Algoritma: Jurnal Ilmu Komputer Informatika **e-ISSN 2598-6341**
Volume: 9, Nomor: 2, November 2025: 72-83
DOI: 10.30829/algoritma.v9i2.25017
https://jurnal.uinsu.ac.id/index.php/algoritma

Studi Pengelompokan Multimetode Provinsi di Sumatera Utara Menggunakan Pendekatan PCA dan K-Means

Fitra Hidayat Lubis*[1], Suthan Farras Ashar[2], OK Mhd Fahri Al-Faruqy M.S[3], Ahmad Boby Amari[4]
[1,2,3,4] Program Studi Ilmu Komputer, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sumetera Utara
E-mail: hidayatfitra80@gmail.com*[1], suthanfarras160@gmail.com[2], fahryalfaruqy@gmail.com[3], bobbyemarie169@gmail.com[4]

Abstrak

Penelitian ini bertujuan untuk mengklasifikasikan wilayah di Sumatera Utara berdasarkan sejumlah indikator sosial dan ekonomi dengan menerapkan pendekatan pengelompokan multi-metode. Analisis Komponen Utama (PCA) digunakan untuk mengurangi dimensi data dan mengidentifikasi variabel-variabel utama yang berpengaruh, sementara algoritma K-Means digunakan untuk membentuk kluster berdasarkan kesamaan karakteristik. Hasil analisis menunjukkan bahwa kombinasi PCA dan K-Means mampu mengelompokkan provinsi atau wilayah secara lebih efisien dan interpretatif. Kluster yang terbentuk mencerminkan pola kesamaan di antara wilayah dalam hal perkembangan sosial dan ekonomi, sehingga dapat menjadi dasar untuk merumuskan kebijakan pengembangan regional yang lebih terarah. Temuan ini menunjukkan bahwa pendekatan multi-metode dapat memberikan hasil yang lebih komprehensif dalam pengelompokan data spasial.

Kata kunci: Klasterisasi, Analisis Komponen Utama (PCA), K-Means, multimetode, Sumatera Utara.

Abstract

This study aims to classify regions in North Sumatra based on a set of social and economic indicators by applying a multi-method clustering approach. Principal Component Analysis (PCA) is employed to reduce data dimensionality and identify the most influential variables, while the K-Means algorithm is used to form clusters based on similarity of characteristics. The results indicate that the combination of PCA and K-Means can cluster provinces or regions more efficiently and interpretably. The resulting clusters reflect patterns of similarity among regions in terms of social and economic development, thus providing a basis for formulating more targeted regional development policies. These findings demonstrate that a multi-method approach can yield more comprehensive results in spatial data clustering.

Keywords: Clustering, Principal Component Analysis (PCA), K-Means, multi-method, North Sumatra.

Algoritma: Jurnal Ilmu Komputer Informatika **e-ISSN 2598-6341**
Volume: 9, Nomor: 1, April 2025: 72-83
DOI: 10.30829/algoritma.v9i2.25017
https://jurnal.uinsu.ac.id/index.php/algoritma

1. PENDAHULUAN

1.1 Latar Belakang

Pulau Sumatra merupakan salah satu pulau terbesar di Indonesia, terdiri dari sepuluh provinsi dengan karakteristik ekonomi, sosial, dan demografis yang beragam. Setiap provinsi memiliki potensi dan tantangan yang berbeda dalam hal pembangunan, industri, tenaga kerja, dan kesejahteraan masyarakat. Oleh karena itu, diperlukan metode yang efektif untuk memahami pola dan klasifikasi karakteristik provinsi guna mendukung perencanaan pembangunan yang lebih baik [1] .

Salah satu metode yang dapat digunakan untuk analisis pola adalah klasterisasi data. Teknik klasterisasi memungkinkan pengelompokkan objek berdasarkan kesamaan karakteristiknya, sehingga wilayah dengan sifat serupa dapat berada dalam kelompok yang sama. Principal Component Analysis (PCA) berfungsi untuk mereduksi dimensi data, mengurangi kompleksitas, dan meningkatkan efisiensi analisis. Sementara itu, algoritma K-Means Clustering digunakan untuk mengelompokkan data berdasarkan centroid yang mewakili pola dominan dalam setiap klaster [2].

Dengan mengintegrasikan PCA dan K-Means, penelitian ini bertujuan untuk melakukan klasterisasi provinsi di Pulau Sumatera berdasarkan berbagai indikator utama yang mempengaruhi pembangunan. Hasil dari klasterisasi ini diharapkan dapat memberikan wawasan yang lebih mendalam mengenai kesamaan dan perbedaan antar provinsi, sehingga dapat menjadi acuan bagi pemerintah daerah dan pemangku kebijakan dalam merancang strategi pembangunan lebih tepat [3].

1.2 Rumusan Masalah

1. Bagaimana penerapan metode PCA dalam mengelompokkan provinsi di Pulau Sumatera berdasarkan indikator utama?
2. Bagaimana hasil klasterisasi dengan algoritma K-Means setelah reduksi dimensi menggunakan PCA?
3. Bagaimana karakteristik utama yang mempengaruhi klasterisasi provinsi di Pulau Sumatera?
4. Apa interpretasi dari klaster yang terbentuk, serta implikasi bagi pembangunan wilayah?

