berdasarkan data mall customers sebanyak 200 data untuk dilakukan uji coba terhadap algoritma K-Harmonic Means dapat dilihat pada tabel 1.

Tabel 1. Data Awal

CustomerID	Genre	Age	Annual_Income_(k$)	Spending_Score
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72

ISSN: 2774-4345, DOI: 10.31284/j.kernel.2026.v7i1.9260

CustomerID	Genre	Age	Annual_Income_(k$)	Spending_Score
.....	.....	.....	.....	.....
200	Male	30	137	83

2. Label dan Normalisasi Data

2.1 Label

Proses selanjutnya adalah proses labeling data dari mall customers menjadi dataset numerik untuk dianalisis. Variabel clustering dalam K-Harmonic Means untuk proses analisis cluster dapat melakukan pemetaan data mall customers dari Gender dengan nilai standar 1 dan 2, sesuai dengan kriteria di bawah ini :

Contoh :

Male = 1 dan Female = 2

Tabel 2. Perubahan Data Menjadi Numerik

CustomerID	Genre	Inisialisasi
1	Male	1
2	Male	1
3	Female	2
4	Female	2
5	Female	2
6	Female	2
7	Female	2
8	Female	2
9	Male	1
10	Female	2
.....	.....	.....
200	Male	1

Label data yang dipakai dalam analisis K-Harmonic Means clustering dapat dilihat pada tabel 2.

2.2 Normalisasi

Kemudian ialah proses normalisasi data dari yang sudah ditransformasikan sebelumnya. Tahapan ini sangat diperlukan untuk mendapatkan hasil yang optimal dan menjadi lebih efisien. Berikut data yang sudah proses normalisasi dapat dilihat pada tabel 3.

Tabel 3. Normalisasi Data

CustomerID	Genre	Age	Annual_Income_(k$)	Spending_Score
1	0.00	0.02	0.00	0.39
2	0.00	0.06	0.00	0.82
3	1.00	0.04	0.01	0.05
4	1.00	0.10	0.01	0.78
5	1.00	0.25	0.02	0.40
6	1.00	0.08	0.02	0.77
7	1.00	0.33	0.02	0.05
8	1.00	0.10	0.02	0.95

ISSN: 2774-4345, DOI: 10.31284/j.kernel.2026.v7i1.9260

CustomerID	Genre	Age	Annual_Income_(k$)	Spending_Score
9	0.00	0.88	0.03	0.02
10	1.00	0.23	0.03	0.72
.....	.....	.....	.....	.....
200	0.00	0.23	1.00	0.84

3. Proses Perhitungan K-Harmonic Means

Contoh perhitungan KHM menggunakan dataset yang terdiri dari 200 data yang memiliki 4 atribut yaitu Genre, Age, Annual_Income_(k$) dan Spending_Score, ditunjukkan pada Tabel 4.

Tabel 4. Dataset Yang Sudah Normalisasikan

CustomerID	Genre	Age	Annual_Income_(k$)	Spending_Score
1	0.00	0.02	0.00	0.39
2	0.00	0.06	0.00	0.82
3	1.00	0.04	0.01	0.05
4	1.00	0.10	0.01	0.78
5	1.00	0.25	0.02	0.40
6	1.00	0.08	0.02	0.77
7	1.00	0.33	0.02	0.05
8	1.00	0.10	0.02	0.95
9	0.00	0.88	0.03	0.02
10	1.00	0.23	0.03	0.72
.....	.....	.....	.....	.....
200	0.00	0.23	1.00	0.84

Langkah pertama, melakukan perhitungan menggunakan metode K-Harmonic Means (KHM) adalah menentukan nilai K sebagai jumlah kelompok/cluster. Nilai K pada contoh perhitungan ini telah ditentukan sebanyak 3.

