Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 810−819
DOI 10.47065/bits.v8i2.9667
https://ejurnal.seminar-id.com/index.php/bits

Analisis Disparitas Data Kebencanaan Indonesia sebagai Dasar Mitigasi Berbasis K-Means

Analysis of Indonesian Disaster Data Disparities as a Basis for K-Means-Based Mitigation

Vanisa Amalia Putri, Fathoni*, Yadi Utama

Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Sriwijaya, Palembang, Indonesia
Email: [1]hello.vanisamalia@gmail.com, [2,*]fathoni@unsri.ac.id, [3]yadiutama@unsri.ac.id
Email Penulis Korespondensi: fathoni@unsri.ac.id
Received: **13/04/2026**, *Revised*: **28/04/2026**, *Accepted*: **01/09/2026**, *Published*: **08/09/2026**

Abstrak−Indonesia merupakan negara dengan frekuensi kejadian bencana yang tinggi, namun perbedaan karakteristik antara dataset nasional dan global menimbulkan tantangan dalam analisis kebencanaan yang konsisten. Penelitian sebelumnya umumnya menggunakan satu sumber data, sehingga belum mampu mengidentifikasi perbedaan representasi data secara menyeluruh. Penelitian ini bertujuan untuk menganalisis disparitas representasi data kebencanaan serta mengidentifikasi pola karakteristik wilayah menggunakan pendekatan data mining. Indeks kebencanaan dibangun melalui normalisasi Min-Max, dan algoritma K-Means diterapkan untuk mengelompokkan wilayah berdasarkan indeks dan jumlah kejadian bencana. Jumlah cluster optimal ditentukan menggunakan Elbow Method dan divalidasi dengan Silhouette Score, sementara dataset global digunakan sebagai pembanding. Hasil penelitian menunjukkan bahwa model optimal diperoleh pada K=4 dengan nilai Silhouette Score sebesar 0,610883 yang mengindikasikan kualitas pemisahan cluster yang baik. Sebagian besar wilayah berada pada kategori karakteristik kebencanaan moderat, sementara sebagian kecil menunjukkan pola ekstrem yang tidak sebanding antara frekuensi kejadian dan tingkat keparahan. Perbedaan distribusi antara dataset nasional dan global mengindikasikan adanya variasi mekanisme pelaporan dan cakupan data. Temuan ini menunjukkan bahwa penggunaan satu sumber data berpotensi menghasilkan interpretasi yang bias, sehingga integrasi multi-sumber data menjadi penting untuk meningkatkan akurasi analisis kebencanaan.

Kata Kunci: K-Means; Data Kebencanaan; Clustering; Disparitas Data; Karakteristik Kebencanaan

Abstract−Indonesia experiences a high frequency of disasters; however, differences in characteristics between national and global datasets create challenges for consistent disaster analysis. Previous studies generally rely on a single data source, limiting the ability to capture disparities in data representation comprehensively. This study aims to analyze disparities in disaster data representation and identify regional disaster patterns using a data mining approach. A Disaster Severity Index was constructed using Min-Max normalization, and the K-Means algorithm was applied to cluster regions based on disaster index and event frequency. The optimal number of clusters was determined using the Elbow Method and validated using the Silhouette Score, while a global dataset was used for comparison. The results indicate that the optimal model is achieved at K=4 with a Silhouette Score of 0.610883, indicating good cluster separation. Most regions exhibit moderate disaster characteristics, while a small number show extreme patterns with disproportionate relationships between frequency and severity. Differences between national and global datasets suggest variations in reporting mechanisms and data coverage. These findings demonstrate that relying on a single data source may lead to biased interpretations, highlighting the importance of multi-source data integration to improve the accuracy of disaster analysis.

Keywords: K-Means; Disaster Data; Clustering; Data Disparity; Disaster Characteristics

1. PENDAHULUAN

Indonesia merupakan wilayah dengan tingkat kejadian bencana yang tinggi akibat kondisi geologis dan klimatologis yang kompleks. Indonesia yang berada pada kawasan *Pacific Ring of Fire* memiliki kerentanan tinggi terhadap bencana geologis seperti gempa bumi dan aktivitas vulkanik [1]. Selain itu, Indonesia juga termasuk negara dengan tingkat risiko bencana tinggi secara global [2]. Karakteristik iklim tropis turut berkontribusi terhadap tingginya kejadian bencana hidrometeorologis seperti banjir dan tanah longsor yang tersebar luas di berbagai wilayah [3], [4]. Kondisi ini menunjukkan bahwa pengelolaan dan analisis data kebencanaan menjadi aspek penting dalam mendukung mitigasi bencana yang efektif.

Dalam perspektif sistem informasi geografis dan pengelolaan data kebencanaan, representasi data sangat dipengaruhi oleh sumber, skala spasial, dan mekanisme pelaporan. Dataset global umumnya mengutamakan kejadian dengan dampak besar, sedangkan dataset nasional cenderung mencatat kejadian secara lebih rinci pada tingkat wilayah administratif. Perbedaan ini berpotensi menghasilkan bias representasi spasial dan ketidakkonsistenan informasi kebencanaan. Dataset global memiliki keterbatasan berupa bias pelaporan dan kecenderungan tidak mencatat kejadian berskala kecil yang dapat menyebabkan distorsi representasi risiko [5]. Permasalahan *missing data* juga masih menjadi kendala utama dalam analisis kebencanaan [6]. Selain itu, peningkatan tren kejadian bencana seringkali dipengaruhi oleh peningkatan kualitas pelaporan, bukan peningkatan kejadian aktual [7]. Bias geografis dalam pelaporan global turut menyebabkan ketimpangan representasi antar wilayah [8]. Dalam perspektif sistem, ketidaksiapan integrasi data nasional dan global menjadi hambatan dalam pengembangan sistem pengurangan risiko bencana berbasis data [9]. Pengukuran risiko bencana juga memerlukan integrasi berbagai dimensi seperti *hazard*, *exposure*, dan *vulnerability* yang tidak selalu tercakup dalam satu sumber data [10]. Selain itu, konsistensi spasial dalam representasi data kebencanaan menjadi faktor penting dalam pengembangan database global [11]. Pada tingkat nasional, keterbatasan

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 810−819
DOI 10.47065/bits.v8i2.9667
https://ejurnal.seminar-id.com/index.php/bits

sistem pelaporan dalam merepresentasikan kejadian skala kecil [12] serta ketidakharmonisan antar level pemerintahan [13] semakin memperkuat pentingnya integrasi multi-sumber data dalam analisis kebencanaan [14].