1.3 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah:

1. Mengidentifikasi karakteristik utama yang membedakan provinsi di Pulau Sumatera.
2. Menganalisis efektivitas metode PCA dalam reduksi dimensi dan pemilihan fitur.
3. Mengimplementasikan algoritma K-Means Clustering untuk mengelompokkan provinsi berdasarkan karakteristik utama.
4. Menginterpretasikan hasil klasterisasi sebagai dasar perencanaan pembangunan wilayah.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat bagi berbagai pihak, antara lain:

1. Pemerintah daerah, dalam memahami pola dan karakteristik provinsi guna merancang kebijakan berbasis data.
2. Akademisi dan peneliti, sebagai referensi dalam studi terkait klasterisasi wilayah dan pemanfaatan teknik PCA-K-Means dalam analisis data.
3. Pelaku bisnis dan investor, dalam mempertimbangkan strategi ekspansi bisnis berdasarkan karakteristik ekonomi dan sosial tiap provinsi.

Algoritma: Jurnal Ilmu Komputer Informatika **e-ISSN 2598-6341**
Volume: 9, Nomor: 1, April 2025: 72-83
DOI: 10.30829/algoritma.v9i2.25017
https://jurnal.uinsu.ac.id/index.php/algoritma

2. INTRODUCTION

2.1 Klasterisasi Data

Klasterisasi adalah metode analisis data yang digunakan untuk mengelompokkan objek berdasarkan kesamaan karakteristiknya. Teknik ini berguna dalam eksplorasi data, segmentasi pasar, analisis demografi, dan berbagai aplikasi lainnya. Beberapa pendekatan klasterisasi yang umum digunakan adalah:

- K-Means Clustering (berbasis centroid).
- Hierarchical Clustering (berbasis hirarki)
- DBSCAN (berbasis kepadatan data) [4].

2.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) adalah teknik reduksi dimensi yang digunakan untuk mengubah variabel yang saling berkorelasi menjadi variabel baru yang tidak berkorelasi. Prinsip utama PCA melibatkan:

1. Transformasi variabel menjadi komponen utama yang merepresentasikan variabilitas data.
2. Pemilihan komponen dengan nilai eigen tertinggi untuk mempertahankan informasi yang paling signifikan.
3. Pemetaan ulang data dalam dimensi yang lebih rendah tanpa kehilangan karakteristik utama [5].

2.3 K-Means Clustering

K-Means adalah algoritma klasterisasi berbasis centroid yang bekerja dengan cara berikut:

1. Menentukan jumlah klaster (k) yang akan digunakan dalam pemisahan data.
2. Menginisialisasi centroid secara acak atau berdasarkan metode tertentu.
3. Mengelompokkan data berdasarkan kedekatan dengan centroid.
4. Memperbarui posisi centroid berdasarkan rata-rata titik dalam klaster.
5. Mengulang proses hingga mencapai konvergensi, di mana tidak ada perubahan signifikan dalam pembentukan klaster [6].

2.4 Studi Terkait

Beberapa penelitian sebelumnya telah menggunakan kombinasi PCA dan K-Means untuk klasterisasi data wilayah. Contoh studi terkait meliputi:

1. Klasterisasi Kabupaten di Sumatera Utara berdasarkan indikator ekonomi menggunakan K-Means.
2. Analisis dampak bencana di Indonesia menggunakan integrasi PCA dan K-Means untuk meningkatkan efektivitas pemetaan wilayah.
3. Segmentasi wilayah di Jawa Timur berdasarkan faktor sosial-ekonomi dengan pendekatan PCA-K-Means [7].

3. METODOLOGI PENELITIAN

3.1 Jenis dan pendekatan penelitian

Penelitian ini termasuk dalam kategori penelitian kuantitatif eksploratif, yang bertujuan untuk mengeksplorasi pola dan struktur tersembunyi dalam data multivariat berdasarkan indikator sosial, ekonomi, dan demografi antar wilayah administratif (kabupaten/kota) di Sumatera Utara dan dianalisis menggunakan metode Principal Component Analysis (PCA) untuk reduksi dimensi dan K-Means Clustering untuk pengelompokan [8].

Algoritma: Jurnal Ilmu Komputer Informatika **e-ISSN 2598-6341**
Volume: 9, Nomor: 1, April 2025: 72-83
DOI: 10.30829/algoritma.v9i2.25017
https://jurnal.uinsu.ac.id/index.php/algoritma

3.1.2 Pendekatan yang Digunakan:

1. Pendekatan Klasterisasi (Clustering)

Klasterisasi adalah bagian dari metode unsupervised learning dalam data mining yang bertujuan mengelompokkan data ke dalam beberapa kelompok (klaster) berdasarkan kesamaan karakteristik antar objek.

