



Pengelompokan Wilayah Bencana Banjir di Indonesia Menggunakan Algoritma K-Means

Wenny Tarisa Oktaviany, Fitri Insani*, Alwis Nazir, Pizaini

Fakultas Sains dan Teknologi, Prodi Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia

Email: ¹12050127565@students.uin-suska.ac.id, ^{2,*}fitri.insani@uin-suska.ac.id,

³alwis.nazir@uin-suska.ac.id, ⁴pizaini@uin-suska.ac.id

Email Penulis Korespondensi: fitri.insani@uin-suska.ac.id

Abstrak—Banjir merupakan salah satu bencana alam yang sering terjadi di Indonesia, terutama pada musim hujan. Bencana ini disebabkan oleh berbagai faktor, baik alami maupun yang disebabkan oleh aktivitas manusia, seperti curah hujan yang tinggi, sistem drainase yang buruk, alih fungsi lahan, dan perencanaan tata ruang yang tidak optimal. Dampak banjir sangat merugikan, baik secara fisik maupun psikologis, termasuk korban jiwa serta kerusakan harta benda. Oleh karena itu, diperlukan suatu metode untuk mengelompokkan wilayah berdasarkan tingkat kerawannya terhadap banjir. Penelitian ini bertujuan untuk mengelompokkan wilayah bencana banjir di Indonesia menggunakan algoritma *K-Means*. Data yang digunakan berasal dari Geoportal BNPB mencakup kejadian banjir dari Januari 2020 hingga Desember 2024, dengan total 7.487 kejadian dari 498 wilayah. Berdasarkan hasil pengujian yang diperoleh dengan menggunakan *Silhouette Coefficient* menunjukkan bahwa 2 *cluster* terpilih menjadi jumlah *cluster* terbaik dengan nilai *Silhouette Coefficient* sebesar 0,8461 yang termasuk struktur *clustering* yang kuat. Dari 2 *cluster* yang diperoleh, *cluster* 1 merupakan kategori rawan tinggi yang terdiri dari 35 wilayah, sedangkan *cluster* 2 merupakan kategori rawan rendah yang terdiri dari 463 wilayah. Hasil dari penelitian ini dapat memberikan informasi bagi pihak terkait untuk meningkatkan efisiensi penanggulangan bencana banjir.

Kata Kunci: Banjir; Indonesia; K-Means; Silhouette Coefficient

Abstract—Floods are one of the natural disasters that often occur in Indonesia, especially during the rainy season. This disaster is caused by various factors, both natural and caused by human activities, such as high rainfall, poor drainage systems, land conversion, and suboptimal spatial planning. The impact of floods is very detrimental, both physically and psychologically, including loss of life and damage to property. Therefore, a method is needed to group areas based on their level of vulnerability to flooding. This study aims to group flood disaster areas in Indonesia using the K-Means algorithm. The data used comes from the BNPB Geoportal covering flood events from January 2020 to December 2024, with a total of 7,487 events from 498 areas. Based on the test results obtained using the Silhouette Coefficient, it shows that 2 clusters were selected as the best number of clusters with a Silhouette Coefficient value of 0.8461 which is included in the strong clustering structure. Of the 2 clusters obtained, cluster 1 is a high-risk category consisting of 35 areas, while cluster 2 is a low-risk category consisting of 463 areas. The results of this study can provide information for related parties to improve the efficiency of flood disaster management.

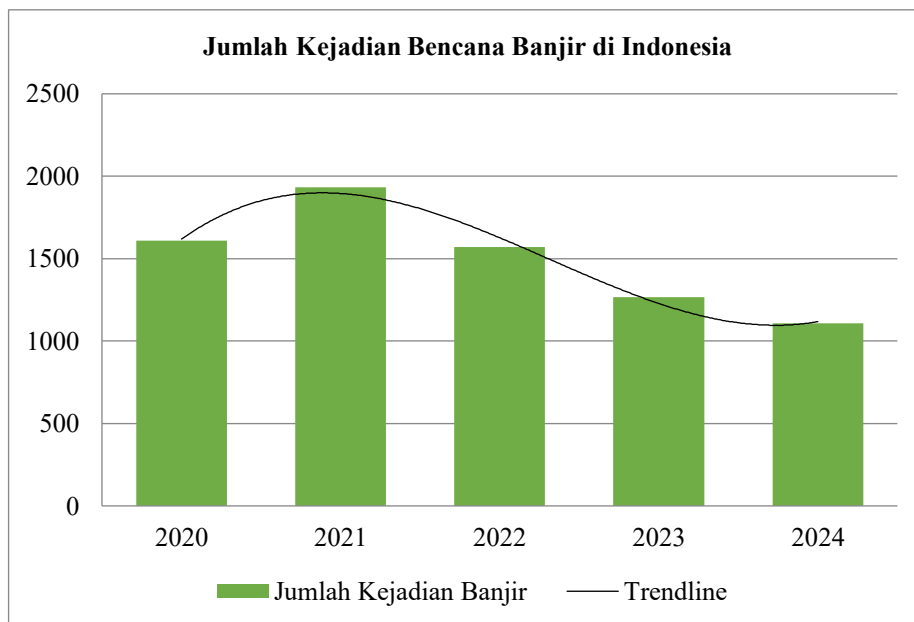
Keywords: Flood; Indonesia; K-Means; Silhouette Coefficient

1. PENDAHULUAN

Indonesia merupakan negara kepulauan yang terletak pada pertemuan empat lempeng tektonik utama, yaitu lempeng benua Australia, lempeng benua Asia, lempeng Samudra Pasifik, dan lempeng Samudra Hindia. Negara ini juga dikenal dengan sebutan Cincin Api Pasifik atau cincin api dunia [1]. Letak geografis Indonesia yang berada di antara lautan dan benua menyebabkan sebagian besar wilayahnya beriklim tropis. Suhu rata-rata di pesisir pantai sekitar 28°C, suhu rata-rata di pedalaman dan pegunungan sekitar 26°C, serta suhu di dataran tinggi dapat mencapai 23°C. Kelembaban udaranya pun relatif lembab, berkisar antara 70% hingga 90%. Pola angin musiman dari selatan dan timur pada bulan Juni hingga September, serta dari barat laut pada bulan Desember hingga Maret, sering kali menyebabkan angin kencang, topan, dan badai besar [2].

Kondisi ini berkaitan dengan keadaan geologis, demografis dan geografis yang meningkatkan terjadinya suatu bencana, baik yang disebabkan oleh faktor alam seperti tanah longsor, gempa bumi, tsunami, letusan gunung berapi, banjir, serta faktor non alam dan manusia [3]. Hal ini menyebabkan Indonesia rawan terhadap bencana alam [4]. Bencana alam merenggut banyak korban jiwa dan menimbulkan kerusakan harta benda di seluruh dunia [5]. Sebagaimana yang disebutkan oleh Pusat Penelitian Epidemiologi Bencana (*Centre of Research on the Epidemiology of Disaster*, 2018) Indonesia secara konsisten berada pada posisi lima teratas dalam daftar negara yang sering mengalami bencana alam [6].

