PROSIDING Seminar Nasional SISFOTEK ISSN 2597-3584 (Media Online)
Sistem Informasi dan Teknologi
website : www.seminar.iaii.or.id | email : seminar@iaii.or.id

Uji Kinerja K-Means Clustering Menggunakan Davies-Bouldin Index Pada Pengelompokan Data Prestasi Siswa

Imam T. Umagapi[1], Basirung Umaternate[2], Hazriani[3], Yuyun[4]
[1,2,3,4]Sistem Komputer, Program Pasca Sarjana Universitas Handayani,
[1,2]Badan Kepegawaian Daerah Kabupaten Pulau Morotai
[4]Badan Riset dan Inovasi (BRIN)
barqoui21@gmail.com

Abstract

This research investigates how the values of clustered datasets, both normalized and non-normalized, influence the computation of Euclidean distance in the K-means algorithm. Additionally, it examines the impact of varying cluster quantities, identified through the elbow method, on the evaluation of the Davies-Bouldin Index (DBI). A dataset comprising 174 records undergoes mining using the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach. In the data preparation phase, the min-max algorithm is applied to ensure that attribute values within the dataset are not diminished relative to each other. Concerning the selection of an optimal K value, the elbow method is employed. In this investigation, two K values exhibit significant mean reduction: the fourth and third cluster quantities. The DBI results for 3 clusters show a smaller value of 0.9250 compared to the DBI result for 4 clusters, which is 1.1584. The fundamental principle of evaluating the Davies-Bouldin Index is that a smaller DBI value (approaching zero but not reaching the minimum) indicates a better cluster. These findings contribute to a better understanding of the evaluation techniques involving the elbow method and Davies-Bouldin Index in clustering analysis and offer insights into the relationship between determining cluster quantities and clustering performance.

Keywords: Clustering, Elbow method, Davies-Bouldin Index.

Abstrak

Penelitian ini menyelidiki bagaimana nilai dataset klasterisasi yang dilakukan normalisasi dan tidak dilakukan normalisasi berpengaruh pada perhitungan Euclidean distance algoritma K-means serta berbagai jumlah klaster yang diidentifikasi melalui metode elbow, memengaruhi nilai evaluasi Indeks Davies-Bouldin (DBI). Dataset yang berjumlah 174 records kemudian dilakukan proses mining dengan pendekatan CRISP-DM (Cross-Industry Standard Process for Data Mining). Pada tahap preparasi data yang dilakukan yaitu penerapan algoritma min-max dengan tujuan agar nilai pada tiap-tiap atribut dalam dataset tidak terreduksi satu sama lain. Terkait pemilihan nilai K yang ideal, digunakan metode elbow untuk menentukan nilai tersebut, pada penyelidikan ini terdapat dua nilai K yang menunjukkan penurunan nilai mean yang signifikan yaitu pada jumlah klaster keempat dan setelah itu nilai klaster ketiga, hasil DBI untuk nilai K=3 (tiga klaster) menujukkan nilai yang lebih kecil yaitu, 0,9250 dibanding hasil DBI untuk nilai K=4 (empat klaster) yaitu 1,1584. Prinsip dasar evaluasi indeks Davies-Bouldin ialah semakin kecil nilai DBI (mendekati nol namun bukan min) adalah klaster terbaik. Temuan ini berkontribusi pada pemahaman yang lebih baik tentang teknik analisis evaluasi metode elbow dan indeks Davies-Bouldin pada klasterisasi dan memberikan wawasan tentang hubungan antara penentuan jumlah klaster dan kinerja pengelompokan.

Kata kunci: Klasterisasi, Metode elbow, Indeks Davies-Bouldin,

1. Pendahuluan

Analisis klaster merupakan teknik dasar dalam bidang analisis data yang digunakan secara luas dalam berbagai domain untuk mengungkap pola dan hubungan yang tersembunyi dalam dataset. Salah satu metode utama untuk melakukan analisis klaster adalah algoritma K-means, yang melibatkan pembagian dataset menjadi klaster-klaster yang berbeda berdasarkan kesamaan mereka. Salah satu aspek kritis dari klasterisasi K-means adalah penentuan jumlah klaster, yang dinyatakan sebagai 'K'. Dalam konteks ini, metode elbow telah muncul sebagai pendekatan populer untuk memilih nilai 'K' yang optimal. Metode ini melibatkan penilaian kinerja klaster dengan berbagai nilai 'K' dan mengidentifikasi titik 'elbow' di mana tingkat peningkatan kualitas klaster mulai melambat. Selanjutnya, evaluasi kualitas klaster memainkan peran penting dalam memahami efektivitas algoritma klasterisasi. Salah satu kriteria umum untuk tujuan ini adalah Indeks Davies-Bouldin (DBI), yang mengukur dissimilaritas rata-rata antara klaster-klasternya dan membantu dalam menilai pemisahan dan kompakness klaster. Evaluasi DBI memberikan wawasan tentang kualitas penugasan klaster dan membantu mengidentifikasi

jumlah klaster yang paling tepat untuk dataset yang diberikan.