Langkah kedua, melakukan inisialisasi posisi centroid awal dimana $C = \{c_j \mid j = 1, ..., K\}$ sebanyak K centroid secara acak dari data yang ada, yaitu :

Cluster	Genre	Age	Annual_Income_(k$)	Spending_Score
C1	0.00	0.02	0.00	0.39
C2	1.00	0.08	0.02	0.77
C3	0.00	0.94	0.03	0.13

Iterasi 1

Ketiga, menghitung jarak data terhadap masing-masing centroid menggunakan rumus jarak euclidean, yaitu :

$\mathbf{d_{1,1}} = \sqrt{(0.00-0.00)^2 + (0.02-0.02)^2 + (0.00-0.00)^2 + (0.39-0.39)^2} = 0.00$

$\mathbf{d_{2,1}} = \sqrt{(0.00-0.00)^2 + (0.06-0.02)^2 + (0.00-0.00)^2 + (0.82-0.39)^2} = 0.43$

$\mathbf{d_{3,1}} = \sqrt{(1.00-0.00)^2 + (0.04-0.02)^2 + (0.01-0.00)^2 + (0.05-0.39)^2} = 1.06$

$\mathbf{d_{4,1}} = \sqrt{(1.00-0.00)^2+(0.10-0.02)^2+(0.01-0.00)^2+(0.78-0.39)^2} = 1.08$
$\mathbf{d_{5,1}} = \sqrt{(1.00-0.00)^2+(0.25-0.02)^2+(0.02-0.00)^2+(0.40-0.39)^2} = 1.03$
.....
$\mathbf{d_{200,1}} = \sqrt{(0.00-0.00)^2+(0.23-0.02)^2+(1.00-0.00)^2+(0.84-0.39)^2} = 1.12$

$\mathbf{d_{1,2}} = \sqrt{(0.00-1.00)^2+(0.02-0.08)^2+(0.00-0.02)^2+(0.39-0.77)^2} = 1.07$
$\mathbf{d_{2,2}} = \sqrt{(0.00-1.00)^2+(0.06-0.08)^2+(0.00-0.02)^2+(0.82-0.77)^2} = 1.00$
$\mathbf{d_{3,2}} = \sqrt{(1.00-1.00)^2+(0.04-0.08)^2+(0.01-0.02)^2+(0.05-0.77)^2} = 0.72$
$\mathbf{d_{4,2}} = \sqrt{(1.00-1.00)^2+(0.10-0.08)^2+(0.01-0.02)^2+(0.78-0.77)^2} = 0.02$
$\mathbf{d_{5,2}} = \sqrt{(1.00-1.00)^2+(0.25-0.08)^2+(0.02-0.02)^2+(0.40-0.77)^2} = 0.41$
.....
$\mathbf{d_{200,2}} = \sqrt{(0.00-1.00)^2+(0.23-0.08)^2+(1.00-0.02)^2+(0.84-0.77)^2} = 1.41$

$\mathbf{d_{1,3}} = \sqrt{(0.00-0.00)^2+(0.02-0.94)^2+(0.00-0.03)^2+(0.39-0.13)^2} = 0.96$
$\mathbf{d_{2,3}} = \sqrt{(0.00-0.00)^2+(0.06-0.94)^2+(0.00-0.03)^2+(0.82-0.13)^2} = 1.12$
$\mathbf{d_{3,3}} = \sqrt{(1.00-0.00)^2+(0.04-0.94)^2+(0.01-0.03)^2+(0.05-0.13)^2} = 1.35$
$\mathbf{d_{4,3}} = \sqrt{(1.00-0.00)^2+(0.10-0.94)^2+(0.01-0.03)^2+(0.78-0.13)^2} = 1.46$
$\mathbf{d_{5,3}} = \sqrt{(1.00-0.00)^2+(0.25-0.94)^2+(0.02-0.03)^2+(0.40-0.13)^2} = 1.24$
.....
$\mathbf{d_{200,3}} = \sqrt{(0.00-0.00)^2+(0.23-0.94)^2+(1.00-0.03)^2+(0.84-0.13)^2} = 1.40$

Langkah keempat, mencari jarak terdekat $d_{i,min}$ dan masukkan X kedalam cluster sesuai dengan kelompok/centroid tersebut.