Bencana hidrometeorologi yang paling sering terjadi di Indonesia ialah bencana banjir [7]. Luapan air yang terjadi karena curah hujan yang meningkat sehingga menyebabkan volume air meningkat di sungai dan danau disebut banjir [8]. Banjir yang pada dasarnya merupakan proses alamiah yang dapat menjadi bencana bagi manusia bila proses itu mengenai manusia dan menyebabkan kehilangan jiwa maupun kerugian materi [9]. Setiap tahun Indonesia sering mengalami bencana banjir. Dalam 10 tahun terakhir, bencana banjir selalu menempati posisi pertama kejadian bencana [10]. Berdasarkan data yang dimiliki oleh Badan Nasional Penanggulangan Bencana (BNPB), teridentifikasi sejak Januari 2020 hingga Desember 2024 terdapat 7.487 bencana banjir yang terjadi di Indonesia. Data jumlah bencana banjir di Indonesia dapat dilihat pada Gambar 1.



Gambar 1. Jumlah Kejadian Bencana Banjir di Indonesia

Menyadari risiko bencana banjir di Indonesia, pemerintah membentuk Badan Nasional Penanggulangan Bencana (BNPB) untuk mengelola tanggap darurat, rehabilitasi dan mitigasi [11]. Untuk memperkuat penanggulangan bencana di setiap daerah, dibentuklah Badan Penanggulangan Bencana Daerah (BPBD) berdasarkan Undang-Undang No. 24 Tahun 2007 yang berisi pernyataan resmi kepada pemerintah daerah untuk menyelenggarakan program penanggulangan bencana. Untuk mengurangi dan meringankan kerugian, kerusakan, dan korban jiwa diperlukan kerja sama berbagai pihak untuk meningkatkan ketahanan ekonomi, fisik, dan sosial. Upaya tersebut dapat dilakukan dengan memanfaatkan sarana dan prasarana, teknologi, informasi, sumber daya manusia, dan pengetahuan [12]. Oleh karena itu, analisis data diperlukan untuk memudahkan proses pengumpulan informasi. Penelitian ini dilakukan untuk merealisasikan penggunaan analisis *cluster* untuk memberikan gambaran berupa *cluster* daerah dengan karakteristik cluster-nya.

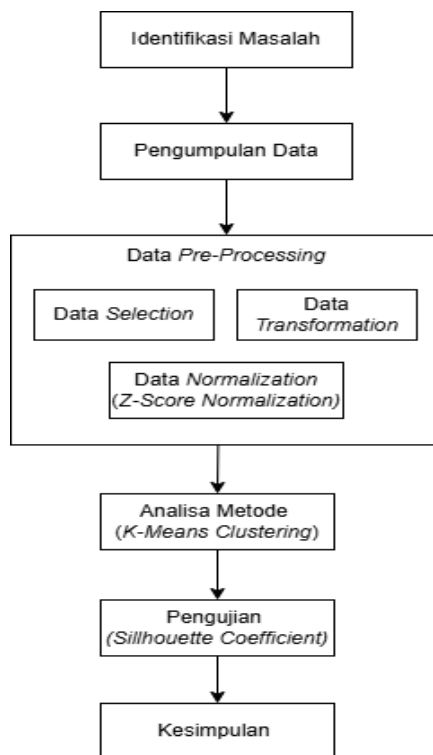
Clustering merupakan suatu proses di mana pola-pola data dikelompokkan dan dibagi menjadi beberapa set data dengan tujuan membentuk pola-pola yang serupa yang dikelompokkan dalam *cluster* yang sama, sementara pola-pola yang berbeda dikelompokkan dalam *cluster* yang berbeda [13]. Dalam penelitian ini, teknik *clustering* yang digunakan ialah algoritma *K-Means*. Algoritma *K-Means* merupakan suatu metode yang digunakan untuk mengelompokkan data dengan cara membagi data tersebut ke dalam beberapa kelompok sehingga data yang memiliki kemiripan tertentu berada pada kelompok yang sama, sedangkan data yang berbeda ditempatkan pada kelompok yang berbeda [14]. Tujuan dari algoritma *K-Means* yaitu meminimalkan fungsi dengan cara meminimalkan variasi antara data dalam satu klaster dan memaksimalkan variasi data pada klaster lain [15].

Berdasarkan penelitian terkait pengelompokan data obat telah diteliti oleh Riva Arsyad Farissa dkk (2021) yang mengelompokkan data obat-obatan di Puskesmas Karangsambung sehingga dapat digunakan sebagai dasar untuk perencanaan obat-obatan di masa mendatang [16]. Penelitian selanjutnya yang dilakukan oleh Diana Dwi Aulia dkk (2023) yang mengelompokkan tingkat pendidikan di Kalimantan Tengah untuk melakukan penyaluran bantuan pendidikan, termasuk dukungan untuk siswa serta penyediaan fasilitas sekolah di desa-desa dengan tingkat pendidikan yang paling rendah [17]. Penelitian selanjutnya yang dilakukan oleh Muhammad Makmun Effendi dkk (2024) yang mengelompokkan wilayah rawan banjir di wilayah Bekasi bertujuan untuk mempermudah penyampaian informasi kepada masyarakat mengenai daerah rawan banjir di wilayah Bekasi [18]. Berdasarkan penelitian-penelitian diatas dapat disimpulkan bahwa algoritma *K-Means* berhasil dalam melakukan pengelompokan data.

Pada penelitian ini, dilakukan pengelompokan wilayah bencana banjir di Indonesia dengan menggunakan algoritma *K-Means*. Data yang digunakan pada penelitian ini adalah data kejadian bencana banjir yang terjadi di Indonesia pada Januari 2020 hingga Desember 2024 yang diperoleh dari situs resmi Geoportal Data Bencana Indonesia milik BNPB. Tujuan dari penelitian ini adalah menerapkan algoritma *K-Means* dalam pengelompokan wilayah bencana banjir di Indonesia. Hasil penelitian ini dapat memberikan informasi bagi pihak terkait untuk meningkatkan efisiensi penanggulangan bencana banjir.

