Seminar Nasional Sains Data 2024 (SENADA 2024)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Klasterisasi Bencana dan Dampaknya di Indonesia: Evaluasi Metode *K-means* dengan Integrasi PCA

Nandya Rezky Utami[1,2], Setia Pramana[3]
[1]*Program Studi Komputasi Statistik, Politeknik Statistika STIS*
[2,3]*Badan Pusat Statistik*
[1]222011485@stis.ac.id
[3]setia.pramana@stis.ac.id
Corresponding author email: 222011485@stis.ac.id

Abstract: *Indonesia, located at the convergence of four major tectonic plates and possessing an active volcanic belt, is highly vulnerable to various natural disasters, making disaster risk reduction a crucial focus in the sustainable development agenda. This study aims to analyze the impact of disasters across 38 provinces in Indonesia during 2023 using k-means cluster analysis, both with and without the integration of Principal Component Analysis (PCA). The data, sourced from "Indonesia Disaster Data 2023" by the National Agency for Disaster Countermeasure (BNPB), includes variables such as the number of disaster events, casualty counts, house damages, and damages to public facilities. The results indicate that k-means with PCA provides more optimal clustering of provinces. An increase in the silhouette score from 0.65 to 0.81 confirms the effectiveness of PCA in reducing components. Specifically, West Java was identified as having unique characteristics, experiencing the highest number of disasters and damages, and consistently being separated from other clusters in both methodologies. This research highlights the importance of understanding the geographical distribution and frequency of disasters for developing more effective and sustainable disaster risk mitigation policies, as well as supporting efforts to achieve the targets of the Sendai Framework and Sustainable Development Goals (SDGs) related to disaster risk reduction.*
Keywords: *disaster, principal component analysis, cluster analysis*

Abstrak: Indonesia, yang terletak di pertemuan empat lempeng tektonik utama dan memiliki sabuk vulkanik aktif, sangat rentan terhadap berbagai bencana alam. Pengurangan risiko bencana menjadi fokus penting dalam agenda pembangunan berkelanjutan. Penelitian ini bertujuan untuk menganalisis dampak bencana di 38 provinsi Indonesia selama tahun 2023 dengan menggunakan metode analisis klaster *k-means*, baik dengan maupun tanpa integrasi Analisis Komponen Utama (PCA). Data yang digunakan bersumber dari "Data Bencana Indonesia 2023" oleh Badan Nasional Penanggulangan Bencana (BNPB), mencakup variabel jumlah kejadian bencana, jumlah korban, kerusakan rumah, dan kerusakan fasilitas umum. Hasil analisis menunjukkan bahwa *k-means* dengan PCA memberikan pengelompokan provinsi yang lebih optimal. Skor *silhouette* yang meningkat dari 0,65 menjadi 0,81 menegaskan efektivitas PCA dalam mereduksi komponen. Secara khusus, Jawa Barat teridentifikasi sebagai provinsi dengan karakteristik unik, mengalami kejadian bencana dan kerusakan yang paling tinggi, dan secara konsisten terpisah dari klaster lainnya dalam kedua metode. Penelitian ini menyoroti pentingnya memahami distribusi geografis dan frekuensi bencana untuk pengembangan kebijakan mitigasi risiko bencana yang lebih efektif dan berkelanjutan, serta mendukung upaya pencapaian target *Sendai Framework* dan *Sustainable Development Goals* (SDGs) terkait pengurangan risiko bencana.
Kata kunci: bencana, analisis komponen utama, analisis klaster

I. PENDAHULUAN

Lokasi geografis Indonesia terletak di wilayah rentan bencana [1]. Hal ini disebabkan karena Indonesia dikelilingi oleh lempeng tektonik aktif, gunung api, serta posisinya yang dilewati garis khatulistiwa [1]. Akibatnya, Indonesia memiliki kerentanan tinggi terhadap berbagai bencana alam seperti angin puting beliung, banjir, banjir bandang, hingga tanah longsor. [2].

Indonesia sendiri mengalami banyak kejadian bencana dalam beberapa tahun terakhir. Berdasarkan data dari World Bank, pada tahun 2022, Indonesia mencapai posisi ke 12 dari 35 negara sebagai negara dengan risiko tertinggi terhadap bencana [3]. Lebih dari itu, sekitar lebih dari 40 persen penduduk Indonesia terancam oleh risiko bencana ini [3]. Pada tahun 2023 sendiri, terjadi 5.400 kejadian bencana di Indonesia dengan 99,35% bencana disebabkan oleh cuaca dan aliran permukaan [2].

Mengurangi risiko bencana merupakan prioritas utama dalam agenda pembangunan berkelanjutan, khususnya tertera pada *Sendai Framework for Disaster Risk Reduction* 2015-2030 serta melalui beberapa target *Sustainable Development Goals* (SDGs). *Sendai Framework*, yang diadopsi negara anggota Perserikatan Bangsa-Bangsa pada tahun 2015, bertujuan untuk secara signifikan mengurangi baik kerugian nyawa, jumlah orang yang terkena dampak, kerugian ekonomi, serta kerusakan pada infrastruktur akibat bencana [4]. Kerangka kerja ini menekankan pentingnya pemahaman risiko bencana yang lebih baik dan memperkuat kerja sama di semua tingkatan.