Dalam konteks analisis klaster, normalisasi dataset sebelum dilakukan klasterisasi dapat secara signifikan memengaruhi hasil. Normalisasi memastikan bahwa nilai atribut berada pada skala yang serupa, mencegah atribut tunggal dominan dalam proses klasterisasi. Oleh karena itu, pemahaman terhadap dampak normalisasi terhadap hasil klasterisasi, terutama ketika menggunakan algoritma K-means dan metode elbow, sangat penting.

Penelitian ini menggali aspek-aspek tersebut dengan menyelidiki bagaimana normalisasi dataset yang dikelompokkan memengaruhi perhitungan jarak Euclidean dalam algoritma K-means. Selain itu, penelitian ini menjelajahi pengaruh berbagai jumlah klaster, yang ditentukan melalui metode elbow, terhadap evaluasi Indeks Davies-Bouldin.

Melalui analisis komprehensif terhadap dataset yang terdiri dari 174 records, penelitian ini memberikan kontribusi dalam memahami interaksi antara normalisasi, penentuan klaster, dan teknik evaluasi klaster. Temuan-temuan ini memberikan wawasan berharga tentang hubungan antara jumlah klaster dan kinerja klasterisasi, memberikan cahaya tentang konfigurasi optimal untuk mencapai penugasan klaster yang bermakna dan akurat.

2. Metode Penelitian

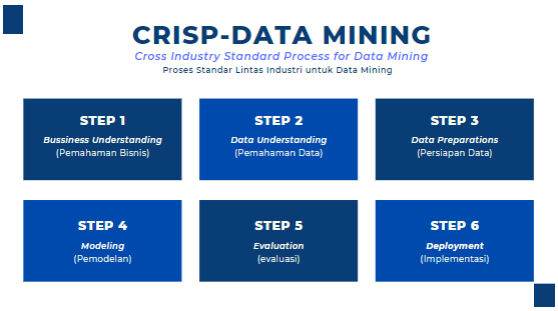
2.1. Objek Penelitian

Sekolah dengan NPSN 60200274 ini terletak di Jln. Siswa desa Darame Kecamatan Morotai Selatan Kabupaten Pulau Morotai. Memiliki luas lahan ± 2 hectare dengan mengadopsi kurikulum K-13. SMP Negeri Unggulan 1 Pulau Morotai mendapat status akreditasi grade A dengan nilai 89 (akreditasi tahun 2015) dari BAN-S/M (Badan Akreditasi Nasional).

Data ini diperoleh melalui petugas atau operator sekolah SMPN Unggluan 1 Kabupaten Pulau Morotai.

2.2. Pendekatan Proses Mining

CRISP-DM (Cross-Industry Standard Process for Data Mining) adalah metodologi panduan fleksibel yang membantu tim proyek untuk mengelola dan menjalankan proyek analisis data dengan efektif. (Budiman et al., 2012) Berikut adalah urutan pendekatan CRISP-DM pada gambar berikut:



Gambar 1. Langkah-langkah CRISP-DM

a. Pemahaman Bisnis: Memahami tujuan bisnis dan tujuan proyek. Identifikasi masalah atau peluang yang ingin diatasi.
b. Pemahaman Data: Mengumpulkan dan memahami data yang tersedia. Evaluasi kualitas, relevansi, dan ketersediaan data.
c. Persiapan Data: Membersihkan, mentransformasi, dan mempersiapkan data untuk analisis. Langkah ini mencakup pemilihan variabel yang relevan dan penggabungan data dari sumber yang berbeda.
d. Modeling: Memilih metode analisis dan membangun model yang sesuai dengan tujuan proyek. Ini melibatkan eksperimen dengan berbagai algoritma dan pendekatan untuk menghasilkan model yang efektif.
e. Evaluasi: Mengukur kinerja model menggunakan metrik yang relevan dengan tujuan bisnis. Model dievaluasi dan dianalisis untuk memastikan bahwa ia memenuhi kriteria keberhasilan yang telah ditetapkan.
f. Implementasi: Solusi yang dikembangkan diterapkan dalam lingkungan bisnis sesuai dengan rencana implementasi yang telah dirancang.