2. Principal Component Analysis (PCA)

PCA adalah teknik reduksi dimensi yang digunakan untuk mengubah sejumlah besar variabel menjadi sejumlah kecil principal components yang masih mempertahankan sebagian besar variasi data. PCA sering digunakan sebelum klasterisasi untuk meningkatkan efisiensi dan visualisasi data berdimensi tinggi.

3. K-Means Clustering

K-Means adalah algoritma klasterisasi partisi yang membagi data menjadi k klaster berdasarkan minimisasi jarak antar data dan pusat klaster (centroid). Metode ini banyak digunakan untuk eksplorasi data karena kesederhanaannya dan efisiensinya [9].

3.2 Sumber Data

Data yang digunakan dalam penelitian ini diperoleh dari situs Kaggle, yang merupakan platform terbuka untuk berbagi dan mengakses dataset. Dataset diunduh dari:

URL Dataset: https://www.kaggle.com/code/wahyuikbalmaulana/analisa-clustering-provinsi-sumatera-indonesia

Dataset tersebut berisi data Analisa Clustering Provinsi Sumatera Indonesia dan indikator pembangunan lainnya dari berbagai kabupaten/kota di Indonesia, yang kemudian difilter untuk wilayah provinsi di Pulau Sumatera Utara. Seluruh atribut numerik dari dataset ini menjadi bahan dasar untuk analisis PCA.

3.3 Tahapan Penelitian

Berikut merupakan alur metodologi yang digunakan dalam penelitian ini:

1. Pengumpulan Data:
   - Mengunduh data dari Kaggle
   - Memfilter data hanya untuk kabupaten/kota di Provinsi Sumatera Utara
2. Reduksi Dimensi dengan PCA:
   - Menurunkan dimensi data dari banyak atribut menjadi 2 komponen utama (PCA1 dan PCA2).
   - Output berupa data 3 dimensi seperti yang ditampilkan pada tabel PCA.
3. Penentuan Jumlah Klaster (k):
   - Menggunakan Metode Elbow untuk menentukan nilai optimal dari k
   - Berdasarkan grafik Elbow (Gambar 2 dan 3), nilai optimal diperoleh pada k = 4
4. K-Means Clustering:
   - Mengelompokkan kabupaten/kota berdasarkan dua komponen utama PCA.
   - Visualisasi klaster ditunjukkan dalam scatter plot 4 klaster (Gambar 4) .

3.4 Alat dan Perangkat

- Bahasa Pemrograman: Python.
- Library: pca dan k-means.
- Visualisasi dilakukan dengan library matplotlib.

3.5 Visualisasi dan Hasil Sementara

Berikut ini adalah beberapa visualisasi dan hasil awal dari tahapan metode:

Algoritma: Jurnal Ilmu Komputer Informatika **e-ISSN 2598-6341**
Volume: 9, Nomor: 1, April 2025: 72-83
DOI: 10.30829/algoritma.v9i2.25017
https://jurnal.uinsu.ac.id/index.php/algoritma

- Tabel Hasil PCA: Menampilkan nilai PCA1 dan PCA2 dari masing-masing kabupaten/kota.
- Grafik Elbow Method: Menentukan jumlah klaster optimal (k = 4).
- Plot Klasterisasi: Menunjukkan pembagian klaster berdasarkan PCA1 dan PCA2.

4. HASIL DAN PEMBAHASAN

Gambar dibawah adalah tabel data mentah dari kabupaten/kota di Provinsi Sumatera Utara, yang memuat sejumlah indikator sosial, ekonomi, dan layanan publik. Data ini digunakan sebagai dasar dalam proses klasterisasi menggunakan PCA dan K-Means.

