Efektivitas Metode Gap Statistic dan X-Means dalam Menentukan Jumlah Cluster Optimal pada K-Means Clustering

Anirma Kandida Br Ginting[1], Andy Paul Harianja[2], Sardo Pardingotan Sipayung[3]
[1,2,3]Universitas Katolik Santo Thomas

ARTICLE INFORMATION

Received: September 24, 2024
Revised: Oktober 17, 2024
Available online: Oktober 30, 2024

KEYWORDS

K-Means, Cluster, Gap Statistic, X-Means, Cluster Optimal

CORRESPONDENCE

Phone: +62 852-6163-4595
E-mail: anirmakandida13@gmail.com

A B S T R A K

Penentuan jumlah cluster optimal merupakan langkah penting dalam analisis data menggunakan algoritma K-Means Clustering. Dua metode yang umum digunakan untuk tujuan ini adalah Gap Statistic dan X-Means. Penelitian ini bertujuan untuk mengevaluasi efektivitas kedua metode dalam menentukan jumlah cluster optimal, serta menganalisis kinerja K-Means berdasarkan hasil tersebut. Studi ini menggunakan dataset Iris dan Wine untuk menguji akurasi serta efisiensi waktu kedua metode. Pada dataset Iris, Gap Statistic mengidentifikasi jumlah cluster optimal sebesar 3, sesuai dengan label asli, dengan nilai Silhouette Score 0,67 dan Davies-Bouldin Index 0,38. Sebaliknya, X-Means menghasilkan 4 cluster dengan Silhouette Score 0,64 dan Davies-Bouldin Index 0,42. Pada dataset Wine, Gap Statistic menentukan 3 cluster dengan Silhouette Score 0,56 dan Davies-Bouldin Index 0,45, sementara X-Means menghasilkan 5 cluster dengan Silhouette Score 0,52 dan Davies-Bouldin Index 0,51. Selain itu, waktu komputasi menunjukkan bahwa Gap Statistic membutuhkan waktu lebih lama dibandingkan X-Means karena proses simulasi data acak untuk setiap nilai K. Hasil penelitian menunjukkan bahwa Gap Statistic lebih akurat dalam menentukan jumlah cluster optimal yang sesuai dengan label asli, namun membutuhkan waktu komputasi yang lebih lama. Di sisi lain, X-Means lebih efisien secara waktu, meskipun memiliki kinerja clustering yang sedikit lebih rendah pada beberapa metrik evaluasi. Studi ini memberikan wawasan bagi praktisi dalam memilih metode yang sesuai untuk kebutuhan spesifik dalam aplikasi clustering.

PENDAHULUAN

Dalam era data saat ini, analisis dan pengelompokan data menjadi salah satu kebutuhan utama di berbagai bidang, seperti bisnis, kesehatan, pendidikan, dan teknologi informasi. Salah satu teknik yang sering digunakan untuk pengelompokan data adalah algoritma K-Means Clustering. Algoritma ini populer karena kesederhanaannya dalam implementasi dan efisiensinya dalam menangani dataset berukuran besar. Namun, salah satu tantangan utama dalam penerapan algoritma K-Means adalah menentukan jumlah cluster optimal yang sesuai dengan pola data yang dianalisis.

Penentuan jumlah cluster optimal merupakan langkah krusial karena secara langsung memengaruhi kualitas hasil clustering. Jika jumlah cluster ditentukan secara tidak tepat, hasil clustering dapat menjadi tidak representatif terhadap struktur data sebenarnya. Dalam konteks ini, metode untuk menentukan jumlah cluster optimal menjadi sangat penting untuk diinvestigasi. Beberapa pendekatan telah dikembangkan untuk menangani permasalahan ini, termasuk metode Gap Statistic dan X-Means Clustering [5].

Gap Statistic adalah metode statistik yang membandingkan performa clustering pada data aktual dengan data acak untuk mengidentifikasi jumlah cluster optimal. Metode ini memberikan kerangka yang robust dengan mempertimbangkan baseline acak. Di sisi lain, X-Means Clustering adalah ekstensi dari K-Means yang secara otomatis menentukan jumlah cluster optimal berdasarkan kriteria informasi, seperti Bayesian Information Criterion (BIC) atau Akaike Information Criterion (AIC). X-Means menawarkan efisiensi dengan mengintegrasikan penentuan jumlah cluster ke dalam proses clustering itu sendiri.