Pendekatan CRISP-DM bersifat siklus, yang berarti bahwa langkah-langkah ini tidak selalu berlangsung dalam urutan linear. Ada kemungkinan untuk kembali ke langkah sebelumnya jika diperlukan perbaikan atau penyesuaian. CRISP-DM metodologi yang digunakan untuk mengatasi proyek analisis data dan data mining.(Ardiada et al., 2019)

2.3. Data

Data yang digunakan dalam proses data mining ini merupakan data rekapan siswa kelas VII pada SMPN Unggluan 1 Kabupaten Pulau Morotai yang berjumlah 208 siswa dengan beberapa atribut yang berjumlah 19 field informasi akademik (nilai), serta informasi yang berkaitan dengan identitas siswa maupun informasi lainnya.

2.4. Metode pada tahap persiapan data

Membersihkan, mentransformasi, dan mempersiapkan data untuk analisis. Langkah ini mencakup pemilihan variabel yang relevan atau yang disebut fitur-selection. Hal ini bertujuan untuk menyesuaikan karakteristik model proses mining yang akan digunakan.

Berikut adalah beberapa metode dalam tahap persiapan data yang digunakan:

a. Normalisasi data dengan algoritma min-max

Proses ini bertujuan agar nilai pada tiap-tiap atribut dalam dataset menjadi seimbang sehingga masing-masing nilai dalam atribut tidak saling mereduksi satu sama lain.

Berikut adalah rumus persamaan dari normalisasi data min-max;

$$\text{Nilai}_{\text{normalisasi}} = \frac{\text{Nilai}_{\text{lama}} - \text{Nilai}_{\text{terendah dalam dataset}}}{\text{Nilai}_{\text{tertinggi dataset}} - \text{Nilai}_{\text{terendah dataset}}}$$

Keterangan :

1. Nilai $_{\text{normalisai}}$ = Nilai yang akan ditransformasi
2. Nilai $_{\text{lama}}$ = masing-masing nilai pada atribut
3. Nilai $_{\text{terendah}}$ = Nilai terkecil dalam atribut dataset
4. Nilai $_{\text{tertinggi}}$ = Nilai tertinggi dalam atribut dataset

b. Seleksi Data
Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam Knowledge Discovery in Database (KDD) dimulai. (Apriliani et al., 2020).
Atribut (kolom) yang terdapat pada dataset ini ialah 19 atribut atau kolom, namun atribut yang akan diambil untuk dilakukan klasterisasi hanya beberapa atribut saja yaitu kolom nama siswa dan beberapa nilai siswa yaitu; nilai UTS dan UAS.

Tabel 1. Dataset setelah dilakukan normalisasi dan seleksi

NO.	SEBELUM		SESUDAH	
	UTS	UAS	UTS	UAS
1	82	89	0,41379	0,65517
2	71	71	0,03448	0,03448
3	74	83	0,13793	0,44828
172	75	75	0,17241	0,17241
173	99	70	1,00000	0,00000
174	91	79	0,72414	0,31034

Adapun data dalam dataset yang dilakukan preparasi yaitu :

a) Normalisasi min-max
Data dari atribut nilai UTS dan UAS dilakukan normalisasi min-max sehingga data tersebut berada dalam range 0 dan 1 hal ini dimaksudkan agar terdapat kesimbangan nilai.
b) Seleksi fitur
Pada tahap ini dilakukan seleksi atribut dimana dari keseluruhan atribut sebanyak 19 atribut yang diambil hanyalah atribut dengan tipe data numerik yaitu atribut nilai UTS dan UAS, hal ini sangat berpengaruh terhadap penentuan model dalam proses mining.

2.5. Modeling

a) K-Means
Berdasarkan hasil preparasi pada tahap sebelumnya maka kemudian dipilih model proses mining unsupervised-learning yaitu klasterisasi. Klasterisasi dengan algoritma K-Means ini dipilih berdasarkan karakteristik dataset. Pada penelitian ini klasterisasi K-Means dilakukan untuk menghitung jarak Euclidean dari dataset yang telah dinormalisasi dan tidak dinormalisasi. (Lestari et al., 2019)
Algoritma klastering K-Means adalah salah satu metode umum dalam analisis data yang digunakan untuk mengelompokkan data menjadi beberapa kelompok atau klaster berdasarkan kesamaan fitur atau atribut. Tujuan dari algoritma K-Means adalah untuk meminimalkan total varian dalam klaster dengan mengalokasikan titik data ke klaster yang sesuai.(Mukrodin1), 2022)
Secara umum, langkah-langkah algoritma K-Means adalah sebagai berikut:

1. Inisialisasi: Pilih jumlah klaster yang diinginkan (K) dan tentukan titik awal (pusat) untuk setiap klaster secara acak atau menggunakan metode lain.
2. Assign Points to Clusters: Untuk setiap titik data, hitung jarak antara titik data dan pusat klaster. Assign (alokasikan) titik data ke klaster dengan pusat terdekat.
3. Update Cluster Centers: Hitung pusat baru untuk setiap klaster dengan menggunakan rata-rata dari semua titik data yang di-assign ke klaster tersebut.
4. Iterasi: Ulangi langkah 2 dan 3 sampai kondisi berhenti terpenuhi. Kondisi berhenti bisa berupa jumlah iterasi yang telah ditentukan, atau perubahan yang kecil dalam posisi pusat klaster.

Formula K-Means untuk menghitung jarak antara dua titik data (x dan y) dalam ruang fitur yang memiliki n dimensi dapat dihitung menggunakan jarak Euclidean, yaitu:

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

Di mana:
a) D(x, y) adalah jarak antara titik data x dan y.
b) n adalah jumlah dimensi (fitur) dari data.
c) x_i dan y_i adalah nilai fitur ke-i dari titik data x dan y.

Selama proses iteratif, pusat klaster baru dihitung dengan mengambil rata-rata dari semua titik data dalam klaster. Jadi, jika C_k adalah klaster ke-k, maka pusat baru m_k dapat dihitung dengan formula:

$$m_k = \frac{1}{N_k}\sum_{x \in c_k} x$$

Di mana:
a) m_k adalah pusat baru klaster ke-k.
b) N_k adalah jumlah titik data dalam klaster ke-k.
c) $\sum_{x \in c_k} x$ adalah penjumlahan dari semua titik data dalam klaster ke-k.

Perlu diingat bahwa algoritma K-Means adalah algoritma iteratif, yang berarti bahwa langkah-langkah 2 dan 3 akan diulang hingga konvergensi atau berhenti setelah jumlah iterasi tertentu.

b) Elbow-Method
Untuk penetuan jumlah klaster, kami gunakan metode Elbow. (Suyanto, 2017) Dengan mengukur jumlah kesalahan kuadrat atau sum of squared error (SSE) antara semua Objek dalam klaster C_i dan centroid C_i yang didefenisikan sebagai berikut:(Anita Fitria Febrianti, 2018)

$$SSE = \sum_{k=1}\sum_{x_i \in S_k} X_i - C_k \quad (2)$$

Keterangan :
K = Jumlah klaster
X_i = Jumlah data
C_k = Jumlah clusteri pada cluster ke K

c) Evaluasi Klaster DBI
Indeks Davies-Bouldin adalah metrik yang digunakan untuk mengukur kualitas klastering dalam analisis data. Tujuannya adalah untuk mengukur seberapa baik klaster yang dihasilkan oleh algoritma klastering memisahkan kelompok data yang berbeda dan mendekati pusat

klasternya. Semakin rendah nilai Davies-Bouldin Index, semakin baik klastering yang dihasilkan.
Rumus Davies-Bouldin Index untuk menghitung kualitas klastering antara dua klaster C_i dan C_j adalah sebagai berikut:

$$R_{ij} = \frac{S_i + S_j}{d_{ij}} \quad (3)$$

Di mana:
1) R_{ij} adalah nilai Davies-Bouldin Index antara klaster C_i dan C_j.
2) S_i adalah ukuran dispersi atau sebaran data dalam klaster C_i, yang dapat dihitung dengan variasi atau jarak rata-rata antara titik data dalam klaster dengan pusat klaster m_i.
3) S_j adalah ukuran dispersi atau sebaran data dalam klaster C_j, yang dapat dihitung dengan variasi atau jarak rata-rata antara titik data dalam klaster dengan pusat klaster m_i.
4) d_{ij} adalah jarak antara pusat klaster m_i dan m_j

Langkah-langkah umum untuk menghitung Davies-Bouldin Index adalah sebagai berikut:
1) Hitung pusat klaster m_i dan m_j untuk masing-masing klaster C_i dan C_j.
2) Hitung ukuran dispersi S_i untuk klaster C_i dan S_j untuk klaster C_j. Ini dapat dilakukan dengan menghitung jarak rata-rata antara titik-titik data dalam klaster dengan pusat klaster.
3) Hitung jarak d_{ij} antara pusat klaster m_i dan m_j.
4) Gunakan rumus Davies-Bouldin Index untuk menghitung nilai R_{ij} antara klaster C_i dan C_j.
5) Ulangi langkah 1 hingga 4 untuk setiap pasangan klaster yang mungkin.