	Kabupaten Kota	KPM	Sanitasi	Minum	Keluhan	Miskin	Kerja	Negeri	Swasta	Nganggur
0	Nias	13962	19.93	47.79	27.37	16.82	81.79	8	3	3.12
1	Mandailing Natal	27257	35.73	73.78	19.77	9.49	69.79	21	3	6.12
2	Tapanuli Selatan	18175	46.41	67.39	20.29	8.80	74.38	10	2	4.00
3	Tapanuli Tengah	32322	57.56	68.81	22.70	12.67	75.05	15	9	7.24
4	Tapanuli Utara	23783	83.79	89.06	12.63	9.72	82.63	18	8	1.54
5	Toba	15152	89.54	95.04	15.66	8.99	80.38	13	3	0.83
6	Labuhan Batu	24883	81.50	94.34	13.42	8.74	61.84	16	16	5.66
7	Asahan	39204	89.09	95.78	19.37	9.35	63.02	17	25	6.39
8	Simalungun	55304	91.75	99.74	25.74	8.81	72.55	20	26	4.17
9	Dairi	19634	92.35	91.90	20.92	8.31	85.73	13	11	1.49
10	Karo	19870	84.33	91.43	18.30	8.79	84.56	13	11	1.95
11	Deli Serdang	56092	96.37	98.18	21.39	4.01	66.78	21	111	9.13
12	Langkat	77946	80.76	92.51	24.56	10.12	69.12	18	49	5.12
13	Nias Selatan	21269	13.14	66.21	17.11	16.92	72.25	49	16	3.91
14	Humbang Hasundutan	19791	91.65	91.95	13.62	9.65	84.17	12	3	1.94
15	Pakpak Bharat	4013	90.14	70.69	24.55	9.35	87.70	5	2	1.36
16	Samosir	12580	91.09	65.64	11.58	12.68	84.38	8	5	0.70
17	Serdang Bedagai	36192	93.19	98.14	27.13	8.30	66.75	18	21	3.93
18	Batu Bara	39369	88.04	97.83	27.64	12.38	70.00	7	16	6.62
19	Padang Lawas Utara	10487	67.17	77.58	16.31	9.92	76.82	9	2	3.19
20	Padang Lawas	12420	59.62	77.84	23.40	8.69	75.23	8	2	4.07
21	Labuhanbatu Selatan	10973	84.85	84.66	22.86	8.53	66.38	10	7	4.71
22	Labuanbatu Utara	16567	79.75	86.75	27.35	10.02	65.73	9	8	5.71
23	Nias Utara	15977	46.09	58.17	31.03	25.66	74.27	13	2	3.00
24	Nias Barat	10498	38.02	71.52	21.39	26.42	82.08	13	2	0.74
25	Sibolga	5619	32.33	92.40	23.60	12.33	71.19	4	5	8.72
26	Tanjungbalai	15255	89.07	87.20	30.16	13.40	66.57	7	4	6.59
27	Pematangsiantar	14536	88.49	99.78	17.12	8.52	68.80	6	21	11.00
28	Tebing Tinggi	10548	95.88	99.35	19.51	10.30	67.19	4	12	8.37
29	Medan	76401	92.71	98.80	13.93	8.34	62.16	21	198	10.81
30	Binjai	11573	95.21	99.76	10.51	5.81	62.77	7	23	7.86
31	Padangsidimpuan	9594	51.33	54.13	32.02	7.53	68.69	8	10	7.18
32	Gunungsitoli	12608	45.13	74.11	31.76	16.45	62.95	6	5	4.80

Gambar 1. Tingkat pengangguran

Algoritma: Jurnal Ilmu Komputer Informatika **e-ISSN 2598-6341**
Volume: 9, Nomor: 1, April 2025: 72-83
DOI: 10.30829/algoritma.v9i2.25017
https://jurnal.uinsu.ac.id/index.php/algoritma

Gambar ini menunjukkan hasil transformasi Principal Component Analysis (PCA) dari dataset yang sebelumnya Anda tampilkan. Hasil ini merupakan output utama dari proses reduksi dimensi, yang menyederhanakan sekumpulan variabel menjadi dua komponen utama: PCA1 dan PCA2 [10].

	Kabupaten Kota	PCA1	PCA2
0	Nias	-3.399684	1.375081
1	Mandailing Natal	-0.373561	1.202457
2	Tapanuli Selatan	-1.148108	0.094683
3	Tapanuli Tengah	-0.504478	1.026380
4	Tapanuli Utara	-0.233816	-1.723751
5	Toba	-0.295234	-2.102947
6	Labuhan Batu	1.399763	-0.311242
7	Asahan	1.790103	0.249773
8	Simalungun	1.517053	0.110938
9	Dairi	-0.396154	-1.888758
10	Karo	-0.376877	-1.748395
11	Deli Serdang	3.837636	0.862380
12	Langkat	2.048825	1.117868
13	Nias Selatan	-1.438708	2.790993
14	Humbang Hasundutan	-0.281895	-2.209793
15	Pakpak Bharat	-1.827086	-1.829633
16	Samosir	-1.556080	-2.022972
17	Serdang Bedagai	1.226922	0.063910
18	Batu Bara	0.909936	0.222079
19	Padang Lawas Utara	-0.912098	-1.023308
20	Padang Lawas	-0.907232	-0.346060
21	Labuhanbatu Selatan	0.164536	-0.405362
22	Labuanbatu Utara	0.185040	0.232742
23	Nias Utara	-3.117149	2.087201
24	Nias Barat	-3.360220	0.715232
25	Sibolga	-0.560825	0.672612
26	Tanjungbalai	-0.022077	0.396920
27	Pematangsiantar	1.674218	-0.548752
28	Tebing Tinggi	1.100294	-0.808815
29	Medan	5.459015	1.968333
30	Binjai	1.983466	-1.332055
31	Padangsidimpuan	-1.165614	1.408811
32	Gunungsitoli	-1.420112	1.703451

Gambar 2. Visualisasi diagram pencar PCA1 versus PCA2

Gambar diatas yaitu menunjukkan hasil analisis Principal Component Analysis (PCA) terhadap data dari 32 Kabupaten/Kota di Sumatera Utara. Berikut adalah penjelasan masing-masing kolom dan makna umumnya:

Penjelasan pada tabel :

- Kabupaten/Kota : Nama wilayah administratif di Sumatera Utara.
- PCA1 : Nilai dari komponen utama pertama (Principal Component 1).
- PCA2 : Nilai dari komponen utama kedua (Principal Component 2).