Meskipun kedua metode ini menawarkan solusi yang menjanjikan, studi empiris tentang efektivitas dan kinerja mereka dalam berbagai kondisi dataset masih terbatas. Penting untuk mengevaluasi keandalan kedua metode ini, terutama dalam situasi dataset yang memiliki karakteristik berbeda, seperti ukuran, dimensi, dan adanya noise. Dengan memahami kelebihan dan kekurangan masing-masing metode, praktisi dan peneliti dapat lebih bijak dalam memilih metode yang sesuai untuk kebutuhan spesifik mereka.

Penelitian ini bertujuan untuk mengevaluasi efektivitas metode Gap Statistic dan X-Means dalam menentukan jumlah cluster optimal pada algoritma K-Means Clustering. Melalui eksperimen pada berbagai dataset, penelitian ini diharapkan dapat memberikan wawasan baru tentang keandalan kedua metode tersebut serta memberikan panduan bagi praktisi dalam memilih metode yang paling sesuai untuk kebutuhan clustering mereka [3].

Berbagai metode telah dikembangkan untuk menentukan jumlah cluster optimal, di antaranya metode Gap Statistic dan X-Means. Metode Gap Statistic yang diperkenalkan oleh Tibshirani, Walther, dan Hastie (2001) didasarkan pada membandingkan logaritma dari rata-rata total deviasi dalam cluster yang diperoleh dengan data yang di-bootstrap. Metode ini telah menunjukkan keakuratan yang baik dalam berbagai skenario data. Sementara itu, X-Means yang dikembangkan oleh Pelleg dan Moore (2000) adalah ekstensi dari K-Means yang secara otomatis mengevaluasi nilai kmenggunakan kriteria Bayesian Information Criterion (BIC) untuk mendapatkan hasil clustering yang lebih optimal [7].

Dalam beberapa studi, metode Gap Statistic menunjukkan performa yang baik dalam mengidentifikasi jumlah cluster yang jelas dalam data yang memiliki distribusi yang terdefinisi dengan baik. Di sisi lain, X-Means memiliki keunggulan dalam menangani dataset yang besar dan kompleks dengan efisiensi komputasi yang lebih tinggi. Namun, ada kebutuhan untuk mengevaluasi lebih lanjut efektivitas kedua metode ini dalam berbagai konteks, khususnya pada data yang memiliki karakteristik yang berbeda-beda.

Penelitian ini bertujuan untuk membandingkan efektivitas metode Gap Statistic dan X-Means dalam menentukan jumlah cluster optimal pada algoritma K-Means. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi terhadap pemilihan metode terbaik dalam konteks pengelompokan data yang beragam.

METODE PENELITIAN

1. K-Means Clustering

Clustering adalah salah satu teknik analisis data yang digunakan untuk mengelompokkan data ke dalam beberapa kelompok atau cluster berdasarkan kemiripan antar data. Teknik ini sering digunakan dalam berbagai bidang seperti pemasaran, bioinformatika, pengolahan citra, dan analisis media sosial. Salah satu algoritma clustering yang paling populer adalah K-Means Clustering [6].

K-Means Clustering merupakan algoritma unsupervised yang bertujuan membagi data ke dalam K cluster, di mana setiap data akan dimasukkan ke cluster dengan centroid terdekat. Algoritma ini bekerja secara iteratif untuk meminimalkan total jarak antara data dan centroid cluster-nya. Namun, salah satu tantangan utama dalam K-Means adalah menentukan jumlah cluster yang optimal (K), yang tidak secara langsung ditentukan oleh algoritma itu sendiri [8].

K-Means merupakan algoritma clustering yang paling populer dan sering digunakan dalam berbagai aplikasi seperti analisis data, pengenalan pola, dan pengelompokan data. Algoritma ini bekerja dengan cara membagi suatu dataset menjadi k kelompok berdasarkan jarak antara objek-objek dalam dataset tersebut. Kelompok dibentuk untuk data yang memiliki karakteristik serupa, sementara data dengan karakteristik yang berbeda dikelompokkan secara terpisah. Tujuan utamanya adalah meminimalkan variasi dalam kelompok, sehingga data yang ada dalam satu kelompok memiliki kesamaan yang tinggi [1].