Tabel 5. Hasil Perhitungan Jarak dan Pengelompokan Data

CustomerID	Jarak C1	Jarak C2	Jarak C3	Jarak Min	Cluster
1	0.00	1.07	0.96	0.00	C1
2	0.43	1.00	1.12	0.43	C1
3	1.06	0.72	1.35	0.72	C2
4	1.08	0.02	1.46	0.02	C2
5	1.03	0.41	1.24	0.41	C2
6	1.07	0.00	1.47	0.00	C2
7	1.10	0.76	1.17	0.76	C2
8	1.15	0.18	1.54	0.18	C2
9	0.94	1.48	0.13	0.13	C3
10	1.07	0.16	1.36	0.16	C2
.....	.....	.....	.....	.....	.....
200	1.12	1.41	1.40	1.12	C1

Kelima, mencari centroid baru sebanyak K sesuai dengan persamaan KHM :

$$M_{1,1} = \frac{0 + \left(\frac{1}{0.43^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}.0.00\right) + \left(\frac{1}{1.06^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}.1.00\right) + \cdots + \left(\frac{1}{1.12^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}.0.00\right)}{0 + \left(\frac{1}{0.43^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}\right) + \left(\frac{1}{1.06^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}\right) + \cdots + \left(\frac{1}{1.12^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}\right)} = 0.25$$

$$M_{1,2} = \frac{0 + \left(\frac{1}{0.43^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}.0.06\right) + \left(\frac{1}{1.06^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}.0.04\right) + \cdots + \left(\frac{1}{1.12^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}.0.23\right)}{0 + \left(\frac{1}{0.43^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}\right) + \left(\frac{1}{1.06^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}\right) + \cdots + \left(\frac{1}{1.12^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}\right)} = 0.32$$

$$M_{1,3} = \frac{0 + \left(\frac{1}{0.43^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}.0.00\right) + \left(\frac{1}{1.06^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}.0.01\right) + \cdots + \left(\frac{1}{1.12^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}.1.00\right)}{0 + \left(\frac{1}{0.43^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}\right) + \left(\frac{1}{1.06^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}\right) + \cdots + \left(\frac{1}{1.12^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}\right)} = 0.42$$

$$M_{1,4} = \frac{0 + \left(\frac{1}{0.43^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}.0.82\right) + \left(\frac{1}{1.06^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}.0.05\right) + \cdots + \left(\frac{1}{1.12^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}.0.84\right)}{0 + \left(\frac{1}{0.43^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}\right) + \left(\frac{1}{1.06^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}\right) + \cdots + \left(\frac{1}{1.12^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}\right)} = 0.52$$

$$M_{2,1} = \frac{0 + \left(\frac{1}{1.00^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}.0.00\right) + \left(\frac{1}{0.72^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}.1.00\right) + \cdots + \left(\frac{1}{1.41^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}.0.00\right)}{0 + \left(\frac{1}{1.00^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}\right) + \left(\frac{1}{0.72^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}\right) + \cdots + \left(\frac{1}{1.41^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}\right)} = 0.92$$

$$M_{2,2} = \frac{0 + \left(\frac{1}{1.00^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}.0.06\right) + \left(\frac{1}{0.72^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}.0.04\right) + \cdots + \left(\frac{1}{1.41^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}.0.23\right)}{0 + \left(\frac{1}{1.00^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}\right) + \left(\frac{1}{0.72^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}\right) + \cdots + \left(\frac{1}{1.41^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}\right)} = 0.38$$

$$M_{2,3} = \frac{0 + \left(\frac{1}{1.00^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}.0.00\right) + \left(\frac{1}{0.72^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}.0.01\right) + \cdots + \left(\frac{1}{1.41^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}.1.00\right)}{0 + \left(\frac{1}{1.00^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}\right) + \left(\frac{1}{0.72^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}\right) + \cdots + \left(\frac{1}{1.41^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}\right)} = 0.40$$

$$M_{2,4} = \frac{0 + \left(\frac{1}{1.00^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}.0.82\right) + \left(\frac{1}{0.72^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}.0.05\right) + \cdots + \left(\frac{1}{1.41^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}.0.84\right)}{0 + \left(\frac{1}{1.00^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}\right) + \left(\frac{1}{0.72^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}\right) + \cdots + \left(\frac{1}{1.41^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}\right)} = 0.54$$

$$M_{3,1} = \frac{0 + \left(\frac{1}{1.12^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}.0.00\right) + \left(\frac{1}{1.35^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}.1.00\right) + \cdots + \left(\frac{1}{1.40^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}.0.00\right)}{0 + \left(\frac{1}{1.12^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}\right) + \left(\frac{1}{1.35^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}\right) + \cdots + \left(\frac{1}{1.40^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}\right)} = 0.33$$