Principal Component Analysis (PCA) adalah metode statistik untuk mereduksi dimensi data yang banyak menjadi lebih sedikit, dengan tetap mempertahankan informasi paling penting (variasi terbesar dari data).

- PCA1 dan PCA2 adalah dua sumbu (komponen utama) hasil transformasi data asli.
- Mereka merepresentasikan variasi terbesar pada data asli dalam dua dimensi.
- PCA1 biasanya menjelaskan variasi terbesar, diikuti PCA2 .

1. Wilayah dengan nilai PCA1 dan PCA2 tinggi, seperti:

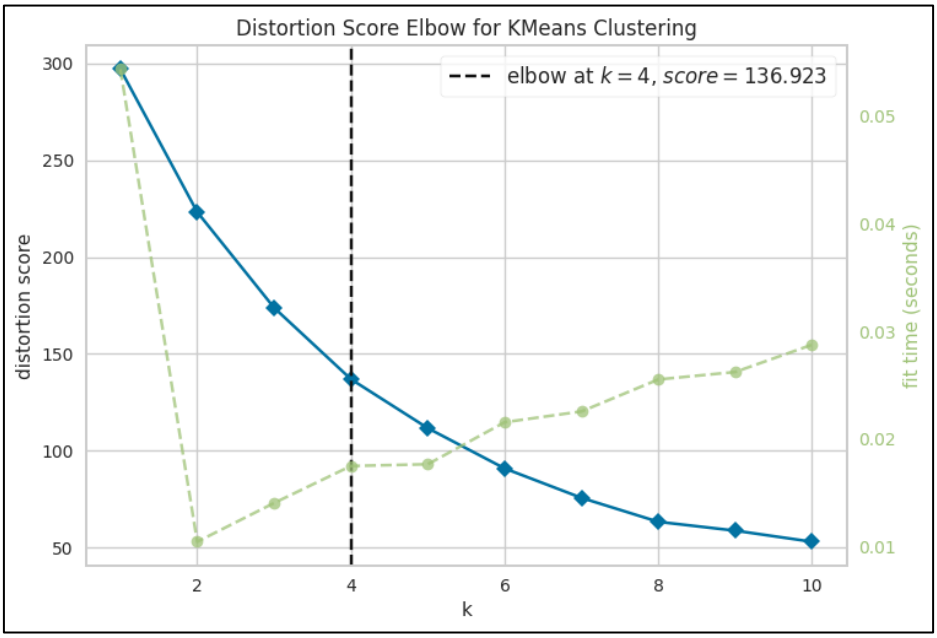
- Medan (5.45, 1.05)
- Deli Serdang (3.83, 0.86)
- Langkat (2.05, 1.18)
- → menunjukkan bahwa mereka memiliki karakteristik yang cukup unik dibanding daerah lain.

Algoritma: Jurnal Ilmu Komputer Informatika **e-ISSN 2598-6341**
Volume: 9, Nomor: 1, April 2025: 72-83
DOI: 10.30829/algoritma.v9i2.25017
https://jurnal.uinsu.ac.id/index.php/algoritma

2. Wilayah dengan nilai PCA1 negatif besar, seperti:
    - Nias (-3.39, 1.37)
    - Nias Barat (-3.38, 1.67)
    - → menunjukkan bahwa mereka cukup jauh dari pusat data dalam arah dimensi pertama (PCA1), bisa jadi karena kondisi yang berbeda ekstrem [11].

3. Wilayah yang dekat ke nol seperti:
    - Tanjungbalai (-0.02, 0.37)
    - Labuhanbatu Selatan (0.16, 0.40)
    - → berarti karakteristiknya cukup rata-rata atau netral, tidak terlalu ekstrim di dimensi manapun.

- Visualisasi: Data bisa divisualisasikan dalam grafik 2D berdasarkan PCA1 dan PCA2 untuk melihat klaster atau outlier.
- Clustering / Klasifikasi: Bisa digunakan untuk pengelompokan kabupaten/kota berdasarkan kemiripan data.
- Pengambilan Keputusan: Pemerintah bisa menggunakan hasil ini untuk membuat kebijakan yang sesuai dengan kelompok [12].



Gambar 3. Analisis setiap kluster

Gambar diatas merupakan grafik Elbow Method untuk menentukan jumlah klaster optimal (k) pada algoritma KMeans Clustering [13].

Metode Siku untuk Clusterin KMeans
Penjelasan Gambar
- Grafik menunjukkan hubungan antara judandistorsi(dua
- Titik siku terlihat di kBahasa Indonesia:35..