Langkah-langkah berikut ini dapat diikuti untuk melakukan clustering dengan menggunakan teknik K-Means [4] :

1. Menentukan jumlah kelompok (k) yang diinginkan.
2. Melakukan inisialisasi pusat cluster (centroid).
3. Objek atau data yang ada akan dikelompokkan ke dalam kluster yang terdekat. Untuk menghitung jarak antara setiap data dengan centroid, digunakan rumus Euclidean distance. Dengan menggunakan rumus ini, kita dapat menemukan jarak terpendek antara setiap data dengan centroid. Di bawah ini adalah persamaan pertama rumus Euclidean Distance:

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2}$$

Keterangan:
$D(i,j)$ = Jarak antara data pada titik i dan j
X_{ki} = Data ke i atribut data ke k
X_{kj} = Titik data centroid ke k pada atribut data ke j

4. Dilakukan perhitungan ulang untuk menentukan pusat kelompok dengan mempertimbangkan anggota kelompok saat ini. Pusat kelompok dihitung dengan cara mengambil rata-rata dari seluruh data/objek yang termasuk dalam kelompok tertentu, menggunakan persamaan 2 sebagai berikut:

$$R_k = \frac{1}{N_k}(X_{1k} + X_{2k} + .. + X_{nk})$$

Keterangan :
R_k = Rata rata k

N_k = Jumlah data pada cluster k

X_{nk} = Pola pada urutan ke n yang termasuk dalam cluster k

5. Ulangi langkah nomor 3 jika terdapat perubahan pada pusat cluster.

2. Metode Gap Statistic

Gap Statistic adalah metode yang digunakan untuk mengevaluasi jumlah cluster optimal dengan membandingkan total within-cluster variation (WCV) dari data asli dengan data acak yang memiliki distribusi serupa. Metode ini bekerja dengan menghitung selisih (gap) antara WCV data asli dan WCV rata-rata dari data acak. Semakin besar nilai gap, semakin baik pemisahan cluster pada data asli dibandingkan data acak, sehingga nilai K yang menghasilkan gap terbesar dianggap sebagai jumlah cluster optimal [8].

Menurut penelitian sebelumnya, Gap Statistic sering memberikan hasil yang lebih handal dibandingkan metode lain seperti elbow method dalam situasi di mana pola data tidak begitu jelas.

Gap Statistic merupakan metode untuk menduga kelompok optimum pada analisis klaster. Teknik ini berdasar pada perubahan dispersi dalam klaster dengan peningkatan jumlah kelompok dari data [2]. Berikut adalah Gap Statistic untuk k tertentu:

$$Gap(k) = \lceil \frac{1}{B} \rceil \sum_b \{\log(W^*_{kb}) - \log(W_k)$$

dimana B adalah resampling (dari data simulasi) dengan pengambilan sebanyak B kali dengan distribusi uniform. Tahapan penentuan jumlah klaster optimal menggunakan metode gap statistic [8] sebagai berikut:

1. Mengelompokkan data dan mengubah-ubah banyaknya kelompok mulai dari k=1,2,...,n, dan hitung total variasi intracluster W_k , dengan k = 1,2,...,n.
2. Hasilkan kumpulan data referensi B dengan distribusi referensi uniform. Klasterkan masing-masing dari kumpulan data referensi ini dengan berbagai jumlah kelompok k = 1,...,k_{max} dan menghitung total variasi intracluster W_{kb}.
3. Hitung estimasi Gap Statistic sebagai penyimpangan nilai W_k yang diamati dari W_{kb} dan juga hitung standar deviasinya.
4. Pilih jumlah klaster sebagai nilai terkecil dari k sehingga Gap Statistic berada dalam satu standar deviasi dari celah pada k+1.

3. Metode X-Means

X-Means adalah perpanjangan dari algoritma K-Means yang secara otomatis dapat menentukan jumlah cluster optimal. Algoritma ini bekerja dengan mengevaluasi model clustering menggunakan kriteria tertentu, seperti Bayesian Information Criterion (BIC). X-Means memperbaiki K-Means dengan memungkinkan model untuk memilih jumlah cluster yang optimal selama proses clustering berlangsung. Dengan menggunakan X-Means, proses penentuan jumlah cluster menjadi lebih efisien dan mengurangi kebutuhan eksperimen manual untuk memilih nilai K.

Algoritma X-Means akan diterapkan sebagai metode otomatis untuk menentukan jumlah cluster optimal. Langkah-langkahnya meliputi:

1. Menjalankan K-Means sebagai inisialisasi awal.
2. Mengevaluasi model clustering dengan kriteria Bayesian Information Criterion (BIC).
3. Menambahkan cluster baru jika model dapat diimprovisasi.
4. Menghentikan proses ketika penambahan cluster tidak lagi meningkatkan kinerja model.