Sejumlah penelitian sebelumnya telah memanfaatkan teknik *data mining* untuk menganalisis data kebencanaan. Algoritma K-Means telah digunakan untuk mengelompokkan wilayah berdasarkan karakteristik bencana [1] dan menunjukkan performa yang baik dalam proses *clustering* data kebencanaan [15]Penggunaan normalisasi seperti Min-Max Scaling juga terbukti meningkatkan kualitas hasil *clustering* [16]. Namun, sebagian besar penelitian tersebut masih berfokus pada satu sumber data, sehingga belum mampu mengidentifikasi perbedaan representasi data antar sistem pencatatan secara komprehensif. Di sisi lain, pendekatan statistik inferensial seperti uji Kolmogorov–Smirnov atau uji paired t-test umumnya digunakan untuk menguji perbedaan distribusi data, tetapi terbatas pada analisis hubungan linier atau distribusi agregat dan tidak mampu menangkap pola karakteristik multidimensi antar wilayah secara eksploratif. Oleh karena itu, pendekatan clustering dipilih dalam penelitian ini karena memungkinkan segmentasi wilayah berdasarkan kemiripan karakteristik data, sehingga lebih sesuai untuk mengidentifikasi pola disparitas yang bersifat kompleks dan tidak linier.

Berdasarkan permasalahan tersebut, terdapat research gap dalam analisis disparitas data kebencanaan lintas sumber, khususnya antara dataset nasional dan global yang memiliki karakteristik pencatatan berbeda. Penelitian ini mengusulkan pendekatan integratif dengan mengombinasikan dataset BNPB sebagai representasi nasional dan GDACS sebagai representasi global untuk menganalisis disparitas representasi data kebencanaan serta mengidentifikasi pola karakteristik wilayah secara lebih komprehensif. Penelitian ini bertujuan untuk menganalisis disparitas data kebencanaan menggunakan pendekatan data mining dengan algoritma K-Means. Kontribusi utama penelitian ini terletak pada integrasi multi-sumber data untuk mengungkap perbedaan sistem pencatatan serta pola karakteristik kebencanaan yang tidak dapat diidentifikasi melalui penggunaan satu dataset saja, sehingga diharapkan dapat mendukung pengambilan keputusan mitigasi bencana berbasis data.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif berbasis data sekunder yang diperoleh melalui teknik dokumentasi. Dataset yang digunakan merupakan data kebencanaan yang berasal dari sumber resmi, yaitu Badan Nasional Penanggulangan Bencana (BNPB) dan Global Disaster Alert and Coordination System (GDACS), yang telah dikompilasi dan distrukturkan melalui platform Kaggle. Penggunaan dataset terstruktur ini bertujuan untuk meningkatkan kemudahan akses, konsistensi format, serta reproducibility penelitian tanpa mengubah sumber asli data. Dataset BNPB merepresentasikan pencatatan kejadian bencana pada tingkat wilayah administratif yang rinci, sedangkan dataset GDACS merepresentasikan kejadian bencana berdampak besar dalam skala global. Penggunaan kedua dataset ini bertujuan untuk mengidentifikasi perbedaan karakteristik pencatatan data serta potensi bias representasi antar sistem [5], [8], [17]. Variabel yang digunakan dalam penelitian ini meliputi tanggal kejadian, lokasi (provinsi/negara), jenis bencana, serta indeks kebencanaan yang telah melalui proses normalisasi. Alur penelitian disusun secara sistematis mulai dari pengumpulan data hingga interpretasi hasil, sebagaimana ditunjukkan pada Gambar 1. Gambar tersebut menggambarkan tahapan penelitian yang terdiri dari data understanding, preprocessing, normalisasi, exploratory data analysis, clustering, hingga tahap evaluasi dan analisis hasil.

Gambar 1. *Flowchart* Alur Penelitian

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 810−819
DOI 10.47065/bits.v8i2.9667
https://ejurnal.seminar-id.com/index.php/bits

2.2 Data Understanding

Tahap data understanding bertujuan untuk memahami struktur data, distribusi variabel, serta karakteristik nilai indeks kebencanaan pada kedua dataset. Proses ini penting karena kualitas dan karakteristik data sangat menentukan validitas analisis, terutama dalam konteks data kebencanaan yang sering mengandung inkonsistensi dan ketidaklengkapan [17]. Selain itu, penggunaan kombinasi dataset nasional dan global mencerminkan pendekatan integratif untuk meningkatkan validitas hasil analisis melalui pemanfaatan multi-sumber data [10], [11].

2.3 Preprocessing Data

Tahap preprocessing dilakukan untuk meningkatkan kualitas data serta memastikan keterbandingan antar dataset. Proses ini meliputi:
a. pembersihan data dari duplikasi dan inkonsistensi
b. penyamaan kategori jenis bencana
c. transformasi data melalui agregasi

Selain itu, dilakukan proses harmonisasi antara dataset BNPB dan GDACS untuk menyelaraskan skala variabel dan unit analisis sehingga dapat dibandingkan secara valid antar wilayah [5], [18]. Agregasi dilakukan dengan menghitung rata-rata indeks kebencanaan dan jumlah kejadian bencana pada setiap wilayah sebagai berikut:

$$IKBprovinsi = \frac{\sum_{i=1}^{n} IKB_i}{n} \quad (1)$$

$$Jumlah_{kejadian} = \sum_{i}^{n} kejadian_i \quad (2)$$

Dimana n adalah jumlah data dalam satu wilayah, IKB_i adalah nilai indeks kebencanaan ke-i, dan $kejadian_i$adalah jumlah kejadian bencana ke-i. Transformasi ini bertujuan untuk menghasilkan representasi data yang lebih stabil [6], [18]. Dataset GDACS tidak melalui agregasi lanjutan karena telah berada pada tingkat representasi wilayah.