2. METODOLOGI PENELITIAN

Metodologi penelitian ini menjelaskan tahapan-tahapan yang akan dilakukan dalam proses penelitian untuk memecahkan rumusan masalah dalam penelitian. Metode penelitian yang digunakan didasarkan pada tahapan-tahapan berikut:



Gambar 2. Metodologi Penelitian

2.1 Identifikasi Masalah

Identifikasi masalah adalah langkah pertama dalam metodologi penelitian ini. Untuk memperjelas masalah dan membentuk landasan penelitian. Badan Penanggulangan Bencana Daerah (BPBD) bertanggung jawab atas tanggap darurat, rehabilitasi, dan mitigasi bencana. Agar ada kesesuaian antara pendekatan umum dan karakteristik setiap wilayah, penting untuk mempertimbangkan karakteristik masing-masing wilayah selama proses penanggulangan bencana, terutama saat mengumpulkan informasi. Analisis data yang dapat membantu proses pengumpulan data diperlukan. Setelah memahami konteks masalah, langkah berikutnya adalah menganalisis literatur dan penelitian sebelumnya di bidang tersebut untuk menemukan celah pengetahuan. Setelah masalah diidentifikasi, rumusan masalah penelitian dibuat ini akan menjadi dasar dari seluruh metodologi penelitian yang akan digunakan untuk menemukan solusi untuk masalah tersebut.

2.2 Pengumpulan Data

Pada penelitian ini data yang digunakan merupakan data kejadian bencana banjir yang terjadi di Indonesia. Data yang diperoleh merupakan data kejadian bencana banjir yang terjadi sejak Januari 2020 hingga Desember 2024, yang diambil dari situs resmi Geoport Data Bencana Indonesia milik BNPB.

2.3 Data Pre-Processing

Dalam penelitian ini, data pre-processing dilakukan setelah pengumpulan data. Pada tahap ini, dimana data mentah disiapkan, dibersihkan, dan diubah agar bisa dianalisis lebih lanjut. Berikut adalah langkah-langkah data *pre-processing*:

2.3.1 Data Selection

Proses data *selection* pada penelitian ini dimulai dengan memilih variabel yang akan digunakan berdasarkan relevansi rumusan masalah penelitian, sehingga mewakili tujuan penelitian.

2.3.2 Data Transformation

Data *transformation* adalah tahap mengubah data menjadi bentuk yang sesuai untuk diproses dalam data mining.

2.3.3 Data Normalization

Penelitian ini menerapkan data normalisasi sebagai tahapan selanjutnya pada data *pre-processing*, yaitu menggunakan *Z-Score normalization*. Normalisasi data adalah langkah pertama dalam proses data, yang dilakukan saat data dikumpulkan untuk meminimalkan redundansi data dan membuat data lebih mudah digunakan atau diolah oleh proses lainnya. Setelah normalisasi selesai, data dapat digabungkan atau digunakan dengan metode lain [19]. Nilai *Z-Score* merupakan ukuran yang menunjukkan seberapa jauh suatu data menyimpang dari nilai rata-rata (\bar{x}) dalam satuan standar deviasinya (SDx).



Z-Score ini juga dikenal sebagai Nilai Standar atau Nilai Baku. Nilai *Z-Score* dari seluruh data dapat dihitung menggunakan rumus yang diberikan pada persamaan berikut [20]:

$$Z = \frac{x - \bar{x}}{SD_x} \quad (1)$$

Keterangan:

Z : Hasil normalisasi

x : Nilai awal yang akan di normalisasi

\bar{x} : Nilai rata-rata

SD_x : Standar Deviasi

Sedangkan, untuk menghitung nilai standar deviasi dapat diselesaikan dengan rumus sebagai berikut:

$$SD_x = \frac{1}{n} \sum_{i=1}^n (x - \bar{x}) \quad (2)$$

SD_x : Standar Deviasi

n : Jumlah sampel

x : Nilai rata-rata

x_i : Nilai x ke i

2.4 Algoritma *K-Means*

Algoritma *K-Means* merupakan salah satu algoritma *unsupervised learning* yang termasuk dalam analisis kluster non hirarki yang digunakan untuk mengelompokkan sejumlah data berdasarkan suatu variable [21]. Teknik *unsupervised learning* mengidentifikasi pola dalam data tanpa bergantung pada label atau *output* yang telah ditentukan sebelumnya [22].

Metode ini mempartisi data ke dalam *cluster*/kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama sedangkan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang berbeda [23]. Langkah-langkah dalam melakukan *clustering* dengan metode *K-Means* adalah sebagai berikut:

- Tentukan nilai k (nilainya bebas) sebagai jumlah *cluster* yang ingin digunakan.
- Tentukan nilai k sebagai *centroid* (*cluster awal*) yang dihasilkan secara acak.
- Untuk setiap objek lakukan perhitungan jarak ke nilai pusat *centroid* yang telah ditentukan sebelumnya. Pengukuran jarak ini dapat dilakukan dengan menggunakan rumus jarak *Euclidean Distance*. Hasil perhitungan jarak tersebut dibandingkan, kemudian dipilih jarak terpendek dari data ke pusat kluster. Jarak terpendek tersebut menunjukkan bahwa data tersebut termasuk dalam kelompok yang sama dengan pusat kluster terdekat. Rumus untuk menentukan jarak kluster dengan menggunakan jarak *Euclidean Distance* adalah:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (3)$$

Keterangan :

$d(x, y)$: Jarak antara nilai pada data setiap variabel (x) dengan nilai *centroid* nya (y)

i : Setiap data

n : Jumlah seluruh data

x_i : Data yang berada di pusat *cluster* ke- i

y_i : Data yang berada di setiap data ke- i

- Mengelompokkan objek dengan mencari jarak terdekat setiap data ke titik pusatnya dengan menentukan keanggotaan *cluster*.
- Tentukan nilai *centroid* baru untuk proses iterasi berikutnya dengan menghitung rata-rata data di setiap *cluster* menggunakan rumus berikut:

$$C_k = \left(\frac{1}{N_k} \right) + \sum_{i=0}^{N_k} d_i \quad (4)$$

Keterangan :

C_k : Rata-rata dari *centroid* baru

N_k : Jumlah data yang ada di dalam *cluster*

i, k : Nilai indeks dari suatu *cluster*

d_i : Nilai dari data ke- i yang ada pada *cluster* untuk setiap variabel

- Ulangi langkah 3 hingga 5 hingga tidak ada anggota setiap *cluster* yang berubah atau berpindah.