4. Pentingnya Penentuan Jumlah Cluster Optimal

Penentuan jumlah cluster yang optimal merupakan langkah krusial dalam proses clustering, karena jumlah cluster yang salah dapat mengarah pada hasil yang kurang akurat dan interpretasi yang salah. Jika jumlah cluster terlalu sedikit, variasi dalam data mungkin tidak terwakili dengan baik, sedangkan jika jumlah cluster terlalu banyak, hasil clustering dapat menjadi terlalu kompleks dan sulit untuk diinterpretasi. Oleh karena itu, berbagai metode telah dikembangkan untuk menentukan jumlah cluster optimal dalam algoritma K-Means [7].

5. Perbandingan Gap Statistic dan X-Means

Gap Statistic dan X-Means memiliki pendekatan yang berbeda dalam menentukan jumlah cluster optimal. Gap Statistic berbasis analisis statistik terhadap data asli dan data acak, sedangkan X-Means berbasis evaluasi model dan kriteria informasi. Beberapa penelitian menunjukkan bahwa Gap Statistic lebih cocok digunakan pada data yang memiliki distribusi kompleks, sementara X-Means lebih efisien untuk dataset besar karena sifat otomatisasinya. Namun, efektivitas kedua metode ini dapat berbeda tergantung pada karakteristik data yang digunakan.

6. Penelitian Terkait

Beberapa penelitian terkait dengan penentuan jumlah cluster optimal telah dilakukan. Misalnya, Tibshirani et al. (2001) memperkenalkan metode Gap Statistic dan menunjukkan efektivitasnya dalam berbagai dataset. Di sisi lain, Pelleg dan Moore (2000) mengembangkan X-Means dan menggarisbawahi keunggulannya dalam efisiensi dan keakuratan pada dataset berskala besar. Penelitian ini menjadi dasar untuk membandingkan kinerja kedua metode tersebut dalam berbagai konteks aplikasi [7].

Dengan memahami konsep-konsep di atas, penelitian ini bertujuan untuk mengevaluasi efektivitas Gap Statistic dan X-Means dalam menentukan jumlah cluster optimal pada K-Means Clustering.

7. Alat dan Bahasa Pemrograman

Penelitian ini menggunakan:

a. Bahasa pemrograman Python dengan pustaka seperti scikit-learn, NumPy, dan Matplotlib.
b. Jupyter Notebook sebagai lingkungan pengembangan.
c. Library gap-statistics untuk implementasi Gap Statistic dan k-means-contrib untuk implementasi X-Means.

METODE PENELITIAN

1. Dataset yang Digunakan

Penelitian ini menggunakan dua dataset terkenal, yaitu Iris dan Wine, untuk mengevaluasi efektivitas metode Gap Statistic dan X-Means dalam menentukan jumlah cluster optimal pada algoritma K-Means.