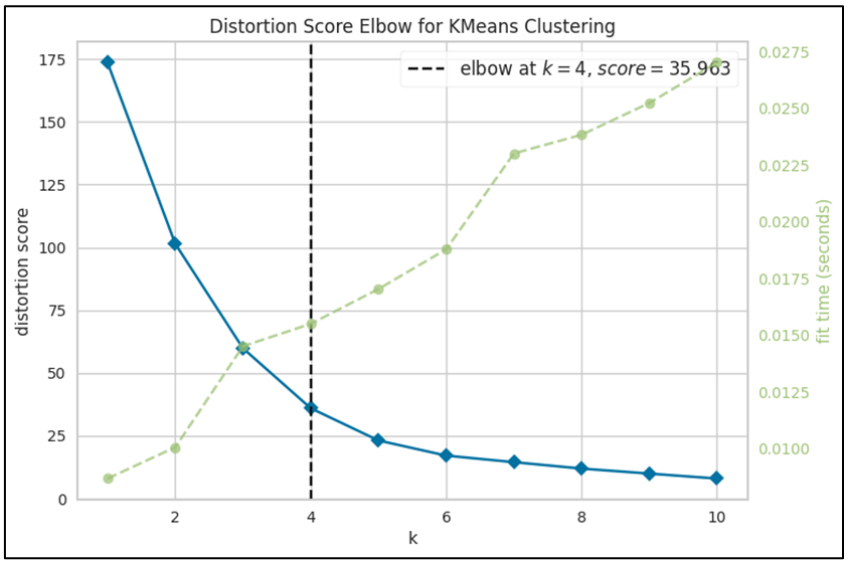
Algoritma: Jurnal Ilmu Komputer Informatika **e-ISSN 2598-6341**
Volume: 9, Nomor: 1, April 2025: 72-83
DOI: 10.30829/algoritma.v9i2.25017
https://jurnal.uinsu.ac.id/index.php/algoritma

- Garis putus-putus hitam menandai titik siku (k
- Garis hijau menunjukkan f(wa

1. Sumbu X
   - Menampilkan jumlah klaster dari 1 sampai 10.
   - k adalah jumlah kelompok/klaster yang digunakan oleh algoritma KMeans.
2. Sumbu Y Kiri: Distortion Score
   - Distortion score menunjukkan seberapa jauh data dari pusat klasternya. Semakin kecil nilainya, semakin baik hasil klastering.
   - Skor ini menurun saat k meningkat karena data dibagi ke dalam lebih banyak kelompok, yang cenderung memperkecil jarak dari pusat.
3. Sumbu Y Kanan (Hijau): Fit Time (seconds)
   - Waktu yang dibutuhkan untuk melakukan training model dengan jumlah klaster tertentu (k).
   - Digambarkan dengan garis putus-putus berwarna hijau muda.
   - waktu komputasi saat nilai k bertambah.
4. Garis Vertikal Hitam Putus-Putus:
   - Menunjukkan titik "elbow" atau titik optimal jumlah klaster yaitu pada k.
   - Tertulis:elbow at k = 4, score = 35.963.
5. Titik Elbow (Optimal k)
   - Ditandai dengan garis vertikal hitam putus-putus pada k = 4.
   - Elbow point adalah tempat di mana penurunan skor distorsi mulai melambat.
   - Pada titik ini, penambahan klaster tidak memberikan peningkatan signifikan dalam kualitas klasterisasi.
   - Skor distorsi di elbow point = 35.963

- Kesimpulan Gambar :
- Jumlah klaster optimal adalah 4.
- Setelah k = 4, penurunan skor distorsi menjadi tidak terlalu signifikan.
- Waktu pelatihan meningkat seiring bertambahnya jumlah klaster [14].

Algoritma: Jurnal Ilmu Komputer Informatika **e-ISSN 2598-6341**
Volume: 9, Nomor: 1, April 2025: 72-83
DOI: 10.30829/algoritma.v9i2.25017
https://jurnal.uinsu.ac.id/index.php/algoritma



Gambar 4. Analisis atau interpretasi isi dari setiap kluster

Gambar ini menunjukkan hasil Elbow Method untuk menentukan jumlah klaster (k) yang optimal dalam algoritma KMeans Clustering, dengan mempertimbangkan distortion score dan waktu pelatihan (fit time).

Distortion Score Elbow for KMeans Clustering

Grafik ini digunakan untuk mengevaluasi seberapa baik hasil klasterisasi dengan berbagai nilai k (jumlah klaster). Grafik ini memperlihatkan bagaimana kualitas hasil klasterisasi dengan metode KMeans berubah seiring bertambahnya jumlah klaster (k). Dua metrik yang ditampilkan adalah distortion score (sumbu Y kiri) dan fit time atau waktu pelatihan (sumbu Y kanan).

Distortion score menggambarkan seberapa dekat titik-titik data terhadap pusat klasternya masing-masing. Pada awalnya, ketika jumlah klaster masih kecil (misalnya k=1 hingga k=3), skor distorsi turun tajam. Artinya, dengan menambah klaster, model menjadi jauh lebih baik dalam mengelompokkan data. Namun, setelah k=4, penurunan skor menjadi lebih lambat dan mulai mendatar. Inilah yang disebut sebagai titik elbow—titik di mana menambah klaster lebih lanjut hanya memberikan sedikit perbaikan pada skor, tetapi dengan biaya komputasi yang lebih tinggi.

Sumbu X (horizontal): k

- Mewakili jumlah klaster dari 1 sampai 10.
- Nilai k adalah parameter utama dalam KMeans yang menentukan berapa banyak kelompok yang akan dibentuk dari data.