Pemahaman tentang kerentanan dan risiko bencana adalah esensial, seperti yang ditunjukkan dalam beberapa studi terkini. Surya Prayoga et al. melakukan pemodelan kerugian yang diakibatkan oleh curah hujan ekstrem dengan metode *Extreme Value Theory* (EVT) dan *Copula* [5]. Temuan ini menyoroti pentingnya model matematik dalam memprediksi kerugian finansial dari bencana alam. Analisis risiko lainnya dilakukan oleh Ulandari et al. dalam memetakan daerah rawan bencana menggunakan algoritma *Locally Scaled Density Based Clustering* (LSDBC) dan *Density-Based Spatial Clustering Algorithm with noise* (DBSCAN) [6]. Sementara itu, faktor kritis lain dalam bencana dipertimbangkan oleh Nooraeni et al. untuk menunjukkan bahwa kesehatan dan keamanan pangan penting dalam manajemen risiko bencana [7]. Tidak hanya menyoroti kejadian bencana dengan dampak langsung, Paramartha et al. telah memanfaatkan data lingkungan dari IQAir untuk pemantauan cuaca sebagai indikator ekonomi hijau Indonesia [8]. Melalui referensi terhadap riset terkait ini, penelitian ini tidak hanya menyoroti pentingnya penilaian dan manajemen risiko bencana tetapi juga mengusulkan metodologi yang dapat meningkatkan pemahaman tentang distribusi dan intensitas bencana di Indonesia.

Pengelompokan data bencana alam di Indonesia sendiri telah dimanfaatkan dalam berbagai penelitian dan berbagai metode. Seperti penelitian yang dilakukan oleh Murdiaty et al. [9], penelitian ini mengelompokkan bencana alam berdasarkan provinsi, bulan, banyaknya korban, dan banyaknya fasilitas yang rusak dari 34 provinsi di tahun 2014-2018. Dengan memanfaatkan metode analisis klaster *k-means*, disimpulkan bahwa Jawa Tengah merupakan provinsi dengan kejadian becana alam paling sering di sepanjang tahun analisis. Pengelompokan bencana juga telah dimanfaatkan di area penelitian Jawa Barat oleh Audy et al. pada tahun penelitian 2017-2022 dengan memanfaatkan algoritma *fuzzy c-means* [10]. Penelitian ini berfokus pada kejadian bencana, korban yang terdampak, dan kerusakan fasilitas yang terjadi pada tahun 2023. Pemanfaatan analisis klaster metode lainnya dimanfaatkan Setiawan et al. untuk mengelompokkan daerah-daerah yang rawan terhadap bencana alam seperti getaran tanah dan gempa bumi menggunakan DBSCAN, *Convolutional Neural Network*, dan *K-Medoids* [11]. Berbeda dengan Murdiaty et al. [9], penelitian ini akan mengklasterisasi provinsi-provinsi di Indonesia berdasarkan jumlah kejadian bencana, jumlah korban, dan kerusakan fasilitas dengan menggabungkan semua variabel dalam satu analisis di 38 provinsi. Dengan begitu, penelitian ini mampu memberikan pandangan menyeluruh terkait perbedaan kejadian dan respon bencana di 38 provinsi di Indonesia.

Penggunaan analisis data dalam konteks bencana ini dapat memainkan peran penting. Salah satu metode yang dapat digunakan dalam pengelompokan provinsi berdasarkan kejadian dan dampak bencana adalah metode *k-means*. Dalam penerapannya, benyaknya variabel menyulitkan interpretasi dari klaster [12]. Salah satu pendekatan alternatif yang dapat digunakan adalah mengurangi dimensi variabel dengan menggunakan *Principal Component Analysis* (PCA). Penelitian lain telah memanfaatkan metode reduksi variabel ini sebelum melakukan analisis klaster [12]-[15]. Untuk itu, penelitian ini bertujuan untuk mengidentifikasi efektivitas dari integrasi PCA sebelum melakukan *k-*

means. Selain itu, dengan mengelompokkan provinsi di Indonesia berdasarkan frekuensi dan dampak bencana, penelitian ini juga bertujuan untuk mengidentifikasi pola risiko dan keparahan bencana yang dapat membantu dalam perencanaan strategis dan operasional dalam manajemen risiko bencana. Dengan begitu, hasil dari penelitian ini diharapkan dapat mendukung pencapaian target yang ditetapkan oleh *Sendai Framework* dan SDGs, serta membantu pemangku kepentingan di Indonesia dalam mengembangkan kebijakan yang lebih efektif dan berkelanjutan dalam menghadapi bencana alam.

II. METODE PENELITIAN

Penelitian ini mengimplementasikan analisis klaster untuk mengevaluasi dampak bencana di 38 provinsi di Indonesia selama tahun 2023. Terdapat dua pendekatan metode yang digunakan, yaitu *k-means* tanpa integrasi *Principal Component Analysis* (PCA) dan dengan integrasi PCA. Analisis ini bertujuan untuk mengidentifikasi pola kejadian bencana dan dampaknya di berbagai provinsi. Pendekatan tanpa PCA menganalisis data dalam bentuk aslinya, sementara pendekatan dengan PCA melibatkan reduksi komponen data untuk menunjukkan struktur data yang paling signifikan.