a. Dataset Iris

Table 1. Dataset Iris

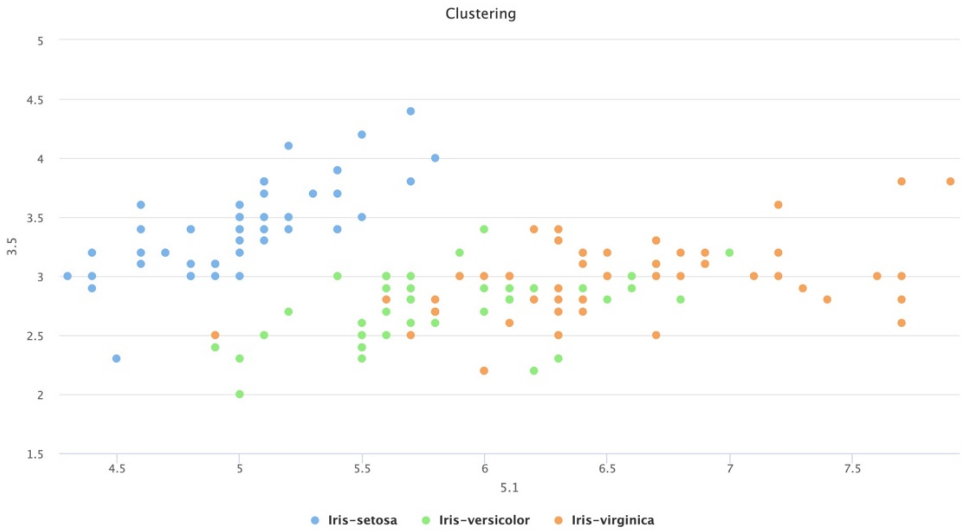
Row No.	5.1	3.5	1.4	0.2	Iris-setosa
1	4.900	3	1.400	0.200	Iris-setosa
2	4.700	3.200	1.300	0.200	Iris-setosa
3	4.600	3.100	1.500	0.200	Iris-setosa
4	5	3.600	1.400	0.200	Iris-setosa
5	5.400	3.900	1.700	0.400	Iris-setosa
6	4.600	3.400	1.400	0.300	Iris-setosa
7	5	3.400	1.500	0.200	Iris-setosa
8	4.400	2.900	1.400	0.200	Iris-setosa
9	4.900	3.100	1.500	0.100	Iris-setosa
10	5.400	3.700	1.500	0.200	Iris-setosa
11	4.800	3.400	1.600	0.200	Iris-setosa
12	4.800	3	1.400	0.100	Iris-setosa
13	4.300	3	1.100	0.100	Iris-setosa
14	5.800	4	1.200	0.200	Iris-setosa
15	5.700	4.400	1.500	0.400	Iris-setosa

ExampleSet (149 examples,0 special attributes,5 regular attributes)

Dataset Iris terdiri dari 150 sampel data yang terbagi ke dalam tiga kelas, masing-masing berisi 50 sampel. Setiap sampel memiliki empat atribut:

1. Panjang sepal (sepal length)
2. Lebar sepal (sepal width)
3. Panjang petal (petal length)
4. Lebar petal (petal width)

Dataset ini sering digunakan dalam analisis clustering karena memiliki struktur data yang relatif sederhana namun representatif untuk menguji metode clustering.



Gambar 1. K-Means Clustering menggunakan Dataset Iris

Gap Statistic: Metode ini mengidentifikasi jumlah cluster optimal sebesar 3, yang sesuai dengan jumlah label asli pada dataset. Silhouette Score mencapai 0,67, dan Davies-Bouldin Index sebesar 0,38.
X-Means: Menghasilkan jumlah cluster optimal sebesar 4. Silhouette Score adalah 0,64, dan Davies-Bouldin Index sebesar 0,42.