2.4 Normalisasi Data

Normalisasi dilakukan untuk menyetarakan skala antar fitur yang digunakan dalam proses clustering, yaitu rata-rata indeks kebencanaan dan jumlah kejadian. Metode yang digunakan adalah Min-Max Scaling dengan persamaan:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

Metode ini dipilih karena mampu mempertahankan hubungan proporsional antar nilai dalam rentang tertentu sehingga memudahkan interpretasi hasil clustering serta mencegah dominasi variabel tertentu dalam perhitungan jarak [7], [19]. Metode normalisasi yang digunakan dalam penelitian ini adalah Min-Max Scaling untuk menyetarakan skala antar fitur yang digunakan dalam proses clustering. Metode ini dipilih karena mampu mempertahankan hubungan proporsional antar nilai serta banyak digunakan dalam pembentukan indeks kebencanaan dan analisis clustering pada penelitian sebelumnya [14], [16].

Pemilihan metode normalisasi dalam penelitian ini disesuaikan dengan kebutuhan standarisasi data numerik agar tidak terjadi dominasi variabel dalam perhitungan jarak pada algoritma K-Means. Dengan demikian, Min-Max Scaling digunakan untuk memastikan setiap variabel memiliki kontribusi yang seimbang dalam proses pengelompokan [7], [19]. Meskipun demikian, hasil analisis dalam penelitian ini diposisikan sebagai pendekatan eksploratif, sehingga interpretasi yang dihasilkan berfokus pada identifikasi pola karakteristik kebencanaan, bukan sebagai representasi risiko yang komprehensif.

2.5 Exploratory Data Analysis (EDA)

Setelah proses normalisasi, dilakukan exploratory data analysis (EDA) untuk memahami distribusi data melalui visualisasi seperti histogram dan boxplot. Tahap ini bertujuan untuk mengidentifikasi pola distribusi, persebaran data, serta potensi outlier sebelum dilakukan clustering. Analisis ini penting untuk memastikan bahwa data telah siap digunakan dalam proses pemodelan [3].

2.6 Penentuan Jumlah Cluster

Penentuan jumlah cluster dilakukan menggunakan metode Elbow dengan menganalisis nilai inertia pada beberapa nilai K. Titik optimal ditentukan pada kondisi ketika penurunan nilai inertia mulai melambat secara signifikan. Berdasarkan hasil pengujian, jumlah cluster optimal dalam penelitian ini adalah empat cluster. Metode ini digunakan untuk mengurangi subjektivitas dalam pemilihan jumlah cluster [1], [20].

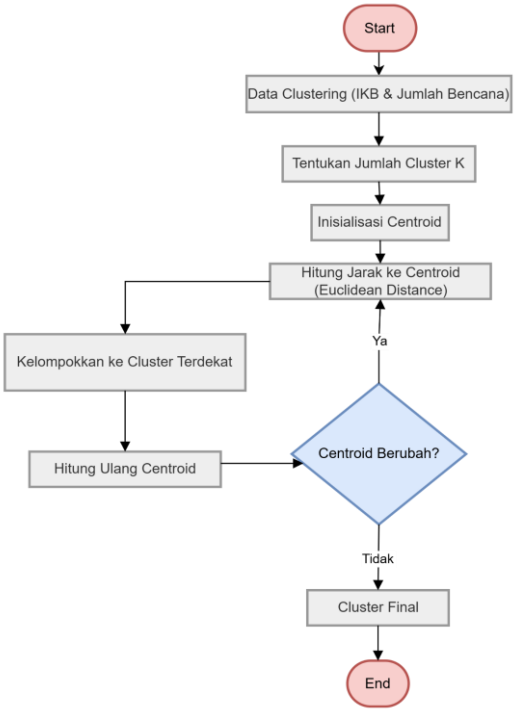
2.7 Clustering K-Means

Clustering dilakukan menggunakan algoritma K-Means dengan dua fitur utama, yaitu indeks kebencanaan dan jumlah kejadian. Perhitungan jarak menggunakan Euclidean Distance:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (4)$$

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 810−819
DOI 10.47065/bits.v8i2.9667
https://ejurnal.seminar-id.com/index.php/bits

Proses K-Means ditunjukkan pada Gambar 2.



Gambar 2. *Flowchart K-Means*

Tahapan algoritma meliputi:
a. penentuan jumlah cluster
b. inisialisasi centroid awal
c. perhitungan jarak data terhadap centroid
d. pengelompokan data berdasarkan jarak minimum
e. pembaruan centroid
f. iterasi hingga centroid stabil

Algoritma ini dipilih karena efektif dalam mengelompokkan data numerik dan menghasilkan cluster yang mudah diinterpretasikan dalam konteks kebencanaan [1], [15], [19]. Pendekatan ini berbeda dengan metode statistik inferensial seperti uji Kolmogorov-Smirnov atau uji *paired t-test* yang berfokus pada pengujian perbedaan distribusi, sedangkan K-Means memungkinkan segmentasi wilayah berdasarkan kemiripan karakteristik multidimensi [1], [15].

Proses clustering ini bertujuan untuk mengelompokkan wilayah berdasarkan kemiripan karakteristik kebencanaan sehingga dapat diidentifikasi pola distribusi data yang terbentuk. Clustering difokuskan pada dataset BNPB sebagai representasi data nasional, sedangkan dataset GDACS digunakan sebagai pembanding untuk menganalisis disparitas representasi data antar sistem.

Dengan demikian, penggunaan K-Means dalam penelitian ini diposisikan sebagai pendekatan eksploratif untuk mengidentifikasi pola karakteristik kebencanaan, bukan untuk menggeneralisasi hubungan kausal atau menguji hipotesis statistik [19].