2.1. Data dan Variabel

Studi ini menggunakan data yang bersumber dari publikasi "Data Bencana Indonesia 2023" oleh Badan Nasional Penanggulangan Bencana (BNPB) [2]. Data mencakup seluruh 38 provinsi di Indonesia dan terdiri dari variabel yang menggambarkan kejadian bencana serta dampaknya pada masyarakat dan infrastruktur. Variabel-variabel ini terbagi menjadi tiga kategori utama:

1. Provinsi
2. Jumlah Kejadian Bencana
3. Dampak Bencana:
   - Jumlah Korban;
   - Kerusakan Rumah; dan
   - Kerusakan Fasilitas Umum.

   Secara lebih rinci, variabel yang digunakan dapat dilihat pada Tabel 1.

Tabel 1. Variabel yang digunakan

Kategori	Variabel	Rincian
Provinsi	x_0	Nama provinsi
Kejadian bencana	x_1	Jumlah kejadian bencana
Jumlah Korban	x_2	Korban meninggal dan hilang
	x_3	Korban luka
	x_4	Korban terdampak
Kerusakan rumah	x_5	Rumah rusak berat
	x_6	Rumah rusak sedang
	x_7	Rumah rusak ringan
	x_8	Rumah terendam
Kerusakan fasilitas umum	x_9	Fasilitas pendidikan
	x_{10}	Fasilitas peribadatan
	x_{11}	Fasilitas kesehatan

2.2 Principal Component Analysis (PCA)

PCA bertujuan untuk menemukan k vektor ortogonal dari n komponen yang paling optimal untuk mewakili data, dengan syarat $k \leq n$ [16]. Dengan cara ini, data asli diproyeksikan ke dalam ruang berdimensi lebih kecil, yang menghasilkan pengurangan komponen. Tahapan dalam mereduksi data menggunakan PCA adalah sebagai berikut [17]:

1. Data dinormalisasikan agar setiap atribut berada dalam rentang yang sama;
2. PCA menghitung k vektor normal yang membentuk basis untuk data yang telah dinormalisasi Vektor-vektor ini adalah vektor-vektor unit yang saling ortogonal dan disebut sebagai komponen utama atau *Principal Components* (PC). Data masukan merupakan kombinasi linear dari komponen utama ini.
3. Komponen utama diurutkan berdasarkan tingkat signifikansi atau kekuatannya secara menurun. Komponen utama berfungsi sebagai sumbu baru bagi data yang memberikan informasi tentang varians.
4. Karena komponen diurutkan berdasarkan signifikansi secara menurun, ukuran data dapat dikurangi dengan menghilangkan komponen-komponen yang memiliki varians lebih rendah.

2.2. K-means Clustering

Algoritma ini menggunakan dasar pengelompokan objek berdasarkan rata-rata klaster terdekat [17]. Tujuan dari metode ini adalah untuk meminimalkan kesalahan yang terjadi saat membagi n objek menjadi k klaster [16]. Secara umum, tahapan dalam melakukan klasterisasi *k-means* adalah [18]:

1. Melakukan standardisasi pada data. Standardisasi dilakukan dengan mengitung skor z. Skor Z dihitung dengan mengurangkan setiap nilai (X) dengan rata-rata (μ) kemudian membaginya dengan standar deviasi (σ) untuk setiap variabel. Secara matematis, standardisasi data dihitung dengan rumus [19]:
$$Z = \frac{X-\mu}{\sigma} \qquad (1)$$
2. Menggunakan matriks data $X = \{x_{ij}\}$ berukuran n × p, perlu menentukan jumlah klaster yang akan dibentuk. Metode *Elbow* dan *silhouette score plot* adalah dua cara yang dapat digunakan untuk menetapkan jumlah klaster optimal. Metode *Elbow* menentukan jumlah klaster optimal dengan mengamati titik di mana penurunan SSE (*Sum of Squared Errors*) mulai melambat, membentuk sebuah siku [20]. SSE dihitung untuk setiap klaster, dan semakin banyak jumlah klaster k, semakin kecil nilai SSE. Rumus SSE pada *k-means* adalah [19]:
$$SSE = \sum_{k=1}^{k} \sum_{x \in S_k} \left\| x_j - c_k \right\|^2 \quad (2)$$
Silhouette score plot, di sisi lain, mengevaluasi setiap pengelompokan dengan menghitung skor silhouette untuk setiap jumlah klaster yang mungkin. Skor *silhouette* untuk setiap pengamatan i dihitung dengan membandingkan jarak rata-rata ke titik terdekat dalam klaster yang sama (a_i) dan jarak rata-rata ke titik-titik dalam klaster terdekat berikutnya (b_i) lalu dibagi dengan nilai maksimum antar keduanya [21]:
$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)} \qquad (3)$$
Semakin tinggi nilai *shilhouette score*, maka klaster yang terbentuk semakin baik.
3. Membagi data secara acak ke dalam klaster
4. Menghitung *centroid* atau pusat klaster dari data dengan persamaan [18]:
$$C_{kj} = \frac{x_{1j} + x_{2j} + \cdots + x_{nj}}{n} \qquad (4)$$
dengan C_{kj} merupakan pusat klaster ke-k pada variabel ke-j (j = 1, 2, …, p) dan n adalah jumlah data pada klaster ke-k
5. Menghitung jarak setiap objek x ke setiap centroid y dengan menggunakan jarak *Euclidian* [18]:

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{p}|x_i - y_i|^2} \qquad (5)$$

6. Setiap objek ditempatkan ke dalam klaster dengan pusat klaster terdekat, sehingga kumpulan objek-objek membentuk sebuah klaster.
7. Menetapkan pusat klaster baru dari klaster yang baru terbentuk, dengan pusat klaster baru dihitung sebagai rata-rata dari semua objek yang berada dalam klaster tersebut.
8. Mengulangi langkah-langkah 3 hingga 7 sampai tidak ada lagi perpindahan objek antara klaster.

2.3. Validasi Klaster

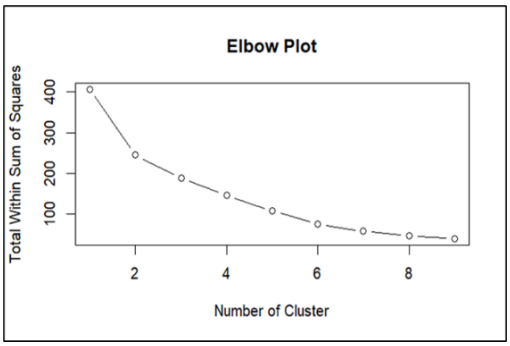
Dalam penelitian ini, validasi klaster dilakukan menggunakan metode validasi internal, khususnya melalui penggunaan skor *silhouette*. Skor *silhouette* yang digunakan adalah skor yang dihitung dengan persamaan (3).

Skor *Silhouette* adalah sebuah nilai dengan rentang dari -1 hingga 1. Apabila nilai mendekati 1 maka pengamatan tersebut cocok dengan klaster tempat objek berada, sedangkan nilai yang semakin dekat ke -1 menandakan pengamatan tersebut seharusnya termasuk dalam klaster lain yang lebih sesuai [16], [21].

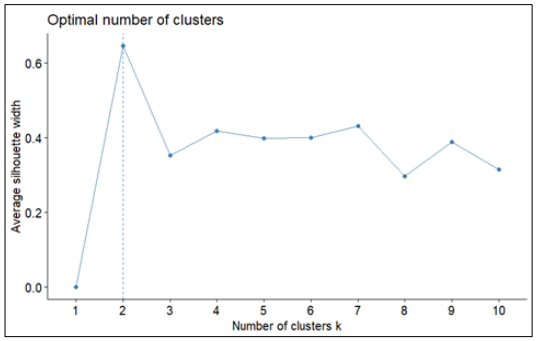
III. HASIL DAN PEMBAHASAN

3.1. Pengelompokan dengan K-means

Pengelompokan pertama dilakukan dengan menerapkan algoritma *k-means* tanpa reduksi komponen dengan PCA. Tahapan pertama dalam algoritma ini adalah menentukan jumlah klaster optimum yang akan digunakan.



Gambar 1a.



Gambar 1b.

Gambar 1. penentuan klaster optimum: (1a) *elbow plot* (1b) *silhouette score plot*

Gambar 1 menunjukkan dua grafik untuk menentukan jumlah klaster optimum dalam analisis *k-means*. *Elbow plot* mulai menunjukkan siku pada klaster ke-2. Begitupun dengan *silhouette score plot* yang menunjukkan skor yang paling tinggi pada 2 klaster. Berdasarkan dua grafik di atas, jumlah klaster yang digunakan adalah 2.

Setelah dilakukan analisis klaster dengan metode *k-means*, didapatkan pengelompokan sebagai berikut:

Tabel 2. Klaster dan anggotanya

Klaster	Provinsi
1	Aceh, Sumatera Utara, Sumatera Barat, Jambi, Sumatera Selatan, Bengkulu Lampung, Kep. Bangka Belitung, Kep. Riau, DKI Jakarta, DI Yogyakarta, Jawa Timur, Banten, Bali, NTB, NTT, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan,

Klaster	Provinsi
2	Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Papua Jawa Barat, Jawa Tengah, Riau

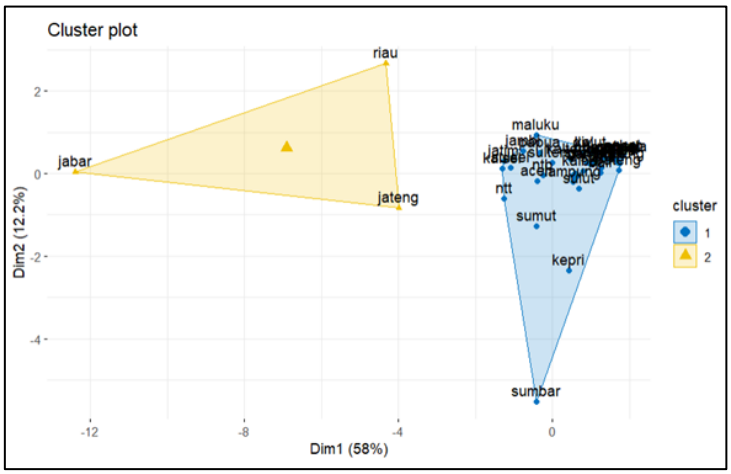
Tabel 2 menunjukkan provinsi-provinsi yang membentuk suatu klaster. Klaster dengan jumlah anggota provinsi terbanyak adalah klaster 1. Artinya, provinsi-provinsi tersebut memiliki kemiripan dalam jumlah kejadian, korban bencana, dan kerusakan fasilitas pada tahun 2023.