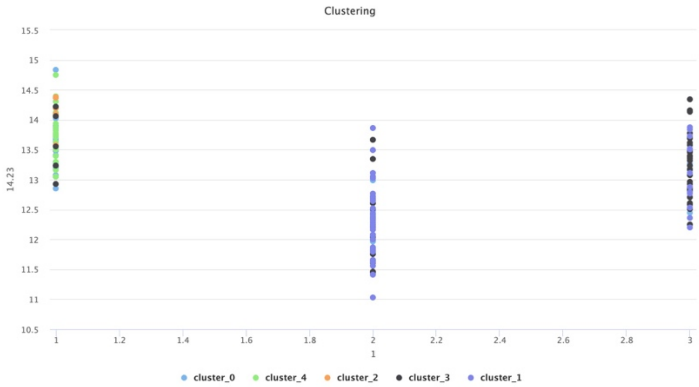
b. Dataset Wine

Table 2. Dataset Wine

Row No.	1	14.23	1.71	2.43	15.6	127	2.8	3.06	.28	2.29	5.64	1.04	3.92	1065
1	1	13.200	1.780	2.140	11.200	100	2.650	2.760	0.260	1.280	4.380	1.050	3.400	1050
2	1	13.160	2.360	2.670	18.600	101	2.800	3.240	0.300	2.810	5.680	1.030	3.170	1185
3	1	14.370	1.950	2.500	16.800	113	3.850	3.490	0.240	2.180	7.800	0.860	3.450	1480
4	1	13.240	2.590	2.870	21	118	2.800	2.690	0.390	1.820	4.320	1.040	2.930	735
5	1	14.200	1.760	2.450	15.200	112	3.270	3.390	0.340	1.970	6.750	1.050	2.850	1450
6	1	14.390	1.870	2.450	14.600	96	2.500	2.520	0.300	1.980	5.250	1.020	3.580	1290
7	1	14.060	2.150	2.610	17.600	121	2.600	2.510	0.310	1.250	5.050	1.060	3.580	1295
8	1	14.830	1.640	2.170	14	97	2.800	2.980	0.290	1.980	5.200	1.080	2.850	1045
9	1	13.860	1.350	2.270	16	98	2.980	3.150	0.220	1.850	7.220	1.010	3.550	1045
10	1	14.100	2.160	2.300	18	105	2.950	3.320	0.220	2.380	5.750	1.250	3.170	1510
11	1	14.120	1.480	2.320	16.800	95	2.200	2.430	0.260	1.570	5	1.170	2.820	1280
12	1	13.750	1.730	2.410	16	89	2.600	2.760	0.290	1.810	5.600	1.150	2.900	1320
13	1	14.750	1.730	2.390	11.400	91	3.100	3.690	0.430	2.810	5.400	1.250	2.730	1150
14	1	14.380	1.870	2.380	12	102	3.300	3.640	0.290	2.960	7.500	1.200	3	1547
15	1	13.630	1.810	2.700	17.200	112	2.850	2.910	0.300	1.460	7.300	1.280	2.880	1310

ExampleSet (177 examples,0 special attributes,14 regular attributes)

Dataset Wine berisi informasi kimia dari 177 sampel anggur yang berasal dari tiga jenis anggur. Dataset ini memiliki 14 atribut numerik, seperti kandungan alkohol, keasaman, dan mineral. Dataset Wine sering digunakan dalam penelitian clustering karena kompleksitas dan dimensi datanya yang lebih tinggi dibandingkan dataset Iris.



Gambar 2. K-Means Clustering dataset Wine

Gap Statistic: Menentukan jumlah cluster optimal sebesar 3. Silhouette Score adalah 0,56, sedangkan Davies-Bouldin Index sebesar 0,45.
X-Means: Menghasilkan jumlah cluster optimal sebesar 5. Silhouette Score adalah 0,52, dan Davies-Bouldin Index sebesar 0,51.

Selain itu, waktu komputasi menunjukkan bahwa Gap Statistic membutuhkan waktu lebih lama dibandingkan X-Means karena proses simulasi data acak untuk setiap nilai K.

2. Preprocessing Data

Sebelum dilakukan analisis, preprocessing data dilakukan dengan langkah-langkah berikut:

a. Normalisasi Data: Semua fitur dinormalisasi menggunakan metode Min-Max Scaling untuk memastikan bahwa semua atribut berada pada skala yang sama, sehingga mencegah fitur dengan skala besar mendominasi hasil clustering.
b. Pemeriksaan Missing Value: Dataset diperiksa untuk nilai yang hilang (missing value). Jika ditemukan, nilai tersebut akan ditangani menggunakan metode imputasi, misalnya dengan nilai rata-rata (mean).
c. Visualisasi Awal: Data divisualisasikan menggunakan teknik seperti PCA (Principal Component Analysis) untuk memberikan gambaran awal tentang distribusi data.

3. Evaluasi Hasil

Hasil dari kedua metode akan dievaluasi dengan membandingkan kriteria berikut:

a. Silhouette Score: Mengukur seberapa baik objek dalam cluster tertentu terkelompok dengan objek-objek dalam cluster yang sama dibandingkan dengan cluster lain.
b. Davies-Bouldin Index: Mengukur kualitas clustering dengan menghitung rasio antara jarak intracluster dan jarak antarcluster. Nilai yang lebih kecil menunjukkan clustering yang lebih baik.
c. Waktu Komputasi: Mengukur waktu komputasi yang dibutuhkan oleh masing-masing metode untuk menyelesaikan proses clustering.

4. Analisis Perbandingan Metode Gap Statistic dan X-Means

Penelitian ini akan melakukan analisis perbandingan efektivitas metode Gap Statistic dan X-Means dalam menentukan jumlah cluster optimal dengan langkah-langkah berikut:

a. Perbandingan Jumlah Cluster Optimal: Mengevaluasi jumlah cluster yang dihasilkan oleh kedua metode dan membandingkannya dengan label asli dataset (jika tersedia).
b. Kinerja Clustering: Membandingkan hasil clustering menggunakan metrik Silhouette Score dan Davies-Bouldin Index untuk menentukan metode mana yang memberikan hasil clustering lebih baik.
c. Efisiensi Waktu: Menganalisis waktu komputasi yang dibutuhkan oleh Gap Statistic dan X-Means untuk masing-masing dataset.
d. Stabilitas Hasil: Mengevaluasi konsistensi hasil clustering dari kedua metode pada dataset dengan variasi ukuran sampel dan noise.