$$M_{3,2} = \frac{0 + \left(\frac{1}{1.12^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}.0.06\right) + \left(\frac{1}{1.35^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}.0.04\right) + \cdots + \left(\frac{1}{1.40^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}.0.23\right)}{0 + \left(\frac{1}{1.12^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}\right) + \left(\frac{1}{1.35^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}\right) + \cdots + \left(\frac{1}{1.40^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}\right)} = 0.60$$

$$M_{3,3} = \frac{0 + \left(\frac{1}{1.12^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}.0.00\right) + \left(\frac{1}{1.35^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}.0.01\right) + \cdots + \left(\frac{1}{1.40^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}.1.00\right)}{0 + \left(\frac{1}{1.12^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}\right) + \left(\frac{1}{1.35^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}\right) + \cdots + \left(\frac{1}{1.40^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}\right)} = 0.42$$

$$M_{3,4} = \frac{0 + \left(\frac{1}{1.12^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}\cdot 0.82\right) + \left(\frac{1}{1.35^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}\cdot 0.05\right) + \cdots + \left(\frac{1}{1.40^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}\cdot 0.84\right)}{0 + \left(\frac{1}{1.12^3\left(\frac{1}{0.43^2}+\frac{1}{1.00^2}+\frac{1}{1.12^2}\right)^2}\right) + \left(\frac{1}{1.35^3\left(\frac{1}{1.06^2}+\frac{1}{0.72^2}+\frac{1}{1.35^2}\right)^2}\right) + \cdots + \left(\frac{1}{1.40^3\left(\frac{1}{1.12^2}+\frac{1}{1.41^2}+\frac{1}{1.40^2}\right)^2}\right)} = 0.40$$

Tabel 6. Centroid Baru

Cluster	Genre	Age	Annual_Income_(k$)	Spending_Score
C1	0.25	0.32	0.42	0.52
C2	0.92	0.38	0.40	0.54
C3	0.33	0.60	0.42	0.40

Lakukan langkah nomor 2 – 4 hingga posisi anggota cluster tidak berubah.

Iterasi 5

Menghitung jarak data terhadap centroid baru menggunakan rumus jarak euclidean :

Tabel 7. Hasil Perhitungan Jarak dan Pengelompokan Data dengan Centroid Baru Iterasi 5

CustomerID	Jarak C1	Jarak C2	Jarak C3	Jarak Min	Cluster
1	0.55	1.09	0.66	0.55	C1
2	0.58	1.11	0.75	0.58	C1
3	1.08	0.68	1.05	0.68	C2
4	0.98	0.52	1.02	0.52	C2
5	0.93	0.39	0.89	0.39	C2
6	0.98	0.52	1.02	0.52	C2
7	1.04	0.59	0.95	0.59	C2
8	1.03	0.62	1.09	0.62	C2
9	0.85	1.24	0.68	0.68	C3
10	0.93	0.42	0.95	0.42	C2
.....	.....	.....	.....	.....	.....
200	0.71	1.20	0.83	0.71	C1

Mencari centroid baru sebanyak K menggunakan metode KHM :

Tabel 8. Centroid Baru Iterasi 5

Cluster	Genre	Age	Annual_Income_(k$)	Spending_Score
C1	0.17	0.34	0.39	0.54
C2	0.96	0.38	0.36	0.52
C3	0.22	0.51	0.38	0.41