Untuk mengetahui karakteristik klaster yang terbentuk, perlu menghitung nilai rata-rata untuk setiap klaster.

Tabel 3. Karakteristik klaster dengan metode *k-means* tanpa PCA

Klaster	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	104,91	7,05	**160,17**	108.655,1	112,71	108,20	445,77	12.731,83	11,94	5,80	2,11
2	**578,33**	**20.33**	63,00	**1,562.786,7**	**308,33**	**7640,00**	**6.887,67**	**103.071,00**	**87,33**	**101,00**	**10,33**

Tabel 3 menunjukkan nilai rata-rata untuk setiap variabel pada masing-masing klaster yang terbentuk. Dapat dilihat bahwa klaster 2 cenderung memiliki nilai yang lebih tinggi untuk setiap variabel kecuali pada variabel x_3 atau korban luka. Hal ini menunjukkan bahwa provinsi Jawa Barat, Jawa Tengah, dan Riau merupakan kelompok provinsi yang cenderung memiliki intensitas kebencanaan, korban jiwa, dan kerusakan fasilitas yang lebih tinggi dibanding provinsi lainnya pada klaster 2.



Gambar 2. Plot Klaster dengan metode *k-means* tanpa PCA

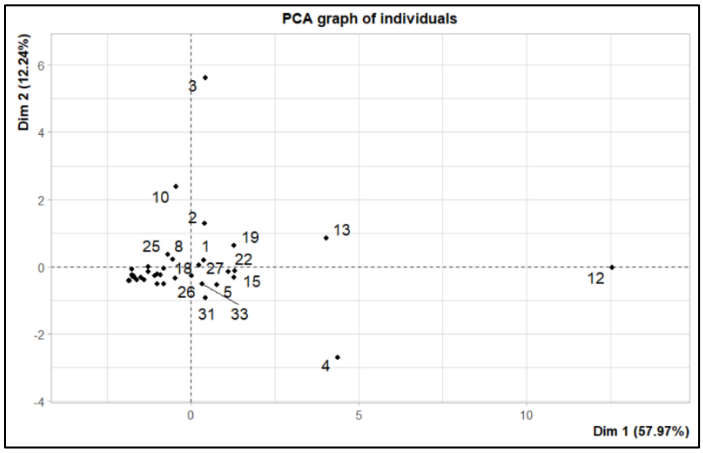
Hasil analisis klaster ditunjukkan pada gambar 2. Gambar tersebut memperlihatkan distribusi setiap provinsi dalam membentuk kalster. Klaster pertama yang disimbolkan dengan warna biru yang cenderung homogen dan memiliki kesamaan karakteristik. Sementara itu, klaster 2 yang memiliki jumlah provinsi lebih sedikit masih mengandung heterogenitas dalam kelompoknya.

Setelah melakukan evaluasi pada dua klaster yang terbentuk dengan metode ini, nilai silhouette score yang didapatkan adalah 0,65. Artinya, kualitas klaster cukup baik karena nilai mendekati 1, di mana observasi dalam klaster cenderung homogen satu sama lain dan heterogen dari klaster lainnya.

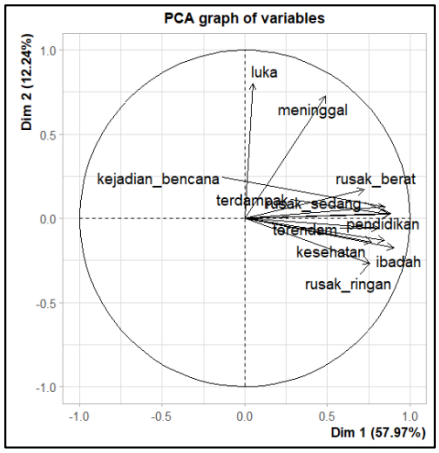
3.2. Principal Component Analysis (PCA)

Untuk melakukan reduksi komponen pada penerapan pengelompokan dengan algoritma k-means, dimanfaatkan metode PCA. Adapun hasilnya dapat dilihat sebagai berikut:

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 3a.



Gambar 3b.

Gambar 3. Grafik hasil analisis PCA: (3a) grafik keterkaitan individu (3b) grafik keterkaitan variabel

Gambar 3a menjelaskan proyeksi observasi dalam hal ini provinsi dalam dua komponen utama yang dihasilkan oleh PCA. Dua komponen utama yang digunakan dalam grafik tersebut adalah komponen utama 1 yang mampu menyumbang 57,97% varians dalam observasi dan komponen utama 2 yang mampu menangkap 12,24% varians. Gabungan kedua komponen utama ini menjelaskan 70,21% varians dalam data.