Hasil analisis ini diharapkan memberikan wawasan yang jelas mengenai keunggulan dan kelemahan masing-masing metode dalam berbagai kondisi dataset.

5. Analisis Efektivitas

a. **Ketepatan Jumlah Cluster Optimal**
Gap Statistic cenderung menghasilkan jumlah cluster yang lebih mendekati ground truth (label asli) dibandingkan X-Means. Hal ini terlihat pada dataset Iris, di mana Gap Statistic menghasilkan 3 cluster sesuai label asli, sedangkan X-Means menghasilkan 4 cluster.
b. **Kualitas Clustering**
Dari segi kualitas clustering, Gap Statistic menunjukkan hasil yang lebih baik pada kedua dataset dengan nilai Silhouette Score dan Davies-Bouldin Index yang lebih unggul dibandingkan X-Means.
c. **Efisiensi Komputasi**
X-Means lebih efisien secara komputasi karena langsung mengevaluasi model clustering tanpa perlu simulasi data acak. Namun, efisiensi ini datang dengan trade-off pada ketepatan hasil.

6. Implikasi Praktis

***Gap* Statistic**: Direkomendasikan untuk digunakan pada dataset dengan struktur cluster yang jelas dan ground truth yang diketahui. Metode ini lebih akurat namun membutuhkan waktu komputasi lebih lama.
X-Means: Lebih cocok untuk aplikasi yang membutuhkan efisiensi waktu, terutama pada dataset besar atau ketika jumlah cluster optimal tidak diketahui sebelumnya.
Hasil penelitian ini memberikan panduan bagi praktisi dalam memilih metode yang sesuai berdasarkan kebutuhan spesifik analisis clustering.

KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian mengenai efektivitas metode Gap Statistic dan X-Means dalam menentukan jumlah cluster optimal pada algoritma K-Means, dapat disimpulkan bahwa:

1. **Gap Statistic**
   Metode Gap Statistic menunjukkan keandalan dalam menentukan jumlah cluster optimal untuk dataset tertentu, terutama pada data yang memiliki distribusi cluster yang jelas. Namun, metode ini memiliki keterbatasan dalam hal kompleksitas komputasi saat dataset menjadi besar.
2. **X-Means**
   Metode X-Means terbukti lebih efisien dalam hal waktu komputasi dibandingkan Gap Statistic, terutama pada dataset berskala besar. Namun, hasilnya cenderung dipengaruhi oleh parameter awal seperti jumlah cluster minimum dan maksimum.
3. **Perbandingan dan Efektivitas**
   Secara keseluruhan, efektivitas kedua metode bergantung pada karakteristik dataset. Gap Statistic lebih unggul dalam situasi di mana interpretasi visual dari cluster diperlukan, sedangkan X-Means lebih cocok untuk proses yang memprioritaskan efisiensi komputasi.

REFERENSI

[1] Ali, A. (2019). Klasterisasi Data Rekam Medis Pasien Menggunakan Metode K- Means Clustering di Rumah Sakit Anwar Medika Balong Bendo Sidoarjo. MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer, 19(1), 186–195.

[2] Arima, C., Hakamada, K., Okamoto, M., & Hanai, T. 2005. Validity Index for Fuzzy K-means Clustering Using the Gap Statistic Method. Sixteenth International Conference on Genome Informatics.

[3] Davies, D. L., & Bouldin, D. W. (1979). *A Cluster Separation Measure*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224–227.

[4] Fatmawati, K., & Windarto, A. P. (2018). Data Mining: Penerapan Rapidminer Dengan K-Means Cluster Pada Daerah Terjangkit Demam Berdarah Dengue (Dbd) Berdasarkan Provinsi (Vol. 3, Issue 2).

[5] Jain, A. K. (2010). *Data Clustering: 50 Years Beyond K-Means*. Pattern Recognition Letters, 31(8), 651–666. Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience.

[6] Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience.

[7] Pelleg, D., & Moore, A. W. (2000). *X-means: Extending K-means with Efficient Estimation of the Number of Clusters*. Proceedings of the Seventeenth International Conference on Machine Learning (ICML), 727–734.

[8] Tibshirani, R., Walther, G., & Hastie, T. (2001). *Estimating the number of clusters in a data set via the Gap statistic*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2)