2.5 *Silhouette Coefficient*

Silhouette coefficient merupakan salah satu metode yang digunakan untuk melihat kualitas dan kekuatan suatu *cluster*, seberapa baik suatu objek ditempatkan dalam suatu *cluster*. Metode ini merupakan gabungan dari metode *cohesion* dan *separation* [24]. Berikut merupakan tahapan-tahapan untuk menghitung *Silhouette coefficient* [25]:



- a. Hitung jarak rata-rata dari data ke- i ke seluruh data lain dalam klaster yang sama, menggunakan persamaan berikut:

$$a(i) = \frac{1}{|A|-1} \sum_{j \in A, j \neq i} d(i, j) \tag{5}$$

di mana,

A = total data di *cluster* A

- b. Untuk mencari nilai $b(i)$, hitunglah jarak rata-rata data ke- i dengan seluruh data di *cluster* yang berbeda, kecuali *cluster* A . Misalnya, untuk menghitung jarak rata-rata data ke- i dengan semua data di *cluster* C , menggunakan persamaan berikut:

$$b(i, C) = \frac{1}{|C|} \sum_{j \in C} d(i, j) \tag{6}$$

di mana,

C = total data di *cluster*

$K = 1, 2, 3, \dots, p$

- c. Setelah menghitung $d(i, C)$ untuk semua *cluster*, langkah selanjutnya adalah memilih jarak minimum sebagai nilai $b(i)$, menggunakan persamaan berikut:

$$b(i) = \min_{C \neq A} d(i, C) \tag{7}$$

Cluster terbaik kedua untuk data ke- i , setelah *cluster* A , adalah *cluster* B , yang disebut sebagai tetangga dari data ke- i jika *cluster* B memiliki nilai jarak minimal, yaitu $d(i, B) = b(i)$. *Silhouette coefficient* ditentukan pada langkah terakhir setelah $a(i)$ dan $b(i)$ diketahui.

- d. Jika nilai indeks *silhouette* mendekati 1, berarti proses *clustering* berhasil, dan jika mendekati 0, berarti proses *clustering* tidak berhasil. Hitung nilai *Silhouette coefficient* pada setiap data ke- i dengan menggunakan persamaan berikut:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i) - b(i)\}} \tag{8}$$

Keterangan:

$s(i)$: nilai *silhouette coefficient* yang berada dalam data ke- i

(i) : object yang akan diteliti

$a(i)$: rata-rata jarak dari data ke- i pada semua data yang berada dalam cluster yang sama.

$b(i)$: rata-rata jarak dari data ke- i pada semua data yang berada dalam cluster yang berbeda.

$d(i, C)$: rata-rata jarak dalam cluster lain pada objek ke- i terhadap seluruh objek

$i \in C$: data ke- i merupakan anggota kelompok C

d_{ij} : jarak data ke- i terhadap j

Nilai $s(i)$ terletak diantara -1 dan 1, di mana setiap nilai ditafsirkan sebagai berikut:

$s(i) = 1 \Rightarrow$ data ke- i digolongkan dengan baik (dalam A)

$s(i) = 0 \Rightarrow$ data ke- i terletak di tengah antara dua *cluster* (A dan B)

$s(i) = -1 \Rightarrow$ data ke- i tergolong lemah (dekat ke *cluster* B daripada A)

Penafsiran nilai *Silhouette Coefficient* ditunjukkan dalam tabel 1 sebagai berikut:

Tabel 1. Interpretasi Nilai *Silhouette Coefficient*

<i>Silhouette Coefficient</i>	Interpretasi
0.71 – 1.00	Struktur yang dihasilkan kuat
0.51 – 0.70	Struktur yang dihasilkan baik
0.26 – 0.50	Struktur yang dihasilkan lemah
≤ 0.25	Tidak terstruktur

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Pada penelitian ini data yang digunakan merupakan data kejadian bencana banjir di Indonesia. Data yang diperoleh merupakan data rekapan kejadian banjir yang terjadi sejak Januari 2020 hingga Desember 2024 yang diperoleh langsung dari situs resmi Geoport Data Bencana Indonesia milik BNPB dengan jumlah keseluruhan data 7,487 record.

Tabel 2. Rekapan data banjir tahun 2020-2024

No	Kode Identitas Bencana	...	Wilayah	Provinsi	...	Rumah Terendam	Fasum Rusak
1	1105 101 20241117 1	...	Aceh Barat	Aceh	...	1312	0
2	1105 101 20241014 1	...	Aceh Barat	Aceh	...	0	0



No	Kode Identitas Bencana	...	Wilayah	Provinsi	...	Rumah Terendam	Fasum Rusak
3	1105 101 20240921 1	...	Aceh Barat	Aceh	...	1578	0
4	1105 101 2023 012 011 1	...	Aceh Barat	Aceh	...	0	0
5	83,17	...	Aceh Barat	Aceh	...	0	0
...
7.485	1206 101 2022 08 027 1	...	Toba	Sumatera Utara	...	0	1
7.486	1206 101 2022 04 011 1	...	Toba	Sumatera Utara	...	4	0
7.487	1206 101 2021 012 31 1	...	Toba	Sumatera Utara	...	79	0

3.2 Data Pre-Processing

Penelitian ini menggunakan tahapan data *pre-processing* yang meliputi data *selection*, data *transformation* dan data *normalization*.

3.2.1 Data Selection

Proses data *selection* pada penelitian ini dimulai dengan memilih variabel yang terdapat pada data berdasarkan relevansi rumusan masalah penelitian, sehingga mewakili tujuan penelitian. Dari keseluruhan variabel yang ada pada data, yang digunakan dalam penelitian ini adalah variabel jumlah korban meninggal, jumlah orang hilang, jumlah orang terluka, jumlah rumah rusak, jumlah rumah terendam dan jumlah fasilitas umum rusak. Hasil dari proses data *selection* tercantum pada Tabel 3 berikut:

Tabel 3. Data kejadian banjir 2020-2024

No	Wilayah	Meninggal	Hilang	Terluka	Rumah Rusak	Rumah Terendam	Fasum Rusak
1	Aceh Barat	0	0	0	0	1312	0
2	Aceh Barat	0	0	0	0	0	0
3	Aceh Barat	2	0	0	1	1578	0
4	Aceh Barat	0	0	0	1	0	0
5	Aceh Barat	0	0	0	0	0	0
...
7.486	Toba	0	0	0	1	0	1
7.486	Toba	0	0	0	0	4	0
7.487	Toba	0	0	0	0	79	0

3.2.2 Data Transformation

Pada tahap ini dilakukan transformasi data ke dalam bentuk yang sesuai dengan proses data mining. Dari total 7.487 record, dilakukan proses transformasi menjadi 498 data agregat berdasarkan wilayah. Transformasi data dilakukan pada wilayah Aceh Barat tercatat mengalami 21 kejadian banjir. Selanjutnya, dibuat variabel baru berupa jumlah kejadian yang diperoleh dari penggabungan data per wilayah. Variabel lain juga dihitung jumlahnya per wilayah sehingga seluruh data dapat direpresentasikan dalam satu baris untuk setiap wilayah. Proses ini bertujuan untuk menyederhanakan data sehingga lebih mudah dianalisis dan dikelompokkan berdasarkan karakteristik bencana banjir di masing-masing wilayah. Berikut transformasi data terdapat pada Tabel 4 sebagai berikut:

Tabel 4. Data yang telah ditransformasikan

No	Wilayah	Jumlah Kejadian	Meninggal	Hilang	Terluka	Rumah Rusak	Rumah Terendam	Fasum Rusak
1	Aceh Barat	21	3	0	0	53	13695	20
2	Aceh Barat Daya	10	0	0	1	6	1045	2
3	Aceh Besar	14	3	3	0	0	1898	0
4	Aceh Jaya	25	0	0	0	0	13221	0
5	Aceh Selatan	39	2	0	0	18	16527	9
...
496	Tapanuli Tengah	3	12	0	22	113	1536	1
497	Tapanuli Utara	4	0	0	0	4	223	0
498	Toba	4	0	0	0	1	143	1

3.2.3 Data Normalisasi

Berdasarkan hasil dari proses data transformasi, selanjutnya melakukan normalisasi data menggunakan *Z-Score normalization*. Harapannya hasil akurasi lebih tepat daripada tanpa melakukan normalisasi data. Data dari hasil proses



data transformasi tersebut dinormalisasikan menggunakan persamaan 1 untuk menghitung nilai *Z-Score* dan persamaan 2 untuk menghitung nilai Standar Deviasi. Berikut perhitungan normalisasi data menggunakan *Z-Score Normalization*. Perhitungan yang dilakukan menggunakan data wilayah Aceh Barat pada variabel jumlah kejadian. Langkah pertama yaitu menghitung nilai rata-rata (\bar{x}) dan Standar Deviasi (SDx), perhitungan nilai rata-rata dan Standar Deviasi untuk variabel jumlah kejadian terhadap keseluruhan data wilayah sebagai berikut:

$$(\bar{x}) = \frac{21+10+14+\dots+4}{498} = 15,034136$$

$$SDx = \sqrt{\frac{(21-15,034)^2+(10-15,034)^2+(14-15,034)^2+\dots+(4-15,034)^2}{498}} = 16.126284$$

Selanjutnya mencari nilai *Z-Score Normalization* menggunakan persamaan 1, perhitungan yang dilakukan menggunakan data wilayah Aceh Barat pada variabel jumlah kejadian.

$$Z = \frac{21 - 15,034136}{16.126284} = \frac{5,965864}{16.126284} = 0,369947$$

Berikut hasil normalisasi keseluruhan data terdapat pada tabel 5 sebagai berikut:

Tabel 5. Hasil *Z-Score Normalization*

No	Wilayah	Jumlah Kejadian	Meninggal	Hilang	Terluka	Rumah Rusak	Rumah Terendam	Fasum Rusak
1	Aceh Barat	0.369947	0.164932	-0.221814	-0.074227	-0.199014	0.095760	-0.036919
2	Aceh Barat Daya	-0.312170	-0.363842	-0.221814	-0.072192	-0.227773	-0.412075	-0.174526
3	Aceh Besar	-0.064127	0.164932	1.739071	-0.074227	-0.231444	-0.377831	-0.189816
4	Aceh Jaya	0.617989	-0.363842	-0.221814	-0.074227	-0.231444	0.076732	-0.189816
5	Aceh Selatan	1.486137	-0.011326	-0.221814	-0.074227	-0.220430	0.209451	-0.121012
...
496	Tapanuli Tengah	-0.746244	1.751254	-0.221814	-0.029451	-0.162302	-0.392364	-0.182171
497	Tapanuli Utara	-0.684233	-0.363842	-0.221814	-0.074227	-0.228996	-0.445074	-0.189816
498	Toba	-0.684233	-0.363842	-0.221814	-0.074227	-0.230832	-0.448286	-0.182171

3.3 Implementasi Algoritma *K-Means*

Setelah data yang didapat telah melalui tahapan data *pre-processing*, dan menghasilkan dataset yang siap untuk diolah maka langkah selanjutnya yaitu menerapkan algoritma *K-Means* pada dataset tersebut.

a. Tentukan jumlah *cluster*

- Jumlah *cluster* : 2
- Jumlah data : 498 record
- Jumlah variabel : 7

b. Menentukan *centroid* awal

Untuk menentukan titik pusat *cluster* atau *centroid* secara acak dilakukan dengan pemilihan data yang ingin dijadikan *centroid* awal. Karena pengelompokan daerah rawan banjir yang telah ditentukan sebelumnya adalah sebanyak 2 *cluster*, maka dialokasikan juga titik pusat *cluster (centroid)* sebanyak 2 data yang terdapat pada data ke-1 dan ke-318 yang diambil secara acak dari data kejadian banjir dari hasil yang telah dinormalisasikan pada tabel 5. Berikut *centroid* awal yang telah ditentukan sebelumnya terdapat pada tabel 6 sebagai berikut:

Tabel 6. *Centroid* Awal

Cluster	Wilayah	V1	V2	V3	V4	V5	V6	V7
1	Aceh Barat	0.36995	0.16493	-0.22181	-0.07423	-0.19901	0.09576	-0.03692
2	Merauke	-0.87026	-0.36384	-0.22181	-0.07423	-0.23144	-0.45403	-0.18982

c. Menghitung jarak antar data terhadap *centroid* setiap *cluster*.

Setelah menentukan *centroid* awal, langkah selanjutnya yaitu menghitung jarak antara data dengan *centroid* menggunakan perhitungan jarak *Euclidean Distance* dengan rumus pada persamaan 3. Berikut ini adalah contoh perhitungan menggunakan *Euclidean Distance*, dimulai dari jarak data-1 sampai jarak data-498. Berikut perhitungan jarak masing-masing data pada data-1 wilayah Aceh Barat untuk iterasi-1 sebagai berikut:

1. Perhitungan jarak dari data-1 terhadap *centroid cluster 1*

$$d(1,1) = \sqrt{(0,36994 - 0,36994)^2 + (0,16493 - 0,16493)^2 + \dots + (-0,03691 - (-0,03691))^2}$$

$$d = 0$$

2. Perhitungan jarak dari data-1 terhadap *centroid cluster 2*

$$d(1,2) = \sqrt{(0,36994 - (-0,87026))^2 + (0,16493 - (-0,36384))^2 + \dots + (-0,03691 - (-0,18981))^2}$$



$$d = 1,46438$$

Kemudian dilanjutkan perhitungan hingga data ke-498.

d. Mengelompokkan objek dengan mencari jarak terdekat

Setelah menghitung jarak setiap data yang ada terhadap setiap pusat *cluster*, langkah selanjutnya adalah mengalokasikan setiap objek ke *centroid* terdekat, dengan cara mengelompokkan berdasarkan jarak minimum objek terhadap pusat *cluster*, berdasarkan hasil perhitungan pada data wilayah Aceh Barat menunjukkan bahwa data-1 merupakan data yang paling dekat dengan *cluster* 1. Dengan demikian, data ke-1 dapat dikategorikan ke dalam kelompok *cluster* 1 sehingga diperoleh iterasi-1. Berikut adalah hasil pengelompokan pada iterasi-1 terdapat pada tabel 7 sebagai berikut:

Tabel 7. Hasil pengelompokan pada iterasi-1

No	Wilayah	C1	C2	Cluster
1	Aceh Barat	0	1,46438	C1
2	Aceh Barat Daya	1,01121	0,55989	C2
3	Aceh Besar	2,06935	2,18639	C1
4	Aceh Jaya	0,60491	1,58006	C1
5	Aceh Selatan	1,13903	2,47425	C1
...
496	Tapanuli Tengah	2.006243	2.121239	C1
497	Tapanuli Utara	1.30678	0.186263	C2
498	Toba	1.307285	0.186278	C2

e. Menentukan nilai *centroid* yang baru untuk proses iterasi

Setelah melakukan perhitungan jarak dan pengelompokan data seperti pada tabel 7. Kemudian lakukan perhitungan iterasi pada iterasi-2 dan diawali dengan menentukan nilai *centroid* yang baru dengan menggunakan persamaan 4 dan tentukan posisi C1 dan C2 baru dengan menghitung rata-rata dari data yang berada pada *centroid* yang sama pada iterasi-1. Kemudian kita tentukan kembali pusat *cluster* dari data yang baru, dengan cara menjumlahkan semua nilai yang menjadi anggota dari *cluster* dan dibagi total dari setiap jumlah anggota *cluster*, maka didapatkan anggota *cluster* yang baru. Berikut adalah perhitungan untuk menentukan *centroid* yang baru:

1. Hitung rata-rata pusat *cluster* baru pada *cluster* 1 :

$$C1 = \frac{(0,36994+(-0,06412)+0,61798+1,4861+\dots+(-0,74624))}{241} = 0,62751$$

2. Hitung rata-rata pusat *cluster* baru pada *cluster* 2:

$$C2 = \frac{(-0,31216+(0,56021)+(-0,87026)+\dots+(-0,68423))}{257} = -0,58844$$

Perhitungan *cluster* yang diperoleh diatas kemudian akan memperoleh *centroid* baru untuk iterasi-2, seperti pada Tabel 8 dibawah ini:

Tabel 8. *Centroid* baru untuk iterasi-2

Cluster	V1	V2	V3	V4	V5	V6	V7
1	0,62751	0,30097	0,14704	0,06356	0,17755	0,4225	0,16166
2	-0,58844	-0,28223	-0,13788	-0,0596	-0,16649	-0,3962	-0,15159

f. Ulangi langkah 3 hingga 5 hingga tidak ada anggota setiap *cluster* yang berubah atau berpindah.

Setelah nilai *centroid* baru diperbarui, kemudian ulangi langkah seperti pada iterasi-1 sebelumnya. Kemudian data hasil *clustering* untuk setiap *cluster* pada iterasi-2 diperoleh pada tabel 9 sebagai berikut:

Tabel 9. Hasil pengelompokan pada iterasi-2

No	Wilayah	C1	C2	Cluster
1	Aceh Barat	0.72655	1.17556	C1
2	Aceh Barat Daya	1.56629	0.30777	C2
3	Aceh Besar	1.99543	2.00101	C1
4	Aceh Jaya	1.00372	1.30336	C1
5	Aceh Selatan	1.12849	2.18064	C1
...
496	Tapanuli Tengah	2.243426	2.041781	C2
497	Tapanuli Utara	1.832805	0.175644	C2
498	Toba	1.833286	0.175733	C2

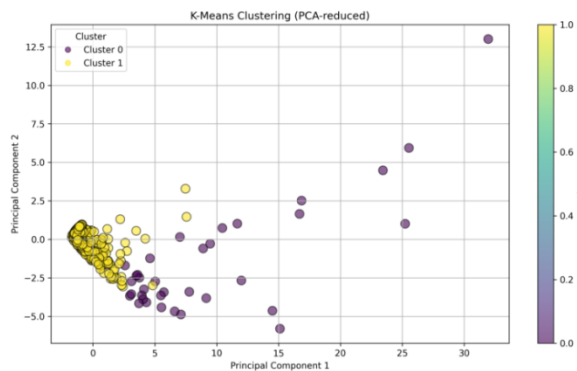


Setelah dilakukan perhitungan jarak dan pengelompokan data seperti pada tabel 9, dapat diketahui bahwa hasil *clustering* pada iterasi-2 masih terdapat perpindahan anggota pada setiap *cluster*. Sehingga perlu dilakukan beberapa kali iterasi lagi untuk dapat membandingkan nilai dari dua iterasi terakhir. Jika nilai dari dua iterasi terakhir sudah sama, maka proses iterasi selesai, dan jika belum, maka ulangi langkah 3 hingga 5 sehingga tidak ada anggota setiap *cluster* yang berubah atau berpindah.

Berdasarkan hasil *clustering* yang diperoleh, proses perhitungan berhenti pada iterasi ke-10. Dengan demikian, hasil *clustering* tetap dan tidak mengalami perubahan atau perpindahan anggota pada setiap *cluster*. Dari hasil akhir 2 *cluster* yang diperoleh, *cluster* 1 merupakan kategori rawan tinggi yang terdiri dari 35 wilayah, sedangkan *cluster* 2 merupakan kategori rawan rendah yang terdiri dari 463 wilayah. Berikut ini adalah hasil pengelompokan pada iterasi ke-10 pada tabel 10 dan visualisasi *K-Means Clustering* pada Gambar 3 sebagai berikut:

Tabel 10. Hasil pengelompokan pada iterasi terakhir

No	Wilayah	C1	C2	Cluster
1	Aceh Barat	4.460762	0.660367	C2
2	Aceh Barat Daya	5.225792	0.404464	C2
3	Aceh Besar	4.542448	1.906244	C2
4	Aceh Jaya	4.739234	0.817247	C2
5	Aceh Selatan	4.358071	1.662843	C2
...
496	Tapanuli Tengah	4.741931	2.032834	C2
497	Tapanuli Utara	5.396307	0.680264	C2
498	Toba	5.396044	0.681109	C2



Gambar 3. Visualisasi *K-Means Clustering*

3.4 Silhouette Coefficient

Dalam penelitian ini, pengujian dilakukan menggunakan metode *Silhouette Coefficient* untuk mengukur seberapa baik pengelompokan data yang dihasilkan oleh algoritma *K-Means*. Pada penelitian ini akan menjabarkan contoh perhitungan pengukuran kualitas *clustering* dengan menerapkan metode *Silhouette Coefficient* pada hasil pengelompokan yang telah dilakukan sebelumnya menggunakan algoritma *K-Means* yang menghasilkan 2 *cluster*.