2.8 Analisis Hasil Clustering

Analisis hasil clustering dilakukan untuk mengidentifikasi karakteristik masing-masing cluster, distribusi wilayah, serta pola karakteristik kebencanaan yang terbentuk. Pendekatan ini memungkinkan identifikasi pola yang tidak dapat diperoleh melalui analisis deskriptif sederhana [21]. Analisis dilakukan berdasarkan dua variabel utama, yaitu indeks kebencanaan dan jumlah kejadian. Pemilihan variabel ini bertujuan untuk merepresentasikan karakteristik dasar kebencanaan secara kuantitatif. Namun demikian, variabel yang digunakan belum mencakup dimensi risiko secara komprehensif seperti *hazard*, *exposure*, dan *vulnerability*, sehingga hasil yang diperoleh lebih bersifat representasi awal dari pola karakteristik kebencanaan.

2.9 Analisis Disparitas Data

Clustering digunakan sebagai pendekatan eksploratif untuk memahami pola karakteristik kebencanaan. Analisis disparitas dilakukan melalui perbandingan distribusi data antara dataset BNPB dan GDACS pada wilayah yang sama, meliputi:
a. distribusi jenis bencana
b. distribusi indeks kebencanaan

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 810−819
DOI 10.47065/bits.v8i2.9667
https://ejurnal.seminar-id.com/index.php/bits

c. pola karakteristik kebencanaan

Perbedaan yang ditemukan mencerminkan adanya bias pelaporan serta perbedaan cakupan data antar sumber [8], [17]. Pendekatan multi-dataset digunakan untuk meningkatkan pemahaman terhadap perbedaan representasi data dan mengurangi potensi bias interpretasi [14].

2.10 Validasi Model

Validasi hasil clustering dilakukan menggunakan Silhouette Score untuk mengukur kualitas pemisahan antar cluster. Persamaan yang digunakan adalah:

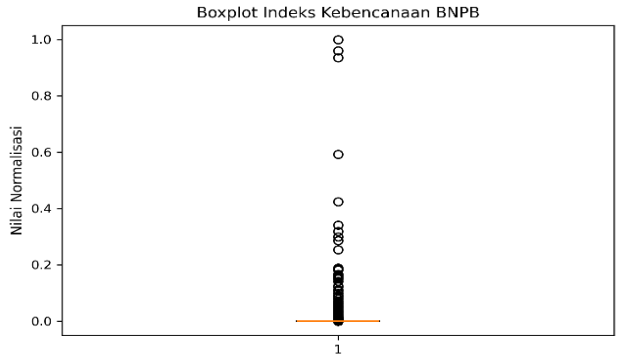
$$s(i) = \frac{b(i)-a(i)}{max(a(i),b(i))} \quad (5)$$

Dimana $a(i)$ adalah rata-rata jarak intra-cluster dan $b(i)$ adalah rata-rata jarak ke cluster terdekat. Hasil pengujian menunjukkan nilai Silhouette Score sebesar 0,610883 yang mengindikasikan bahwa cluster yang terbentuk memiliki kualitas pemisahan yang baik [1], [15], [21].

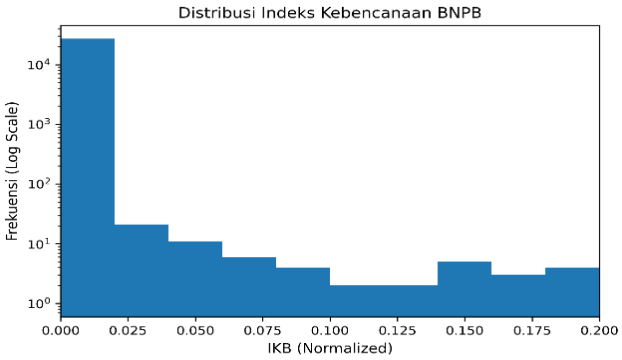
3. HASIL DAN PEMBAHASAN

3.1 Analisis Distribusi Data

3.1.1 Distribusi Data BNPB



Gambar 3. *Boxplot* Indeks Kebencanaan BNPB



Gambar 4. *Histogram* BNPB

Visualisasi distribusi indeks kebencanaan pada dataset BNPB ditunjukkan pada Gambar 3 dan Gambar 4. Sebagian besar nilai berada pada rentang yang sangat rendah dengan konsentrasi mendekati nol, sementara terdapat sejumlah nilai ekstrem yang menyimpang dari distribusi utama. Pola ini menunjukkan bahwa data kebencanaan nasional memiliki distribusi yang tidak merata dan cenderung bersifat *skewed*.

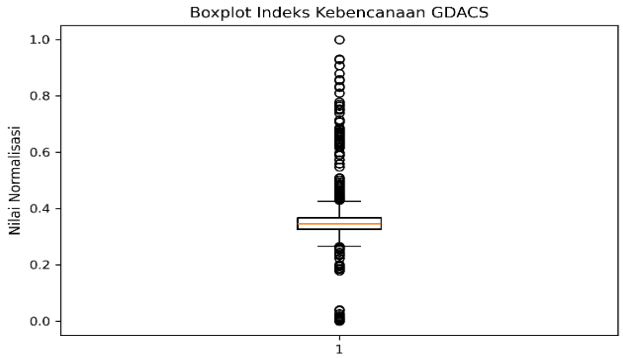
Fenomena ini mencerminkan kondisi umum data kebencanaan, di mana sebagian besar wilayah memiliki tingkat kejadian rendah, sementara sebagian kecil wilayah menunjukkan intensitas yang jauh lebih tinggi. Pola distribusi yang tidak normal ini menunjukkan adanya heterogenitas karakteristik kebencanaan antar wilayah serta ketimpangan spasial yang signifikan [3], [18], [21]. Temuan ini konsisten dengan penelitian sebelumnya yang menyatakan bahwa data kebencanaan cenderung memiliki distribusi tidak merata dan dipengaruhi oleh faktor geografis serta kondisi lingkungan.