Setelah semua nilai R_{ij} dihitung, Davies-Bouldin Index keseluruhan dapat dihitung dengan rumus:

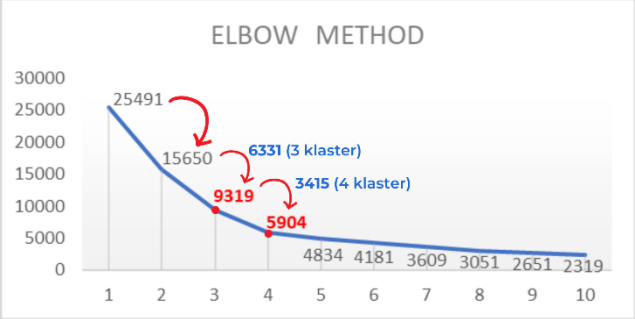
$$DB = \frac{1}{N}\Sigma_{i=1}^{N} \max_{j \neq I} (R_{ij}) \quad (4)$$

Di mana:
- DB adalah Davies-Bouldin Index keseluruhan.
- N adalah jumlah total klaster.

Nilai Davies-Bouldin Index yang lebih rendah menunjukkan kualitas klastering yang lebih baik, karena menunjukkan bahwa klaster-klasternya lebih terpisah dan lebih dekat dengan pusat klasternya.

3. Hasil dan Pembahasan

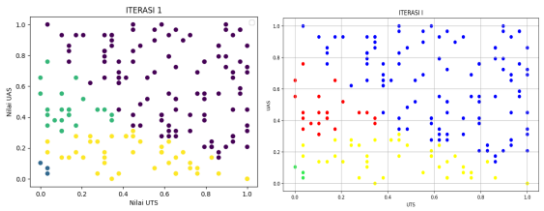
3.1. Analisis Perbandingan Penentuan Jumlah klaster
Dalam penelitian ini, pertama-tama kami membandingkan hasil perhitungan metode elbow dalam menentukan jumlah klaster yang akan dilakukan proses mining dengan algoritma K-Means. Berikut visualisasi hasil dari metode elbow:



Gambar 2. Grafik Elbow

Grafik diatas menujukkan bahwa secara kasat mata terdapat dua siku yang paling dominan, dari nilai k=1 sampai dengan nilai k=10, didapati titik siku pada nilai k = 3 dan k = 4, disini perlu dianalisis secara mendalam terkait dengan hasil tersebut, sebab apabila penentuan nilai k hanya ditentukan berdasarkan penglihatan (secara kasat mata), ini akan mempengaruhi hasil dari evaluasi klaster.
Prinsip dasar dari metode elbow yaitu apabila titik di grafik menunjukkan penurunan varians yang tajam dan menyerupai bentuk siku (elbow), titik inilah yang merupakan indikasi jumlah klaster yang optimal atau ideal untuk dataset tersebut.
Berdasarkan prinsip tersebut maka dapat diketahui bahwa penurunan varians yang signifikan terdapat pada klaster 2 ke 3 dengan nilai selisihnya sebesar 6331 sedangkan selisih penurunan varians dari klaster 3 ke 4 yaitu 3415. Maka seharusnya jumlah klaster ideal adalah 3 klaster.

3.2. Analisis perbandingan terhadap normalisasi data
Setelah didapati nilai k dari hasil perhitungan pada metode elbow, selanjutnya dihitung jarak Euclidean. Pada tahap ini yang akan dianalisa atau diamati yaitu antara dataset yang telah dilakukan normalisasi dengan dataset yang tidak dinormalisasi. Hasil analisis perbandingan terhadap dataset yang dinormalisasi dan tidak dinormalisasi menunjukkan bahwa tidak terdapat perbedaan sama sekali, namun ada hal lain yang dapat dipertimbangkan yaitu, apabila dilihat pada tabel 2 diatas, nilai dari masing-masing atribut (UTS dan UAS) pada dataset yang belum di normalisasi tidak terdapat selisih, sehingga kalaupun tidak dinormalisasi pun kedua nilai dari masing-masing atribut tidak saling mereduksi.