Untuk menghitung nilai *Silhouette Coefficient*, langkah pertama adalah menghitung jarak rata-rata dari data ke-1 ke seluruh data lain dalam *cluster* yang sama. Data yang digunakan pada perhitungan ini adalah data ke-37 yaitu wilayah Lebak yang berada pada *cluster* 1, yang berjumlah 35 anggota *cluster*. Perhitungan ini dilakukan menggunakan persamaan 5 sebagai berikut:

$$a(i) = \frac{\sqrt{(5,64084-0,86603)^2+(0,16493-2,80880)^2+\dots+(-0,18982-0,53644)^2}}{34}$$

$$= 9.891221$$

Setelah nilai $a(i)$ diketahui, selanjutnya nilai $b(i)$ dihitung dari rata-rata jarak antara data ke-37 dengan anggota di luar *cluster* 1 yang berjumlah 463 anggota *cluster*. Perhitungan ini dilakukan menggunakan persamaan 6 sebagai berikut:

$$b(i) = \frac{\sqrt{(0,36994-0,86603)^2+(0,16493-2,80880)^2+\dots+(-0,18217-0,53644)^2}}{463}$$

$$= 6.474981$$

Setelah nilai $a(i)$ dan $b(i)$ dihitung, nilai SC untuk data wilayah Lebak akan dihitung menggunakan persamaan 8 sebagai berikut:



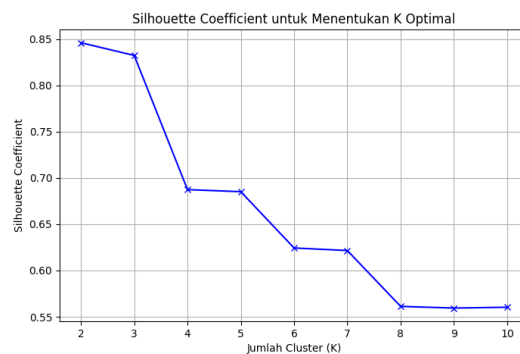
$$s(i) = \frac{6.474981 - 9.891221}{\max(9.891221; 6.474981)} = \frac{-3.41624}{9.891221} = -0.345381$$

Hasil perhitungan *Silhouette Coefficient* menunjukkan bahwa nilai untuk data ke-37 adalah -0.345381. nilai *Silhouette Coefficient* keseluruhan diperoleh dengan menghitung rata-rata dari hasil perhitungan *Silhouette Coefficient* untuk keseluruhan data wilayah, seperti yang ditunjukkan pada Tabel 11 sebagai berikut:

Tabel 11. Hasil nilai keseluruhan data *Silhouette Coefficient*

No	Wilayah	$a(i)$	$b(i)$	$s(i)$
1	Aceh Barat	1.604912	12.400650	0.870578
2	Aceh Barat Daya	1.270068	13.018337	0.902440
3	Aceh Besar	2.868814	11.348100	0.747199
4	Aceh Jaya	1.583567	12.329892	0.871567
5	Aceh Selatan	2.531644	11.312899	0.776216
...
496	Tapanuli Tengah	3.016646	11.298748	0.733011
497	Tapanuli Utara	1.275470	12.835521	0.900630
498	Toba	1.272924	12.834909	0.900823
			Rata-rata	0.846110

Adapun hasil akhir pengujian *cluster* dengan menggunakan perhitungan *Silhouette Coefficient*, nilai rata-rata keseluruhan *Silhouette Coefficient* yang dihasilkan adalah 0.846110. Berdasarkan hasil evaluasi yang diperoleh dengan menggunakan *Silhouette Coefficient* menunjukkan bahwa 2 *cluster* terpilih menjadi jumlah *cluster* terbaik dengan nilai *Silhouette Coefficient* sebesar 0.846110 yang termasuk dalam kategori struktur kuat. Semakin besar nilai *Silhouette Coefficient* maka semakin baik pula hasil dan menunjukkan *cluster* yang optimal. Berikut visualisasi hasil pengujian yang ditunjukkan pada Gambar 4.



Gambar 4. Visualisasi *Silhouette Coefficient*

4. KESIMPULAN

Berdasarkan hasil pengujian yang diperoleh dengan menggunakan *Silhouette Coefficient* menunjukkan bahwa 2 *cluster* terpilih menjadi jumlah *cluster* terbaik dengan nilai *Silhouette Coefficient* sebesar 0,846110 yang termasuk struktur *clustering* yang kuat. Semakin besar nilai *Silhouette Coefficient* maka semakin baik pula hasil dan menunjukkan *cluster* yang optimal. Dari 2 *cluster* yang diperoleh, *cluster* 1 merupakan kategori rawan tinggi yang terdiri dari 35 wilayah, sedangkan *cluster* 2 merupakan kategori rawan rendah yang terdiri dari 463 wilayah. Hasil penelitian ini dapat memberikan informasi bagi pihak terkait untuk meningkatkan efisiensi penanggulangan bencana banjir.

REFERENCES

- [1] S. Ulya, H. Hapidin, and Z. Akbar, "SIGANA Banjir: Game Edukasi Kesiapsiagaan Bencana Banjir Untuk Anak Usia 5-6 Tahun," *Murhum J. Pendidik. Anak Usia Dini*, vol. 4, no. 2, pp. 151–164, Aug. 2023, doi: 10.37985/murhum.v4i2.311.
- [2] M. Ridwan, A. Zainuddin, M. Kasim, M. Yahya, P. Studi Destinasi Pariwisata Politeknik Pariwisata Makassar Jalan Gunung Rinjani, and S. Selatan, "PEMETAAN DAERAH BENCANA PADA DESTINASI KOTA PALOPO (STUDI KASUS BENCANA BANJIR DAN LONGSOR) (Disaster Space Mapping in Palopo City Destinations (Case Study of Flood and Landslide Disaster)," 2022.
- [3] D. I. Ramadhani, O. Damayanti, O. Thaushiyah, and A. R. Kadafi, "Penerapan Metode K-Means Untuk Clustering Desa Rawan Bencana Berdasarkan Data Kejadian Terjadinya Bencana Alam," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 3, p. 749, Jun. 2022, doi: 10.30865/jurikom.v9i3.4326.
- [4] N. Amaliyah Wibowo and A. Maman Abadi, "ANALISIS TINGKAT KERAWANAN BENCANA ALAM BANJIR DI KABUPATEN PURBALINGGA DENGAN FUZZY LOGIC ANALYSIS OF FLOOD VULNERABILITY LEVEL IN PURBALINGGA REGENCY WITH FUZZY LOGIC," 2022.