Sumbu Y Kiri (vertical): Distortion Score

- Mewakili *distortion score*, yaitu total jarak (biasanya jarak Euclidean) antara titik data dan pusat klasternya.
- Skor ini diwakili oleh garis biru dengan penanda berbentuk belah ketupat.
- Semakin kecil nilai distortion score, semakin baik kualitas klasterisasi.
- Terlihat menurun drastis dari k=1 hingga k=4, kemudian menurun perlahan—ini menunjukkan adanya "elbow point".

Algoritma: Jurnal Ilmu Komputer Informatika **e-ISSN 2598-6341**
Volume: 9, Nomor: 1, April 2025: 72-83
DOI: 10.30829/algoritma.v9i2.25017
https://jurnal.uinsu.ac.id/index.php/algoritma

Titik Elbow (Optimal k)

- Ditandai dengan garis vertikal hitam putus-putus pada k = 4.
- Elbow point adalah tempat di mana penurunan skor distorsi mulai melambat.
- Pada titik ini, penambahan klaster tidak memberikan peningkatan signifikan dalam kualitas klasterisasi.
- Skor distorsi di elbow point = 35.963.

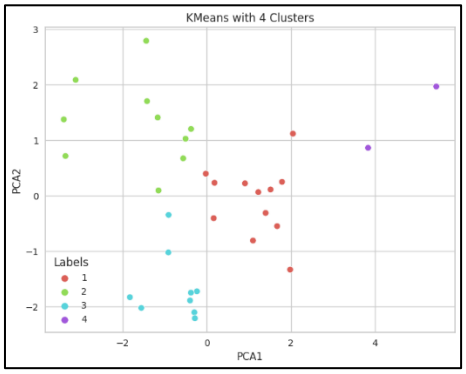
Kesimpulan Gambar :

- Grafik ini menunjukkan bahwa **4 klaster** adalah jumlah yang ideal untuk data yang dianalisis.
- Setelah k = 4, terjadi *diminishing returns* dalam menurunkan distortion score.
- Waktu pelatihan bertambah saat k meningkat, jadi k = 4 memberi keseimbangan antara kualitas klasterisasi dan efisiensi waktu.

Tujuan

Untuk menunjukkan hasil pengelompokan data menjadi 4 cluster secara visual. Dari gambar ini kita bisa lihat bagian tersebut

- Data dibagikan dalam kelompok yang cukup jelas.
- Setiap warna mewakili kategori/kelompok yang diprediksi oleh model.



Gambar 5. Membantu memahami distribusi spasial antara kluster

Penjelasan Elemen pada Gambar:

- "KMeans with 4 Clusters" mengindikasikan bahwa hasil klasterisasi ini menggunakan 4 klaster (sesuai hasil dari metode elbow sebelumnya).
- Sumbu X dan Y: Sumbu horizontal adalah PCA1, dan sumbu vertikal adalah PCA2. PCA digunakan untuk mereduksi dimensi data (misalnya dari banyak fitur menjadi hanya 2 komponen utama) agar bisa divisualisasikan dalam 2D tanpa kehilangan terlalu banyak informasi penting.
- Titik-titik berwarna: Setiap titik mewakili satu data. Warna titik menunjukkan klaster mana data tersebut tergolong, berdasarkan hasil dari KMeans. Warna:
- Merah: Label klaster 1
- Hijau: Label klaster 2
- Biru muda: Label klaster 3
- Ungu: Label klaster 4
- Legenda di kiri bawah menunjukkan label numerik dari masing-masing klaster (1, 2, 3, dan 4), dan setiap label dikaitkan dengan warna tertentu [15].

Interpretasi Visual:

- Hasil ini menunjukkan bahwa sebagian besar data berhasil diklasifikasikan ke dalam kelompok yang cukup terpisah dalam ruang dua dimensi. Kita bisa melihat:
- Klaster merah (1) terkonsentrasi di tengah kanan, menunjukkan bahwa data dalam kelompok ini memiliki pola yang saling berdekatan.
- Klaster hijau (2) menyebar di sisi kiri atas, membentuk kelompok yang cukup padat juga.
- Klaster biru muda (3) terkonsentrasi di sisi kiri bawah.
- Klaster ungu (4) hanya terdiri dari dua titik dan terletak agak terisolasi di bagian kanan atas—kemungkinan mewakili *outlier* atau kelompok kecil yang sangat berbeda dari lainnya.

- **Kesimpulan Gambar :**
- Visualisasi ini memperkuat hasil sebelumnya bahwa 4 klaster adalah jumlah yang tepat. Penggunaan PCA membantu memetakan data berdimensi tinggi ke bentuk 2D agar mudah dianalisis. Kita dapat melihat bahwa setiap klaster memiliki ruang yang relatif terpisah, menandakan pemisahan yang baik oleh algoritma KMeans.