Gambar 3. Grafik dataset normalisasi dan tidak normalisasi

3.3. Analisis Perbandingan Hasil Evaluasi DBI
Pada tahap ini, akan ditampilkan hasil evaluasi perhitungan indeks Davies-Bouldin untuk 3 klaster dan 4 klaster berdasarkan hasil perhitungan dengan metode elbow sebelumnya. Berikut tabel rangkuman hasil klaster K-Means sebanyak 3 klaster dengan total perhitungan jarak Euclidean hingga dinyatakan konvergen sebanyak 16 iterasi

Tabel 2. Perhitungan 3 klaster

Iterasi	Anggota Klaster			C1		C2		C3	
	K1	K2	K3	C x	C y	C x	C y	C x	C y
Ke-1				0,41	0,66	0,03	0,03	0,14	0,45
Ke-2	120	20	34	0,64	0,57	0,29	0,10	0,21	0,39
Ke-3	98	35	41	0,69	0,59	0,42	0,13	0,18	0,52
Ke-4	69	52	53	0,76	0,63	0,51	0,16	0,21	0,61
Ke-5	54	63	57	0,80	0,67	0,53	0,19	0,23	0,64
Ke-6	47	68	59	0,81	0,71	0,55	0,19	0,25	0,64
Ke-7	44	70	60	0,81	0,72	0,57	0,20	0,24	0,63
Ke-8	42	71	61	0,81	0,74	0,58	0,20	0,24	0,63
Ke-9	42	71	61	0,81	0,74	0,59	0,20	0,24	0,62
Ke-10	41	70	63	0,81	0,75	0,60	0,21	0,24	0,62
Ke-11	43	69	62	0,79	0,76	0,61	0,21	0,23	0,60
Ke-12	44	65	65	0,78	0,76	0,63	0,21	0,22	0,57
Ke-13	46	61	67	0,77	0,76	0,66	0,21	0,22	0,54
Ke-14	46	60	68	0,77	0,76	0,67	0,21	0,22	0,53
Ke-15	47	60	67	0,76	0,76	0,67	0,21	0,21	0,53
Ke-16	**47**	**60**	**67**	**0,76**	**0,76**	**0,67**	**0,21**	**0,21**	**0,53**

Berikut tabel rangkuman hasil klaster K-Means sebanyak 4 klaster dengan total perhitungan jarak Euclidean hingga dinyatakan konvergen sebanyak 6 iterasi :

Tabel 3. Perhitungan 4 klaster

Iterasi	Anggota Klaster				C1		C2		C3		C4	
	K1	K2	K3	K4	Cx	Cy	Cx	Cy	Cx	Cy	Cx	Cy
Ke-1					0,4	0,7	0,0	0,0	0,1	0,4	0,2	0,2
Ke-2	106	3	23	42	0,6	0,6	0,0	0,1	0,1	0,5	0,5	0,1
Ke-3	76	9	36	53	0,7	0,7	0,1	0,1	0,2	0,4	0,6	0,2
Ke-4	69	14	36	55	0,6	0,8	0,2	0,1	0,2	0,5	0,7	0,2
Ke-5	59	20	39	56	0,7	0,8	0,2	0,2	0,2	0,6	0,7	0,2
Ke-6	51	27	43	53	0,7	0,8	0,2	0,2	0,2	0,6	0,7	0,2

Dari hasil perhitungan algoritma K-Means diatas kemudian akan diambil nilai dari masing-masing centroid pada iterasi terakhir, baik nilai centroid konvergen iterasi terakhir pada hasil perhitungan algoritma K-Means 3 klaster maupun nilai centroid konvergen iterasi terakhir pada perhitungan K-Means 4 klaster.

3.4. Perhitungan Nilai SSW

Untuk mengetahui kohesi dalam sebuah cluster ke-i adalah dengan menghitung nilai dari Sum of Square Within-cluster (SSW). Kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap titik pusat cluster dari sebuah cluster yang diikuti. Persamaan yang digunakan untuk memperoleh nilai Sum of Square Within cluster adalah sebagai berikut:

rumus-ssw (sum square within):

Berikut tabel rekapan nilai centroid pada iterasi terakhir terhadap kedua klaster tersebut yang telah konvergen untuk selanjutnya dihitung nilai SSW:

$$SSW_i = \frac{1}{m_i} \sum_{j=i}^{m_i} d(x_j, c_i) \quad (5)$$

Tabel 4. Centroid konvergen dari 3 dan 4 klaster

CENTROID	X	Y	CENTROID	X	Y
CENTROID 1	0,76	0,77	CENTROID 1	0,71	0,79
CENTROID 2	0,67	0,21	CENTROID 2	0,21	0,19
CENTROID 3	0,21	0,53	CENTROID 3	0,23	0,62
			CENTROID 4	0,73	0,23