Cluster berhasil ditemukan pada iterasi ke-5

4. Evaluasi Cluster Menggunakan SSE dan DBI

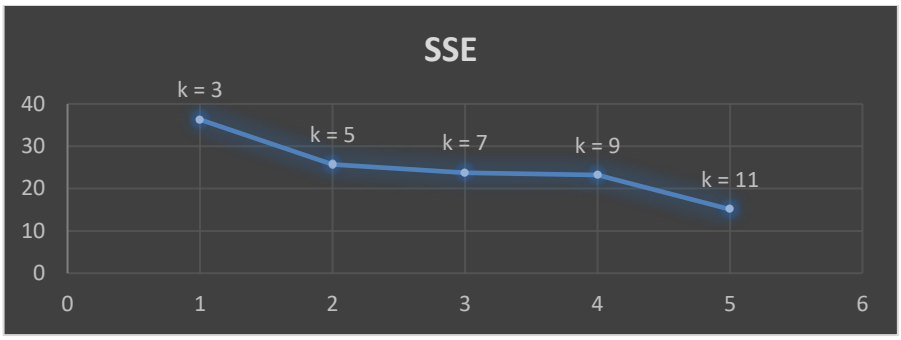
Penelitian ini menggunakan 5 kali perubahan jumlah cluster. Untuk menentukan jumlah cluster yang paling optimal, dilakukan evaluasi cluster dengan Sum Of Square Error (SSE) dan Davies Bouldin Index (DBI). Setiap pengujian menghasilkan nilai jarak yang

saling dibandingkan sesuai ketentuan. Adapun nilai SSE dan DBI evaluasi ditampilkan oleh tabel 9.

Tabel 9. Nilai Evaluasi Cluster SSE dan DBI

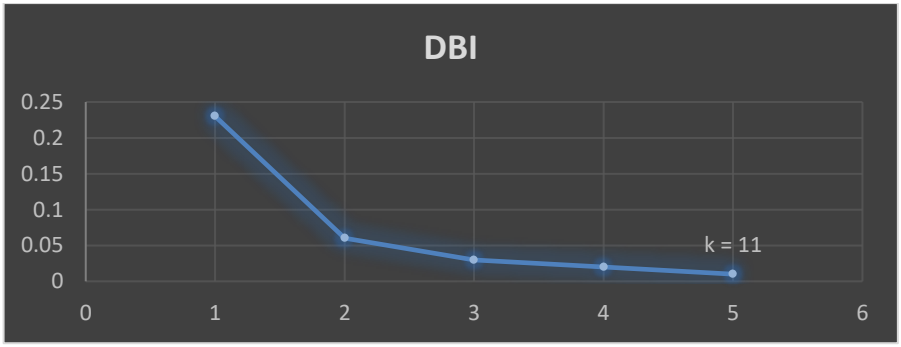
Jumlah k	Nilai SSE	Nilai DBI
k = 3	36.29	0.23
k = 5	25.68	0.06
k = 7	23.73	0.03
k = 9	23.20	0.02
k = 11	15.10	0.01

Nilai SSE didasarkan pada besarnya error yang dihasilkan untuk dapat membentuk siku dan menjadi jarak terbesar. Pada Tabel 9, jarak terbesar terdapat pada k = 11. Grafik hasil perhitungan SSE dapat dilihat pada Gambar 1.



Gambar 1. Cluster Optimal SSE

Sedangkan untuk teknik DBI, semakin kecil nilai index yang terbentuk, semakin optimal cluster tersebut. DBI terkecil juga terdapat pada k = 11. Grafik DBI dapat ditampilkan seperti pada Gambar 2.



Gambar 2. Cluster Optimal DBI

ISSN: 2774-4345, DOI: 10.31284/j.kernel.2026.v7i1.9260

KESIMPULAN

Evaluasi Cluster SSE dan DBI mampu menentukan jumlah cluster yang paling optimal dari serangkaian k yang diujikan pada data uji. Berdasarkan hasil pengujian, SSE dan DBI membentuk jumlah cluster optimal pada k = 11. Namun performa DBI lebih baik dalam meminimalkan jarak antara cluster sehingga informasi yang terbentuk lebih lengkap dan terpola. Penelitian ini berbeda karena komparasi disertai dengan analisa hasil setiap k yang diujikan. Komparasi performa SSE dan DBI masih dapat dikembangkan melalui sejumlah data uji yang dibedakan dari segi jumlah maupun penggunaan pengukuran jarak dalam algoritma K-Harmonic Means. Hasil penelitian ini diharapkan dapat memberikan pandangan baru dalam penggunaan evaluasi cluster sehingga mempermudah pengguna untuk menentukan teknik yang sesuai dengan data yang ada dalam menghasilkan jumlah cluster yang paling optimal dan informasi yang benar.

DAFTAR PUSTAKA

[1] T. Anwar, T. Siswantining, D. Sarwinda, S. M. Soemartojo, dan A. Bustamam, "A study on missing values imputation using K-Harmonic means algorithm: Mixed datasets," dalam AIP Conference Proceedings, American Institute of Physics Inc., Des 2019. doi: 10.1063/1.5141651.