Distribusi yang tidak merata ini juga menjadi faktor penting dalam pembentukan cluster, karena algoritma K-Means sangat dipengaruhi oleh kepadatan data. Wilayah dengan nilai ekstrem cenderung terpisah menjadi cluster tersendiri, sementara wilayah dengan nilai rendah terkonsentrasi dalam cluster mayoritas. Hal ini menunjukkan bahwa

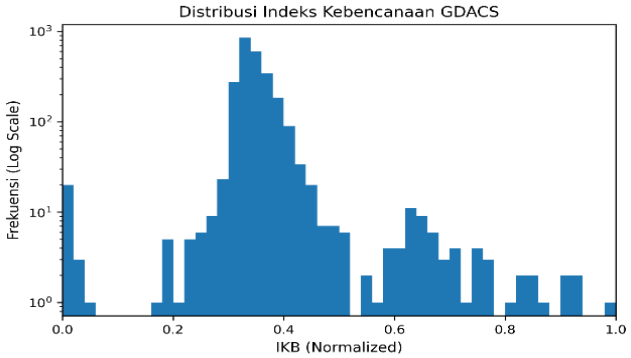
Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 810−819
DOI 10.47065/bits.v8i2.9667
https://ejurnal.seminar-id.com/index.php/bits

struktur distribusi data memiliki peran langsung dalam menentukan hasil segmentasi.

3.1.2 Distribusi Data GDACS



Gambar 5. *Boxplot* Indeks Kebencanaan GDACS

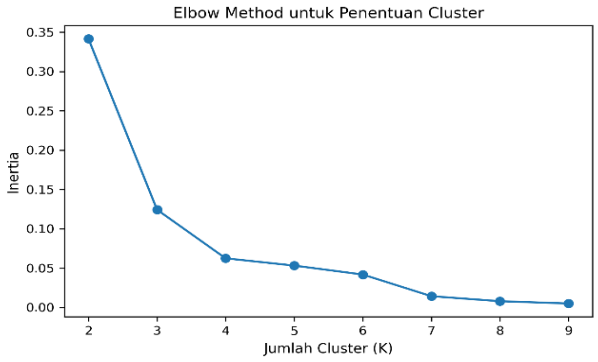


Gambar 6. *Histogram* GDACS

Distribusi indeks kebencanaan pada dataset GDACS (Gambar 5 dan Gambar 6) menunjukkan pola yang lebih terpusat pada rentang nilai menengah dibandingkan dataset BNPB. Meskipun terdapat outlier, distribusi data relatif lebih stabil dan tidak mengalami distorsi signifikan akibat nilai ekstrem.

Hal ini menunjukkan bahwa dataset GDACS memiliki karakteristik pencatatan yang lebih terstandarisasi karena berfokus pada kejadian bencana dengan dampak signifikan dalam skala global. Sistem global cenderung mencatat kejadian besar secara konsisten, sementara kejadian berskala kecil sering tidak terdokumentasi. Temuan ini sejalan dengan karakteristik sistem pelaporan global yang menekankan respons terhadap kejadian darurat berskala besar [8], [22]. Perbedaan pola distribusi antara BNPB dan GDACS mengindikasikan bahwa kedua dataset memiliki representasi yang berbeda terhadap realitas kebencanaan. Hal ini menjadi dasar penting dalam analisis disparitas, karena perbedaan struktur data akan mempengaruhi hasil analisis lanjutan, termasuk clustering.

3.2 Penentuan Jumlah Cluster



Gambar 7. *Elbow Method*

Penentuan jumlah cluster dilakukan menggunakan metode Elbow sebagaimana ditunjukkan pada Gambar 7. Hasil analisis menunjukkan penurunan nilai *inertia* yang signifikan hingga K=4, kemudian kurva mulai melandai. Berdasarkan pola tersebut, jumlah cluster optimal ditetapkan sebanyak empat cluster.

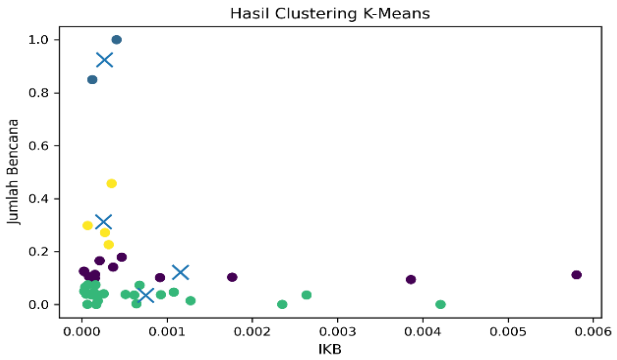
Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 810−819
DOI 10.47065/bits.v8i2.9667
https://ejurnal.seminar-id.com/index.php/bits

Pendekatan ini bertujuan untuk menjaga keseimbangan antara kompleksitas model dan kemampuan representasi data. Selain itu, hasil ini diperkuat melalui evaluasi menggunakan Silhouette Score yang menunjukkan kualitas pemisahan cluster yang baik. Kombinasi metode Elbow dan Silhouette Score memberikan landasan analitis yang lebih kuat dalam menentukan jumlah cluster serta mengurangi subjektivitas dalam proses pemilihan parameter [1], [15], [20].

Penentuan jumlah cluster ini tidak hanya bersifat teknis, tetapi juga mencerminkan struktur alami data. Nilai K=4 menunjukkan bahwa data kebencanaan dapat direpresentasikan ke dalam empat kelompok karakteristik utama, yang kemudian dianalisis lebih lanjut untuk memahami pola kebencanaan antar wilayah.

3.3 Hasil Clustering K-Means

3.3.1 Visualisasi Clustering



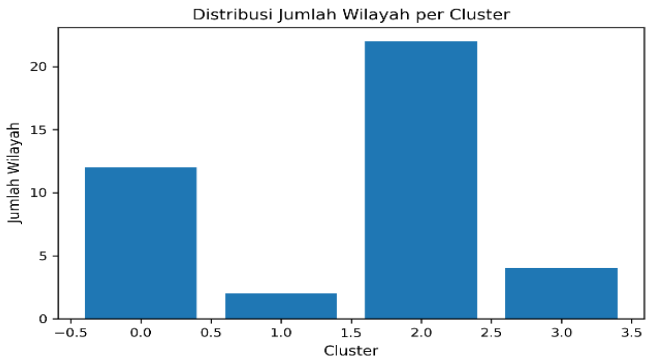
Gambar 8. *Scatter Clustering dan Centroid*

Hasil clustering divisualisasikan dalam bentuk scatter plot pada Gambar 8, di mana setiap warna merepresentasikan cluster yang berbeda. Pemisahan antar cluster menunjukkan adanya variasi karakteristik kebencanaan berdasarkan nilai indeks kebencanaan dan jumlah kejadian bencana. Pola tersebut menunjukkan bahwa algoritma K-Means mampu mengelompokkan data secara efektif berdasarkan kemiripan karakteristik kebencanaan. Hal ini sejalan dengan penelitian sebelumnya yang menyatakan bahwa K-Means efektif dalam mengelompokkan data numerik serta menghasilkan cluster yang mudah diinterpretasikan [1], [15].