Tabel 5. Nilai SSW dari masing-masing klaster

KLASTER	AGT	SSW	KLASTER	AGT	SSW
KLASTER 1	47	0,77	KLASTER 1	51	0,79
KLASTER 2	60	0,21	KLASTER 2	27	0,19
KLASTER 3	67	0,53	KLASTER 3	43	0,62
			KLASTER 4	53	0,23

3.5. Perhitungan Nilai SSB

Perhitungan Sum of Square Between-cluster (SSB) bertujuan untuk mengetahui separasi antar klaster. Persamaan yang digunakan untuk menghitung nilai Sum of Square Between cluster adalah sebagai berikut.

$$SSB_{i,j} = d(c_i, c_j) \quad (6)$$

Tabel 6. Nilai SSB dari masing-masig klaster

SSB	CENTROID			SSB	CENTROID			
	C1	C2	C3		C1	C2	C3	C4
1	0	0,565	0,594	1	0,000	0,779	0,509	0,555
2	0,565	0	0,556	2	0,779	0,000	0,435	0,522
3	0,594	0,556	0	3	0,509	0,435	0,000	0,637
				4	0,555	0,522	0,637	0,000

3.6. Perhitungan Nilai Rasio

Perhitungan ini Bertujuan untuk mengetahui nilai perbandingan antara cluster ke-i dan cluster ke-j. Untuk menghitung nilai rasio yang dimiliki oleh masing-masing cluster, digunakan persamaan berikut:

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (7)$$

Hasil perhitungan matriks SSW dan SSB, kemudian diambil nilai keduanya dan dihitung nilai rasio menggunakan rumus persamaan diatas untuk dicari nilai rasio maksimalnya atau nilai Rmax, sehingga hasilnya dapat dilihat pada tabel berikut :

Tabel 7. Hasil perhitungan Rasio

R	1	2	3	Rmax	R	1	2	3	4	Rmax
1	**0**	0,87	0,95	**0,95**	1	0,00	1,01	1,23	1,16	**1,23**
2	0,87	**0**	0,69	**0,87**	2	1,01	0,00	0,65	0,52	**1,01**
3	0,95	0,69	**0**	**0,95**	3	1,23	0,65	0,00	0,51	**1,23**
					4	1,16	0,52	0,51	0,00	**1,16**

3.7. Perhitungan Nilai Indeks Davies-Bouldin

Untuk mengevaluasi kalster menggunakan pendekatan ini terlebih dahulu diambil nilai centroid pada iterasi terakhir yang telah dinyatakan kovergen atau tidak ada lagi perpindahan anggota klaster ke titik pusat klaster tetangganya, setelah itu dihitung nilai SSW, SSB dan rasio. Setelah didapatkan nilai rasio maksimalnya barulah kemudian dapat dihitung nilai Indeks Davies-Bouldin melalui persamaan berikut :

$$DB = \frac{1}{N} \Sigma_{i=1}^{N} \max_{j \neq I} (R_{ij})$$

Di mana:

- DB adalah Davies-Bouldin Index keseluruhan.
- N adalah jumlah total klaster.

Untuk N = 3 maka :

$$DB = \frac{1}{3} (0{,}951 + 0{,}872 + 0{,}951) = 0{,}9250$$

N = 3 (klaster)

DB = 0,9250

Untuk N = 4, maka:

$$DB = \frac{1}{4}(1{,}2327 + 1{,}0115 + 1{,}2327 + 1{,}1570) = 1{,}1584$$

N = 4 (klaster)
DB = 1,1584
Prinsip dasar indeks Davies-Bouldin yaitu apabila nilai DBI mendekati nol dan bukan min (DBI>=0) maka klaster tersebut dinggap baik.
Berdasarkan hasil diatas maka klaster terbaik adalah klaster 3 dengan nilai evaluasi DBI=0,9250 hal ini sesuai dengan penurunan nilai varian yang lebih besar dibanding nilai penurunan varian untuk titik klaster 4 pada grafik elbow.

4. Kesimpulan

Hasil evaluasi khususnya evaluasi internal tanpa menggunakan ground-truth sebagai parameter sangat bergantung pada tahap perisapan data serta penetuan jumlah klaster diawal proses data mining. Persiapan data bergantung pada karakteristik dan kondisi data. Terkait dengan normalisasi data pada penelitian ini tidak ada pengaruh dikarenakan nilai data dalam dataset masih dalam kategori seimbang sehingga tidak berdampak pada perhitungan jarak Euclidean.