REFERENCES

[1] A. I. Silitonga, Z. A. Nabila, C. Rizkia, Z. Lubis, and N. Safitri, "KLASTERISASI GIZI BURUK DAN STUNTING DI PROVINSI SUMATERA UTARA MENGGUNAKAN K-MEANS CLUSTERING," *METHODIKA*, vol. 10, no. 2, pp. 13–18, 2024.

[2] S. R. Nasution, R. F. Sari, and R. Widyasari, "Analisis Klaster dengan Metode K-Means Pada Penyebaran Kasus Covid-19 Berdasarkan Kabupaten/Kota di Sumatera Utara," *G-Tech J. Teknol. Terap.*, vol. 7, no. 3, pp. 1308–1314, 2023, doi: 10.33379/gtech.v7i3.2904.

[3] Edisman Rahul Gonjales Siahaan, "Implementation Of The K-Means Method In Grouping Districts And Cities In North Sumatra On Social Welfare Problems," *J. Artif. Intell. Eng. Appl.*, vol. 1, no. 2, pp. 168–173, 2022, doi: 10.59934/jaiea.v1i2.86.

[4] D. Nasution, D. N. Sirait, I. Wardani, and Dwiyanto, "Optimasi Jumlah Cluster Metode K-Medoids Berdasarkan Nilai DBI Pada Pengelompokkan Data Luas Tanaman Dan Produksi Kelapa Sawit Di Sumatera Utara," *Kumpul. J. Ilmu Komput.*, vol. 9, no. 2, p. 381, 2022.

[5] A. A. Lubis and T. M. Diansyah, "K-Means Cluster as a Reading Interest Analysis Tool in the North Sumatra Provincial Library," *J. Artif. Intell. Eng. Appl.*, vol. 4, no. 2, 2025.

[6] E. Yolanda, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Data Pasien Rehabilitasi Narkoba," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 1, pp. 182-`191, 2023, doi: 10.30865/klik.v4i1.1107.

[7] R. Kurniawan, M. S. Hasibuan, and R. Hasibuan, "Klasterisasi Wilayah Prioritas Vaksin Menggunakan Algoritma K-MeansClustering," *Media Online*, vol. 4, no. 3, pp. 1585–1592, 2023, doi: 10.30865/klik.v4i3.1334.

[8] H. Sibarani, Solikhun, W. Saputra, I. Gunawan, and Z. M. Nasution, "Penerapan Metode K-Means Untuk Pengelompokkan Kabupaten/Kota Di Provinsi Sumatera Utara Berdasarkan Indikator Indeks Pembangunan Manusia," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 1, pp. 154–161, 2022, doi: 10.36040/jati.v6i1.4590.

[9] S. U. Tarigan, S. Saniman, and M. Yetri, "Klasterisasi Data Penanganan Dan Pelayanan Kesehatan Masyarakat Menggunakan Algoritma K-Means," *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, vol. 1, no. 3, p. 193, 2022, doi: 10.53513/jursi.v1i3.5223.

[10] W. Andriyani, A. H. Nasyuha, Y. Syahra, and B. Triaji, "Clustering Analysis of Poverty Levels in North Sumatra Province Using the Application of Data Mining with the K-Means Algorithm," *J. Media Inform. Budidarma*, vol. 7, no. 4, p. 1971, 2023, doi: 10.30865/mib.v7i4.6867.

[11] S. E. Wardani, S. Z. Harahap, and R. Muti'ah, "Implementation of the K-Means Method for Clustering Regency/City in North Sumatra based on Poverty Indicators," *Sinkron*, vol. 8, no. 3, pp. 1429–1442, 2024, doi: 10.33395/sinkron.v8i3.13720.

[12] R. I. Syahputri, "Fuzzy C-Means Clustering Technique Analysis of North Sumatra Province's District/City Classification Based on Community Social Welfare Level," *JISTech (Journal Islam. Sci. Technol.*, vol. 9, no. 1, pp. 53–57, 2024.

[13] R. F. Purba, A. A. Panjaitan, L. T. Butar-butar, J. F. E. D. L. Sitorus, R. Rumapea, and I. M. S. S, "IMPLEMNTASI K-MEANS CLUSTERING UNTUK MENENTUKAN TINGKAT BENCANA RAWAN BANJIR DI WILAYAH SUMATERA UTARA," *TAMIKA*, vol. 4, no. 2, pp. 210–215, 2024.

[14] C. J. Silalahi, A. Situmorang, and J. F. Naibaho, "Implementasi Metode K-Means Clustering Untuk Memetakan Daerah Potensial Penghasil Padi di Provinsi Sumatera Utara," *Methotika J. Ilm. Tek. Inform.*, vol. 2, no. 2, pp. 49–57, 2022, [Online]. Available: http://ojs.fikom-methodist.net/index.php/methotika

[15] S. P. Simanjorang and M. Yanti, "Pengelompokan Kabupaten/Kota di Sumatera Utara Menggunakan Algoritma Average Linkage dan K-Means Berdasarkan Indikator Pendidikan," *Emerg. Stat. Data Sci. J.*, vol. 1, no. 3, pp. 406–414, 2023, doi: 10.20885/esds.vol1.iss.3.art46.