Namun demikian, pemisahan cluster tidak hanya mencerminkan perbedaan nilai antar data, tetapi juga dipengaruhi oleh struktur distribusi data yang tidak merata. Dataset kebencanaan yang cenderung bersifat *skewed* menyebabkan sebagian besar wilayah terkonsentrasi pada nilai rendah, sementara hanya sebagian kecil wilayah memiliki nilai ekstrem, sehingga membentuk cluster minoritas. Kondisi ini menunjukkan bahwa hasil clustering tidak hanya dipengaruhi oleh karakteristik variabel, tetapi juga oleh distribusi data itu sendiri [3], [18].

Selain itu, wilayah yang berada pada cluster ekstrem cenderung memiliki nilai indeks kebencanaan tinggi dengan jumlah kejadian yang tidak selalu sebanding. Hal ini mengindikasikan bahwa tingkat keparahan bencana tidak selalu berbanding lurus dengan frekuensi kejadian, melainkan dipengaruhi oleh faktor lain seperti kerentanan dan kapasitas wilayah dalam menghadapi bencana [17], [23].

3.3.2 Distribusi Cluster



Gambar 9. *Distribusi Cluster*

Distribusi jumlah wilayah pada setiap cluster ditunjukkan pada Gambar 9. Hasil menunjukkan adanya ketidakseimbangan jumlah anggota cluster, di mana sebagian besar wilayah terkonsentrasi pada cluster tertentu, sementara cluster lainnya hanya terdiri dari sedikit wilayah. Kondisi ini menunjukkan bahwa distribusi data tidak homogen dan terdapat wilayah dengan karakteristik ekstrem. Cluster mayoritas merepresentasikan kondisi umum

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 810−819
DOI 10.47065/bits.v8i2.9667
https://ejurnal.seminar-id.com/index.php/bits

wilayah dengan karakteristik kebencanaan moderat, sedangkan cluster minoritas mencerminkan wilayah dengan deviasi signifikan terhadap pola umum.

Ketidakseimbangan ini juga mengindikasikan bahwa clustering lebih sensitif dalam mengidentifikasi anomali dibandingkan distribusi umum data. Oleh karena itu, interpretasi hasil clustering perlu mempertimbangkan dominasi cluster mayoritas serta keterbatasan representasi pada cluster minoritas [3], [21].

3.4 Evaluasi Model

Evaluasi model clustering dilakukan menggunakan Silhouette Score dengan nilai sebesar 0,610883. Nilai ini menunjukkan bahwa struktur cluster memiliki tingkat separasi yang baik, sehingga setiap cluster memiliki karakteristik yang relatif berbeda satu sama lain. Namun demikian, nilai ini tidak hanya menunjukkan kualitas pemisahan cluster, tetapi juga mencerminkan bahwa struktur data memiliki pola yang cukup jelas untuk dipisahkan. Data dengan distribusi yang tidak merata cenderung menghasilkan nilai separasi yang tinggi karena adanya perbedaan yang signifikan antara kelompok data.

Oleh karena itu, validitas model dalam penelitian ini perlu diinterpretasikan dalam konteks analisis eksploratif, di mana hasil clustering digunakan untuk memahami pola karakteristik data, bukan sebagai model prediktif yang bersifat generalisasi [1], [15].

3.5 Analisis Karakteristik Cluster

Hasil clustering merepresentasikan karakteristik kebencanaan berdasarkan indeks kebencanaan dan jumlah kejadian, namun tidak secara langsung mencerminkan risiko bencana secara komprehensif. Secara konseptual, risiko bencana dipengaruhi oleh interaksi antara *hazard*, kerentanan, dan kapasitas wilayah, yang tidak sepenuhnya tercakup dalam variabel penelitian [8], [17].

Tabel 1. Statistik Rata-rata Karakteristik Cluster

Variabel	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Rata-rata Indeks Kebencanaan	0,000746	0,000265	0,000230	0,001089
Rata-rata Jumlah Bencana	0,034490	0,924595	0,342282	0,129946

Berdasarkan Tabel 1, setiap cluster menunjukkan karakteristik yang berbeda. Cluster 1 didominasi oleh wilayah dengan frekuensi kejadian bencana yang tinggi, sedangkan cluster 3 mencerminkan wilayah dengan tingkat keparahan bencana yang lebih tinggi. Di sisi lain, cluster 0 merepresentasikan wilayah dengan karakteristik kebencanaan rendah, sementara cluster 2 berada pada kategori menengah.

Perbedaan ini menunjukkan bahwa hubungan antara frekuensi kejadian dan tingkat risiko tidak bersifat linear. Wilayah dengan frekuensi tinggi tidak selalu memiliki tingkat keparahan tinggi, dan sebaliknya. Hal ini mengindikasikan adanya faktor lain seperti kerentanan wilayah, kapasitas mitigasi, dan kondisi geografis yang mempengaruhi dampak bencana. Dengan demikian, clustering dalam penelitian ini tidak hanya berfungsi sebagai teknik pengelompokan data, tetapi juga sebagai alat untuk mengidentifikasi pola hubungan antar variabel yang tidak dapat diamati secara langsung melalui analisis deskriptif sederhana [17], [23].

3.6 Analisis Disparitas Data

Perbandingan antara dataset BNPB dan GDACS menunjukkan adanya perbedaan signifikan dalam karakteristik distribusi data. Dataset BNPB memiliki distribusi yang tidak merata dengan dominasi nilai rendah dan outlier ekstrem, sedangkan dataset GDACS lebih stabil dan terpusat pada nilai menengah [22].