Plot menunjukkan adanya beberapa kelompok atau *outlier* yang jelas. Plot di atas menampilkan kode provinsi. Dalam hubungannya antara komponen utama pertama dan kedua, terlihat jelas outlier pada observasi 3 dan 12. Observasi tersebut merupakan Provinsi Sumatra Barat dan Jawa Barat yang terpisah jauh dari kelompok utama. Sementara itu, observasi-observasi yang mengumpul merupakan provinsi-provinsi yang memiliki kesamaan dalam karakteristik komponen utama.

Gambar 3b menunjukkan vektor dari variabel yang menunjukkan kekuatan pengaruh variabel terhadap komponen utama dalam hal ini komponen utama 1 dan 2. Variabel rumah rusak berat, rumah rusak sedang, dan rumah rusak ringan memiliki arah yang relatif sama dan berdekatan satu sama lain. Hal ini menunjukkan ketiga variabel saling berkorelasi positif dan lebih memberi pengaruh pada komponen utama 1. Sementara itu, variabel kejadian becana dan terdampak memiliki posisi yang cukup berdekatan yang menunjukkan korelasi positif antar keduanya.

Tabel 4. Distribusi Varians Komponen Utama

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Varians	6,377	1,346	1,100	0,911	0,413	0,387	0,200	0,108	0,086	0,054	0,018
% Varians	57,972	12,241	9,998	8,283	3,754	3,515	1,812	0,982	0,780	0,495	0,160
Kumulatif	57,972	70,212	80,210	88,492	92,247	95,761	97,582	98,564	99,344	99,840	10,000

Tabel 4 menunjukkan nilai variasi, persentase variasi, dan variasi kumulatif dari setiap komponen utama (PC) yang terbentuk. Varians menunjukkan variabilitas data yang mampu dijelaskan oleh setiap komponen utama. Sementara itu, persentase varians memberikan gambaran seberapa penting setiap komponen dalam menjelaskan struktur data. Lebih lanjut, varians kumulatif merupakan total persentase varians yang mampu dijelaskan dari PC1 hingga komponen terkait. Untuk mempertimbangkan berapa banyak komponen utama yang akan digunakan, penelitian ini mempertimbangkan kumulatif varians yang dimiliki oleh komponen utama.

PC1 menjelaskan sebesar 57,972% dari total varians yang artinya, komponen ini mampu menangkap mayoritas informasi dalam data. Kemudian, PC2 dan PC3 memberi kontribusi yang cukup signifikan untuk 12,241% dan 9,998% varians dalam data. Secara kumulatif, penggunaan

komponen utama dari PC1 hingga PC3 telah mampu menjelaskan 80,210% dari total varians. Dengan kata lain, tiga komponen pertama cukup untuk menangkap sebagian besar informasi terkait kejadian dan dampak bencana di Indonesia pada tahun 2023.

3.3. Pengelompokan dengan K-means dan integrasi PCA

Untuk membentuk analisis klaster dengan menggunakan komponen utama yang terbentuk melalui PCA, penelitian ini menerapkan beberapa jumlah komponen utama yang dijelaskan pada bagian sebelumnya. Adapun jumlah klaster optimum yang digunakan dalam analisis *k-means* ditentukan melalui pertimbangan *elbow plot* dan *silhouette score plot*. Untuk membandingkan jumlah komponen utama terbaik yang akan digunakan dalam analisis, digunakan skor *silhouette*. Semakin tinggi skor silhouette maka pengelompokan semakin baik. Hasil analisis terlihat pada tabel 5.

Tabel 5. Jumlah klaster, skor *silhouette*, dan ukuran klaster untuk setiap jumlah PC yang digunakan

Jumlah PC	Jumlah klaster optimum	Skor *silhouette*	Ukuran cluster
2	2	0,81	37, 1
3	2	0,78	37, 1
4	2	0,69	3, 35
5	2	0,67	35, 3

Tabel 5 menunjukkan hasil analisis klaster dengan jumlah komponen utama yang berbeda. Semua percobaan jumlah komponen utama memberikan jumlah klaster optimum sebanyak 2 melalui pertimbangan *elbow plot* dan *silhouette score* plot. Berdasarkan skor *silhouette* pada tabel, semakin banyak jumlah komponen utama yang digunakan, skor *silhouette* semakin berkurang. Skor *silhouette* tertinggi ditunjukkan oleh klaster yang dibentuk oleh 2 komponen utama dengan skor 0,81. Untuk itu, analisis klaster selanjutnya menggunakan 2 komponen utama. Adapun klaster yang terbentuk memberikan 37 provinsi pada klaster pertama dan 1 provinsi di klaster kedua.