Catatan yang dianggap penting terkait dengan normalisasi data ialah apabila dataset tidak ditransformasi ke range 0 dan 1 akan berpengaruh pada evaluasi DBI, sebab pada prinsipnya untuk nilai SSW, jarak antar anggota klaster semakin kecil semakin baik, sedangkan untuk SSB, jarak antar klaster semakin besar semakin baik, dan prinsip yang terakhir dalam evaluasi DBI adalah apabila nilai DBI tidak mendekati nol maka klaster tersebut dianggap belum baik, padahal bisa saja disebabkan oleh nilai atribut dalam dataset yang tidak dinormalisasi atau ditransformasi.

Catatan berikut, Sangat penting kiranya jika dalam menetukan jumlah klaster kemudian yang digunakan ialah metode elbow maka disarankan agar tidak menentukan secara kasat-mata atau dengan mengandalkan penilaian tanpa menghitung penurunan nilai varian apabila terdapat dua atau lebih titik siku pada grafik elbow.

Pada penelitian ini ada dua titik pada grafik elbow yang menunjukkan lengkungan yang tajam, secara kasat mata lengkungan itu tepat pada titik klaster ke 4 dari 10 titik klaster sehingga, setelah dilakukan proses mining dengan K-Means klastering sampai pada perhitungan evaluasi barulah keliahatan bahwa nilai DBI menunjukkan bahwa klaster 4 tidak lebih baik dari klaster 3 dengan nilai DBI-nya mendekati nol.

Ucapan Terimaksih

Terhadap tulisan ini, masihlah teramat banyak kekurangan padanya, sehingga penulis membebaskan kepada sesiapapun kiranya meluangkan kritik dan saran agar belenggu kebenaran tunggal (egosentrisme maupun logosentrime) tulisan ini menjadi tergugurkan. Terima kasih diucapkan kepada;
Ibu Hazriani Zainudin dan Bapak Yuyun Wabula selaku Author yang telah membimbing, Ibu Jacklyn Anindya Syah, S.Pd, Gr., guru SMP N Unggulan 1 Pulau Morotai yang telah memberikan data siswa serta terima kasih kepada kawan-kawanku.

Daftar Rujukan

[1] Suyanto. 2017. Data Mining untuk Klasifikasi dan Klasterisasi Data. Edisi Pertama. Bandung: Penerbit Informatika, 2017. p. 342. 978-602-6232-36-6
[2] Anita Fitria Febrianti, A. H. C. A. (2018). K-Means Clustering Denganmetode Elbow Untuk Pengelompokan Kabupaten Dan Kota Di Jawa Timur Berdasarkan Indikator Kemiskinan. Jurnal Teknologi Dan Pendidikan, 8(978-602-5793-40–0), 863–870.
[3] Apriliani, A., Zainuddin, H., Hasanuddin, Z. B., Handayani Makassar, S., & Pembangunan Nasional Veteran Jawa Timur, U. (2020). Peramalan Tren Penjualan Menu Restoran Menggunakan Metode Single Moving Average. 7(6), 1161–1168. https://doi.org/10.25126/jtiik.202072732
[4] Ardiada, I. M. D., Sudarma, M., & Giriantari, D. (2019). Text Mining pada Sosial Media untuk Mendeteksi Emosi Pengguna Menggunakan Metode Support Vector Machine dan K-Nearest Neighbour. Majalah Ilmiah Teknologi Elektro, 18(1), 55. https://doi.org/10.24843/mite.2019.v18i01.p08
[5] Budiman, I., Prahasto, T., & Christyono, Y. (2012). Data Clustering Menggunakan Metodologi Crisp-Dm Untuk Pengenalan Pola Proporsi Pelaksanaan Tridharma. In Seminar Nasional Aplikasi Teknologi Informasi.
[6] Kurniawan, S., Gata, W., Ayu Puspitawati, D., Tabrani, M., Novel, K., Sarjana, P., Komputer, I., Nusa Mandiri Jakarta, S., & Informasi, S. (2017). Perbandingan Metode Klasifikasi Analisis Sentimen Tokoh Politik Pada Komentar Media Berita Online. Masa Berlaku Mulai, 1(3), 176–183.
[7] Lestari, W., Bina, S., & Kendari, B. (2019). Clustering Data Mahasiswa Menggunakan Algoritma K-Means Untuk Menunjang Strategi Promosi (Studi Kasus : STMIK Bina Bangsa Kendari). In SIMKOM (Vol. 4, Issue 2). http://e-jurnal.stmikbinsa.ac.id/index.php/simkom35
[8] Mukrodin1), R. T. D. S. R. E. (2022). Data Mining Clustering Data Obat-Obatan Menggunakan Algoritma K-Means pada RSU An Ni'Mah Wangon. JIKA (Jurnal Informatika), 7, 165–172.