Perbedaan ini tidak hanya menunjukkan variasi nilai, tetapi juga mencerminkan perbedaan sistem pencatatan data. Dataset BNPB bersifat lebih granular dan mencatat kejadian secara detail, sementara GDACS berfokus pada kejadian berdampak besar dalam skala global. Perbedaan struktur data ini mempengaruhi hasil clustering. Data BNPB yang bersifat *skewed* lebih sensitif terhadap nilai ekstrem, sedangkan GDACS menghasilkan pola yang lebih stabil. Hal ini menunjukkan bahwa kualitas dan struktur data memiliki pengaruh signifikan terhadap hasil analisis clustering [5], [7], [19].

Dengan demikian, disparitas yang ditemukan dalam penelitian ini tidak hanya berasal dari perbedaan nilai data, tetapi juga dari perbedaan pendekatan pencatatan antara sistem nasional dan global. Oleh karena itu, penggunaan satu sumber data saja berpotensi menghasilkan interpretasi yang bias.

3.7 Pembahasan

Hasil penelitian menunjukkan bahwa analisis kebencanaan sangat dipengaruhi oleh kualitas dan integrasi data. Disparitas antara dataset BNPB dan GDACS disebabkan oleh perbedaan mekanisme pelaporan, cakupan data, dan tujuan sistem, di mana BNPB bersifat lebih granular, sedangkan GDACS berfokus pada kejadian berdampak besar [5], [7], [8], [9], [14].

Dalam konteks global, integrasi multi-sumber data menjadi penting karena risiko bencana bersifat multidimensional, sehingga pendekatan berbasis dua variabel dalam penelitian ini masih bersifat eksploratif [10], [19], [24]. Hasil penelitian ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa algoritma K-Means efektif dalam mengelompokkan data kebencanaan [1], [15], namun penelitian ini memberikan kontribusi tambahan melalui

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 810−819
DOI 10.47065/bits.v8i2.9667
https://ejurnal.seminar-id.com/index.php/bits

integrasi multi-sumber data yang mampu mengungkap disparitas representasi data yang tidak terlihat pada pendekatan *single dataset*.

Selain itu, algoritma K-Means mampu mengidentifikasi pola awal karakteristik kebencanaan, namun hasilnya sensitif terhadap distribusi data dan tidak merepresentasikan risiko secara komprehensif. Oleh karena itu, hasil clustering perlu diinterpretasikan secara hati-hati dan tidak dapat digeneralisasi secara absolut [6], [7], [8]. Dalam konteks ini, analisis yang dilakukan diposisikan sebagai pendekatan eksploratif yang bertujuan untuk mengidentifikasi pola representasi data berdasarkan karakteristik distribusi, bukan sebagai pengujian perbedaan statistik antar dataset.

4. KESIMPULAN

Berdasarkan hasil penelitian, terdapat disparitas signifikan antara dataset kebencanaan nasional BNPB dan dataset global GDACS yang dipengaruhi oleh perbedaan cakupan data, mekanisme pelaporan, serta karakteristik sistem pencatatan. Penerapan algoritma K-Means menghasilkan empat cluster optimal dengan nilai Silhouette Score sebesar 0,610883 yang menunjukkan kualitas pemisahan cluster yang baik, di mana sebagian besar wilayah berada pada kategori karakteristik kebencanaan moderat, sementara sebagian kecil menunjukkan pola ekstrem yang mencerminkan ketidakseimbangan antara frekuensi kejadian dan tingkat keparahan. Hasil analisis juga menunjukkan bahwa struktur distribusi data yang tidak merata *skewed* berpengaruh terhadap pembentukan cluster, sehingga hasil clustering lebih tepat dipahami sebagai pendekatan eksploratif dalam mengidentifikasi pola karakteristik kebencanaan. Perbedaan distribusi antara kedua dataset menegaskan bahwa penggunaan satu sumber data berpotensi menghasilkan interpretasi yang bias, sehingga integrasi data lintas sumber menjadi kebutuhan penting dalam analisis kebencanaan. Oleh karena itu, diperlukan pendekatan integratif melalui standarisasi format data, penyelarasan skala variabel, serta pengembangan sistem data kebencanaan yang terintegrasi berbasis *multi-source data* untuk meningkatkan konsistensi dan akurasi analisis. Selain itu, penguatan dimensi analisis melalui integrasi variabel risiko seperti *hazard*, *vulnerability*, dan *exposure* dapat meningkatkan representasi karakteristik kebencanaan secara lebih komprehensif serta mendukung pengambilan keputusan mitigasi bencana yang lebih tepat dan berbasis data.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: V.A.P., F., dan Y.U.;
Metodologi: V.A.P., F.;
Perangkat Lunak: V.A.P.;
Validasi: V.A.P., F., dan Y.U.;
Analisis Formal: V.A.P., F.;
Investigasi: V.A.P., Y.U.;
Sumber Daya: V.A.P., F., dan Y.U.;
Kurasi Data: V.A.P.;
Penulisan Draf Awal: V.A.P., Y.U.;
Penulisan, Peninjauan, dan Penyuntingan: V.A.P., F., dan Y.U.;
Visualisasi: V.A.P., F., dan Y.U.;
Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis pertama.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan eksternal.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable).

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable).

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 810−819
DOI 10.47065/bits.v8i2.9667
https://ejurnal.seminar-id.com/index.php/bits

Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.

REFERENCES

[1] A. Wibowo, N. Rohman, Rusdah, A. H. Achadi, and I. Amri, "Clustering Indonesian Provinces by Disaster Intensity using K-Means Algorithm: a Data Mining Approach," *Disaster Adv.*, vol. 17, no. 12, pp. 1–8, Dec. 2024, doi: 10.25303/1712da0108.

[2] World Economic Forum, *The Future of Jobs Report 2023*. Geneva, Switzerland: World Economic Forum, 2023. [Online]. Available: https://www.weforum.org/reports/the-future-ofjobs-report-2023/

[3] M. A. Y. Pratama, A. R. Hidayah, and T. Avini, "Clustering K-Means Untuk Analisis Pola Persebaran Bencana Alam Di Indonesia," *J. Inform. Dan Teknol. Komput. JITEK*, vol. 3, no. 2, pp. 108–114, Jul. 2023, doi: 10.55606/jitek.v3i2.1506.