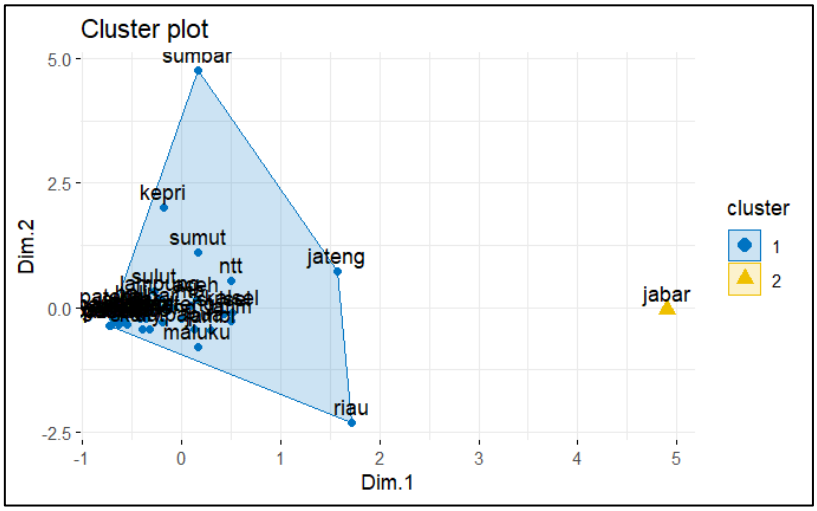
Tabel 6. Karakteristik klaster dengan metode *k-means* dengan PCA

Klaster	PC1	PC2
1	-0,339	0,000946
2	12,549	-0,035

Tabel 6 menunjukkan karakteristik klaster yang terbentuk berdasarkan dua komponen utama (PC1 dan PC2). Berdasarkan tabel tersebut, rata-rata PC1 memiliki nilai yang sangat rendah pada klaster 1 dibandingkan dengan klaster 2. Nilai rata-rata PC2 pada klaster 1 sebesar 0,000946 memiliki nilai yang lebih tinggi dibanding klaster 2. Hal ini menunjukkan bahwa klaster 1 memiliki posisi yang sangat dekat pada titik pusat *origin* dalam dua komponen utama. Sementara itu, klaster 2 berada jauh dari pusat di PC1 dan dekat dengan pusat di PC2.

Nilai ini mengindikasikan bahwa klaster pertama mencakup provinsi-provinsi yang memiliki karakteristik lebih homogen atau seragam dalam hal variabel yang diwakili oleh dua komponen utama ini, sedangkan klaster kedua menunjukkan variabilitas yang lebih tinggi atau karakteristik yang sangat berbeda di komponen utama PC1.

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 4. Plot klaster dengan metode *k-means* dengan PCA

Sesuai dengan tabel 6, gambar 4 menjelaskan distribusi provinsi di Indonesia berdasarkan dua komponen utama. Klaster 1 terlihat membentuk satu kelompok yang mencerminkan nilai PC yang lebih rendah dan distribusi yang lebih homogen. Sementara klaster 2 hanya mencakup satu provinsi, yaitu Jawa Barat yang terletak jauh dari klaster utama. Klaster ini menggambarkan nilai PC yang sangat tinggi untuk PC1.

Berdasarkan tabel 5, nilai *silhouette score* yang dihasilkan oleh pengelompokan *k-means* dengan integrasi PCA dengan 2 klaster adalah 0,81. Artinya, kualitas pengelompokan sangat baik dalam membedakan antar klaster dan homogen antar observasi di dalam klaster. Dibandingkan dengan nilai *silhouette score* pada metode *k-means* tanpa integrasi PCA, metode *k-means* dengan integrasi PCA memberikan nilai yang jauh lebih baik yang artinya pengelompokan juga lebih baik.

IV. KESIMPULAN

Analisis klaster menggunakan *k-means* tanpa dan dengan PCA menunjukkan hasil berbeda, namun keduanya mengidentifikasikan dua klaster sebagai jumlah klaster optimum berdasarkan *elbow plot* dan *silhouette score plot*. Tanpa PCA, klaster terbesar terdiri dari mayoritas provinsi, sementara Jawa Barat, Jawa Tengah, dan Riau membentuk klaster dengan karakteristik kejadian bencana dan kerusakan yang paling tinggi. Sementara itu, dengan PCA, Jawa Barat membentuk klaster sendiri, menegaskan karakteristik uniknya. Penggunaan PCA yang meningkatkan skor *silhouette* dari 0,65 menjadi 0,81, menunjukkan klasterisasi yang lebih optimal dengan reduksi komponen. Jawa Barat konsisten terpisah dalam kedua metode, menandakan perlunya analisis lebih dalam untuk memahami faktor-faktor yang menyebabkannya memiliki tingkat bencana dan kerusakan yang tinggi akibat bencana.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Politeknik Statistika STIS yang telah memberikan dukungan dana penelitian dalam menyelesaikan kajian ini.

REFERENSI

1. T. Yunuarto, S. Pinuji, A. C. Utomo, and I. T. Satrio, *Tanggap Tangkas Tangguh Menghadapi Bencana*, vol. 4. Jakarta Timur: Pusat Data Informasi dan Humas BNPB, 2019. Accessed: Jun.