[2] "STUDI KOMPARASI METODE KLASTERISASI DATA K-MEANS DAN K-HARMONIC MEANS."

[3] Y. Syahra, R. Imanta Ginting, M. Yetri, dan S. Triguna Dharma, "Implementasi Data Mining Untuk Penyusunan Tata Letak Data Obat-Obatan Dengan Menggunakan Algoritma K-Harmonic Means Pada Apotek Inti Fada Sidamanik." [Daring]. Tersedia pada: http://www.jurnal.umb.ac.id/index.php/JTIS

[4] M. Nishom dan M. Y. Fathoni, "Implementasi Pendekatan Rule-Of-Thumb untuk Optimasi Algoritma K-Means Clustering," vol. 03, no. 02, 2018.

[5] A. Izzah dan N. Hayatin, "Seminar Nasional Matematika dan Aplikasinya 2013 Imputasi Missing data Menggunakan Algoritma Pengelompokan Data K-Harmonic Means."

[6] A. Salsabila, T. Widiharih, D. Statistika, dan F. Sains dan Matematika, "METODE K-HARMONIC MEANS CLUSTERING DENGAN VALIDASI SILHOUETTE COEFFICIENT (Studi Kasus : Empat Faktor Utama Penyebab Stunting 34 Provinsi di Indonesia Tahun 2018)," vol. 11, no. 1, hlm. 11–20, 2022.

[7] M. Awaludin, "PENERAPAN ALGORITMA K-MEANS CLUSTERING PADA K-HARMONIC MEANS UNTUK SCHEDULE PREVENTIVE MAINTENANCE SERVICE."

[8] D. I. Yunistya, R. Goejantoro, F. Deny, dan T. Amijaya, "The Application Of K-Harmonic Means Method In District/City Grouping (Case Study: Poverty in Kalimantan Island in 2020) Penerapan Metode K-Harmonic Means Dalam Pengelompokan Kabupaten/Kota (Studi Kasus: Kemiskinan di Pulau Kalimantan Tahun 2020)," vol. 19, no. 1, hlm. 51–64, 2022, doi: 10.20956/j.v19i1.21116.

[9] "STUDI KOMPARASI METODE KLASTERISASI DATA K-MEANS DAN K-HARMONIC MEANS."

[10] Z. Güngör dan A. Ünler, "K-Harmonic means data clustering with tabu-search method," Appl Math Model, vol. 32, no. 6, hlm. 1115–1125, Jun 2008, doi: 10.1016/j.apm.2007.03.011.

[11] "BAB II LANDASAN TEORI 2.1 Data Mining."

[12] L. Petra Refialy, H. Maitimu, dan M. Soyano Pesulima, "Perbaikan Kinerja Clustering K-Means pada Data Ekonomi Nelayan dengan Perhitungan Sum of Square Error (SSE) dan Optimasi nilai K cluster," 2021.

[13] N. T. Hartanti, "Metode Elbow dan K-Means Guna Mengukur Kesiapan Siswa SMK Dalam Ujian Nasional," Jurnal Nasional Teknologi dan Sistem Informasi, vol. 6, no. 2, hlm. 82–89, Agu 2020, doi: 10.25077/teknosi.v6i2.2020.82-89.

[14] D. Jollyta, S. Efendi, M. Zarlis, dan H. Mawengkang, "Prosiding Seminar Nasional Riset Information Science (SENARIS) Optimasi Cluster Pada Data Stunting: Teknik Evaluasi Cluster Sum of Square Error dan Davies Bouldin Index," 2019.

[15] M. -Normalisasi Data Untuk Efisiensi K-Means Pada Pengelompokan Wila-yah Berpotensi Kebakaran Hutan Dan Lahan Berdasarkan Sebaran Titik Panas, A. Harmain, H. Kurniawan, D. Maulina, dan M. Universitas Amikom Yogyakarta, "DATA NORMALIZATION FOR K-MEANS EFFICIENCY ON GROUPS OF AREAS WITH POTENTIAL FORES AND /LAND FIRE BASED ON HEAT SPOTS DISTRIBUTION."