[4] W. T. Oktaviany, F. Insani, A. Nazir, and Pizaini, "Pengelompokan Wilayah Bencana Banjir di Indonesia Menggunakan Algoritma K-Means," *Bull. Comput. Sci. Res.*, vol. 5, no. 4, pp. 542–553, Jun. 2025, doi: 10.47065/bulletincsr.v5i4.608.

[5] D. Delforge *et al.*, "EM-DAT: the Emergency Events Database," *Res. Sq.*, Dec. 2023, doi: 10.21203/rs.3.rs-3807553/v1.

[6] R. L. Jones, A. Kharb, and S. Tubeuf, "The Untold Story of Missing Data in Disaster Research: A Systematic Review of the Empirical Literature Utilising the Emergency Events Database (EM-DAT)," *Environ. Res. Lett.*, vol. 18, no. 10, Oct. 2023, doi: 10.1088/1748-9326/acfd42.

[7] N. Joshi, R. Roberts, and A. Tryggvason, "Incompleteness of Natural Disaster Data and Its Implications on the Interpretation of Trends," *Environ. Hazards*, vol. 24, no. 3, pp. 237–253, 2024, doi: 10.1080/17477891.2024.2377561.

[8] I. Kong and R. S. Purves, "Analyzing Geographic Bias of Newspaper Articles Reporting Global Climate Disasters," *Ann. Am. Assoc. Geogr.*, 2025, doi: 10.1080/24694452.2025.2564220.

[9] A. Kumar, R. Sengupta, S. R. Jegillos, R. Sharma, and S. Bang, *Data and Digital Maturity for Disaster Risk Reduction Informing the Next Generation of Disaster Loss and Damage Databases*. United Nations Office for Disaster Risk Reduction (UNDRR), 2022. [Online]. Available: https://reliefweb.int/node/3911893

[10] K. Poljansek *et al.*, *INFORM Severity Index*. Joint Research Centre (European Commission), 2026. doi: 10.2760/2266313.

[11] K. Teber, M. Weynants, F. Gans, and M. D. Mahecha, "Geo-Disasters: Geocoding Climate-Related Events in the International Disaster Database EM-DAT," *ArXiv Prepr.*, Jun. 2025, doi: 10.48550/arXiv.2506.03797.

[12] Syugiarto, "Disaster Management System in Indonesia," *Sumatra J. Disaster Geogr. Geogr. Educ.*, vol. 5, no. 2, Nov. 2021, doi: 10.24036/sjdgge.v5i2.377.

[13] D. Saputra, A. Wibowo, C. S. Marnani, P. Widodo, H. R. Juni, and K. Saragih, "Strategy and Formulation of Disaster Management Policies in Indonesia," *Int. J. Progress. Sci. Technol. IJPSAT*, vol. 38, no. 1, pp. 520–529, 2023, doi: 10.52155/ijpsat.v38.1.5246.

[14] L. E. Norris, M. Lemin, and L. A. Kelly-Hope, "Spatial-temporal analysis of natural hazards and disasters in the Greater Horn of Africa between 2010 and 2024 to inform disaster risk reduction, and surveillance and control strategies for climate and environmentally sensitive diseases," *BMJ Open*, vol. 15, no. 11, Nov. 2025, doi: 10.1136/bmjopen-2025-104998.

[15] N. Dwitiyanti, S. A. Kumala, and S. D. Handayani, "Comparative Study of Earthquake Clustering in Indonesia Using K-Medoids, K-Means, DBSCAN, Fuzzy C-Means and K-AP Algorithms," *J. RESTI*, vol. 8, no. 6, pp. 768–778, Dec. 2024, doi: 10.29207/resti.v8i6.5514.

[16] S. Supriyadi and A. Gamot Malau, "Clustering dengan Menggunakan K-Means untuk Analisa Dampak Banjir," *Indones. Impr. J. JII*, vol. 4, no. 9, 2025, doi: 10.58344/jii.v4i9.6998.

[17] E. L. Rosvold and H. Buhaug, "GDIS, a global dataset of geocoded disaster locations," *Sci. Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1038/s41597-021-00846-6.

[18] V. Jayawardene, T. J. Huggins, R. Prasanna, and B. Fakhruddin, "The role of data and information quality during disaster response decision-making," *Prog. Disaster Sci.*, vol. 12, Dec. 2021, doi: 10.1016/j.pdisas.2021.100202.

[19] L. Zhao *et al.*, "Disaster loss index development and comprehensive assessment: A case study of Shanghai," *Ecol. Indic.*, vol. 166, Sep. 2024, doi: 10.1016/j.ecolind.2024.112497.

[20] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *Eurasip J. Wirel. Commun. Netw.*, vol. 2021, no. 1, Dec. 2021, doi: 10.1186/s13638-021-01910-w.

[21] H. Hafid and I. Muthahharah, "Clustering of Disaster Risk in Indonesia Regions Using Self-Organizing Maps and K-Means," *JOMTA J. Math. Theory Appl.*, vol. 7, no. 2, pp. 107–114, 2025, doi: 10.31605/jomta.v7i2.5365.

[22] D. Masante *et al.*, *GDACS User Consultation 2023-2024*. Publications Office of the European Union, 2024. doi: 10.2760/435644.

[23] Md. K. Hasan, T. K. Aunto, T. Ahmed, K. R. Calma, and J. Ranse, "Trends and Advancements in Disaster Preparedness Research: A Global Bibliometric Analysis (1951-2024)," *Nat. Hazards Res.*, Dec. 2025, doi: 10.1016/j.nhres.2025.12.004.

[24] F. A. Shaik, A. Holappa, O. Myllymäki, J. Sotaniemi, and M. Oussalah, "Network-Based Simulation of Flood Alert Dissemination using GDACS and Synthetic Communication Logs," in *UbiComp Companion 2025 - Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Association for Computing Machinery, Inc, Dec. 2025, pp. 181–185. doi: 10.1145/3714394.3754440.