07, 2024. [Online]. Available: https://bnpb.go.id/storage/app/media/uploads/24/buku-data-bencana/6-buku-saku-cetakan-4-2019.pdf
2. BNPB, *Buku Data Bencana Indonesia 2023*, vol. 3. Jakarta: Pusat Data Informasi dan Komunikasi Kebencanaan BNPB, 2024.
3. World Bank, "Membangun Masa Depan yang Tahan Bencana." World Bank Jakarta, Jakarta, 2010. Accessed: Jun. 07, 2024. [Online]. Available: https://documents1.worldbank.org/
4. BNPB, "Kerangka Kerja Sendai untuk Pengurangan," Jakarta, 2015.
5. I. Surya Prayoga and A. Ahdika, "Pemodelan Kerugian Bencana Banjir Akibat Curah Hujan Ekstrem Menggunakan EVT dan Copula," *Jurnal Aplikasi Statistika dan Komputasi Statistik*, vol. 13, no. 1, pp. 35–46, 2021, Accessed: Jun. 06, 2024. [Online]. Available: https://jurnal.stis.ac.id/index.php/jurnalasks/article/view/273
6. F. Ulandari and R. Kurniawan, "Perbandingan Algoritma LSDBC dan DBSCAN pada Pemetaan Daerah Rawan Kebakaran Hutan (Studi Kasus di Pulau Sumatera, Kalimantan, Sulawesi, dan Papua)," *Jurnal Aplikasi Statistika dan Komputasi Statistik*, vol. 12, no. 2, pp. 25–30, 2020, Accessed: Jun. 08, 2024. [Online]. Available: https://jurnal.stis.ac.id/index.php/jurnalasks/article/view/281
7. R. Nooraeni, N. P. Yudho, and S. Pramana, "Mapping the socio-economic vulnerability in Aceh to reduce the risk of natural disaster," in *AIP Conference Proceedings*, American Institute of Physics Inc., Oct. 2018. doi: 10.1063/1.5062736.
8. D. Y. Paramartha, A. L. Fitriyani, and S. Pramana, "Development of Automated Environmental Data Collection System and Environment Statistics Dashboard," *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 2, pp. 314–325, Jun. 2021, doi: 10.29244/ijsa.v5i2p314-325.
9. Murdiaty, A. Angela, and C. Sylvia, "Pengelompokkan Data Bencana Alam Berdasarkan Wilayah, Waktu, Jumlah Korban dan Kerusakan Fasilitas Dengan Algoritma K-Means," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 3, p. 744, Jul. 2020, doi: 10.30865/mib.v4i3.2213.
10. I. Nabilla Audy, T. Nur Padilah, and B. Nurina Sari, "Pengelompokan Daerah Rawan Bencana Alam di Jawa Barat Menggunakan Algoritma Fuzzy C-Means," *Jurnal Mahasiswa Teknik Informatika*, vol. 7, no. 4, pp. 2799–2803, 2023, Accessed: Jun. 07, 2024. [Online]. Available: https://ejournal.itn.ac.id/index.php/jati/article/view/7205
11. I. N. Setiawan, D. Krismawati, S. Pramana, and E. Tanur, "Klasterisasi Wilayah Rentan Bencana Alam Berupa Gerakan Tanah Dan Gempa Bumi Di Indonesia," *Seminar Nasional Official Statistics*, vol. 2022, no. 1, pp. 669–676, 2022, Accessed: Jun. 09, 2024. [Online]. Available: https://prosiding.stis.ac.id/index.php/semnasoffstat/article/view/1538
12. G. Gunawan, "Disaster Event, Preparedness, and Response in Indonesian Coastal Areas: Data Mining of Official Statistics," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 249–264, Jul. 2024, doi: 10.12785/ijcds/160120.
13. S. Mulyaningsih and J. Heikal, "K-Means Clustering Using Principal Component Analysis (PCA) Indonesia Multi-Finance Industry Performance Before and During Covid-19," *Asia Pacific Management and Business Application*, vol. 011, no. 02, pp. 131–142, Dec. 2022, doi: 10.21776/ub.apmba.2022.011.02.1.
14. G. Rahayu, "Principal Component Analysis untuk Dimensi Reduksi Data Clustering Sebagai Pemetaan Persentase Sertifikasi Guru di Indonesia," *Seminar Nasional Teknologi Informasi, Komunikasi dan Industri (SNTIKI)*, vol. 9, pp. 201–208, 2017, Accessed: Jun. 08, 2024. [Online]. Available: https://ejournal.uin-suska.ac.id/index.php/SNTIKI/article/view/3265
15. N. M. Noor Mathivanan, N. A. MdGhani, and R. M. Janor, "A comparative study on dimensionality reduction between principal component analysis and k-means clustering," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 2, pp. 752–758, 2019, doi: 10.11591/ijeecs.v16.i2.pp752-758.
16. S. Pramana, B. Yuniarto, S. Mariyah, I. Santoso, and R. Nooraeni, *Data mining dengan R konsep serta implementasi*. Bogor: In Media, 2018.

17. J. Han, M. Kamber, and J. Pei, *Data Mining Concept and Techniques*, Third Edition. Waltham: Elsevier, 2012.
18. J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, Seventh. Pearson Prentice Hall, 2009.
19. S. Pramana, R. Yordani, R. Kurniawan, and B. Yuniarto, *Dasar-dasar Statistika dengan Software R konsep dan aplikasi*. Bogor: In Media, 2016.
20. T. S. Madhulatha, "An Overview On Clustering Methods," *IOSR Journal of Engineering*, vol. 2, no. 4, pp. 719–725, 2012, Accessed: Jun. 07, 2024. [Online]. Available: https://arxiv.org/abs/1205.1117
21. R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analytics*, Sixth Edition. New Jersey: Pearson Education, 2007.