- [5] M. D. H. Rahiem and F. Widiastuti, "Pembelajaran Mitigasi Bencana Alam Gempa Bumi untuk Anak Usia Dini melalui Buku Bacaan Bergambar," *J. Obs. J. Pendidik. Anak Usia Dini*, vol. 5, no. 1, p. 36, Apr. 2020, doi: 10.31004/obsesi.v5i1.519.
- [6] E. Kurniati, V. Adriany, M. Mirawati, R. M. El-Seira, and I. Winangsih, "Identifikasi Kesiapsiagaan Guru PAUD sebagai Upaya Pengurangan Risiko Bencana Banjir di Bandung," *J. Obs. J. Pendidik. Anak Usia Dini*, vol. 4, no. 2, p. 840, Feb. 2020, doi: 10.31004/obsesi.v4i2.388.
- [7] W. Wirmando, F. Patarru', and J. L. Saranga', "MENINGKATKAN PENGETAHUAN DAN KESIAPSIAGAAN MASYARAKAT DALAM MENGHADAPI BENCANA BANJIR MELALUI EDUKASI DAN SIMULASI MENGGUNAKAN TABLETOP DISASSTER EXERCISE," *JMM (Jurnal Masy. Mandiri)*, vol. 6, no. 3, p. 2166, Jun. 2022, doi: 10.31764/jmm.v6i3.8244.
- [8] A. Irfan Abdurrahman, B. Yuwono, and Y. Fauziah, "PENERAPAN METODE MULTI ATTRIBUTE UTILITY THEORY (MAUT) DALAM PEMETAAN TINGKAT DAMPAK BENCANA BANJIR DI KABUPATEN BANTUL," 2020.
- [9] H. Setiawan *et al.*, "ANALISIS PENYEBAB BANJIR DI KOTA SAMARINDA," 2020. [Online]. Available: <https://ejournal.upi.edu/index.php/gea>
- [10] B. Banjir, P. dan Pengendalian Pemanfaatan Ruang, P. U. dan Pengendalian Pemanfaatan Ruang Berdasarkan Penataan Ruang dan RUU Cipta Kerja, S. Nurhayati Qodriyatun, and P. R. Penelitian Badan Keahlian DPR Jl Gatot Subroto, "Sri Nurhayati Qodriyatun", doi: 10.22212/aspirasi.v11i1.1590.
- [11] N. Agusdianita and V. Karjiyati, "STUDI DESKRIPTIF SIKAP KESIAPSIAGAAN BANJIR ANAK SD DITINJAU DARI SEGI ETNIS DI DAS KOTA BENGKULU." [Online]. Available: <https://www.jurnalfai-uikabogor.org/attadib>
- [12] F. A. Setyorini, "Menakar Paradigma Penanggulangan Bencana Melalui Analisis Undang-Undang No. 24 Tahun 2007 Tentang Penanggulangan Bencana," *J. Soc. Polit. Gov.*, vol. 5, no. 2, pp. 97–113, 2023, doi: 10.24076/jspg.v5i2.1376.
- [13] R. A. Indraputra and R. Fitriana, "K-Means Clustering Data COVID-19".
- [14] W. Utomo, "The comparison of k-means and k-medoids algorithms for clustering the spread of the covid-19 outbreak in Indonesia," *Ilk. J. Ilm.*, vol. 13, no. 1, pp. 31–35, Apr. 2021, doi: 10.33096/ilkom.v13i1.763.31-35.
- [15] T. Hardiani, "Analisis Clustering Kasus Covid 19 di Indonesia Menggunakan Algoritma K-Means," *J. Nas. Pendidik. Tek. Inform.*, vol. 11, no. 2, pp. 156–165, Aug. 2022, doi: 10.23887/janapati.v11i2.45376.
- [16] A. Supriyadi *et al.*, "PERBANDINGAN ALGORITMA K-MEANS DENGAN K-MEDOIDS PADA PENGELOMPOKAN ARMADA KENDARAAN TRUK BERDASARKAN PRODUKTIVITAS."
- [17] D. D. Aulia, "COMPARISON PERFORMANCE OF K-MEDOIDS AND K-MEANS ALGORITHMS IN CLUSTERING COMMUNITY EDUCATION LEVELS Jurnal Nasional Pendidikan Teknik Informatika : JANAPATI | 274," vol. 12, no. 2, pp. 273–282, 2023.
- [18] M. M. Effendi and A. Siswandi, "Analysis Prediksi Wilayah Rawan Banjir dengan Algoritma K-Means," vol. 5, no. 2, pp. 697–703, 2024, doi: 10.47065/josh.v5i2.4770.
- [19] A. Ahyuna, M. Lasena, R. Aminuddin, and Z. Azhar, "Pembentukan Pola Peminjaman Buku Pada Perpustakaan Dengan Menerapkan Metode CART dan Normalisasi Z-Score," *Technol. Sci.*, vol. 6, no. 1, 2024, doi: 10.47065/bits.v6i1.5238.
- [20] S. E. Saqila, I. P. Ferina, and A. Iskandar, "Analisis Perbandingan Kinerja Clustering Data Mining Untuk Normalisasi Dataset," *J. Sist. Komput. dan Inform.*, vol. 5, no. 2, p. 356, Dec. 2023, doi: 10.30865/json.v5i2.6919.
- [21] F. P. Ferdy Pangestu, N. Y. Nur Yasin, R. C. Ronald Chistover Hasugian, and Y. Yunita, "Penerapan Algoritma K-Means Untuk Mengklasifikasi Data Obat," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 12, no. 1, pp. 53–62, Mar. 2023, doi: 10.32736/sisfokom.v12i1.1461.
- [22] S. A. D. Darmawan and Karmilasari, "PENERAPAN METODE K-MEANS CLUSTERING DAN SIMPLE MOVING AVERAGE UNTUK MEMPREDIKSI JENIS PENYAKIT DI PROVINSI JAWA TIMUR," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 4, pp. 877–886, Aug. 2024, doi: 10.25126/jtiik.1148703.
- [23] S. Kasus, : Hoyweapstore, D. Triyansyah¹, and D. Fitriannah², "Analisis Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing", doi: 10.22441/incomtech.v8i2.4174.
- [24] S. Multi Fani and R. Santoso, "PENERAPAN TEXT MINING UNTUK MELAKUKAN CLUSTERING DATA TWEET AKUN BLIBLI PADA MEDIA SOSIAL TWITTER MENGGUNAKAN K-MEANS CLUSTERING," vol. 10, pp. 583–593, [Online]. Available: <https://ejournal3.undip.ac.id/index.php/gaussian/>
- [25] N. Nugroho and F. D. Adhinata, "Penggunaan Metode K-Means dan K-Means++ Sebagai Clustering Data Covid-19 di Pulau Jawa," *Teknika*, vol. 11, no. 3, pp. 170–179, Oct. 2022, doi: 10.34148/teknika.v11i3.502.