



Analisis Pengelompokan Wilayah Berdasarkan Frekuensi Kejadian Banjir Menggunakan K-Means Clustering

Cinta Aurelya*, Yunus Widjaja

Fakultas Teknologi dan Desain, Program Studi Sistem Informasi, Universitas Pembangunan Jaya, Tangerang Selatan, Indonesia

Email: ^{1,*}cinta.aurelya@student.upj.ac.id, ²yunus.widjaja@upj.ac.id

Email Penulis Korespondensi: cinta.aurelya@student.upj.ac.id

Abstrak—Banjir merupakan salah satu bencana yang paling sering terjadi di Provinsi Jawa Barat dan berdampak pada berbagai aspek sosial maupun ekonomi. Meskipun pemerintah telah menyediakan data frekuensi kejadian banjir hingga tingkat desa dan kelurahan, pemanfaatannya untuk analisis kerawanan masih terbatas. Penelitian ini bertujuan untuk mengelompokkan wilayah berdasarkan tingkat frekuensi kejadian banjir menggunakan metode *K-Means Clustering* sebagai pendekatan analitik untuk menghasilkan pemetaan risiko yang lebih komprehensif. Data yang digunakan merupakan catatan kejadian banjir periode 2022–2024 di Provinsi Jawa Barat dari sumber resmi pemerintah. Analisis dilakukan berdasarkan satu variabel utama, yaitu frekuensi kejadian banjir, sesuai dengan ketersediaan data pada sumber resmi. Proses pengolahan data dilakukan melalui tahapan *Knowledge Discovery in Database* dengan fokus pada penerapan algoritma *K-Means*, sedangkan evaluasi model dilakukan menggunakan *Elbow Method* dan *Silhouette Score* untuk memastikan jumlah kluster yang optimal. Hasil penelitian menunjukkan bahwa tiga kluster memberikan struktur pengelompokan terbaik dengan distribusi risiko yang jelas. Kluster rendah terdiri dari wilayah yang tidak mengalami banjir sama sekali dengan frekuensi nol dan mencakup 12.454 wilayah. Kluster sedang berisi 3.404 wilayah dengan frekuensi kejadian banjir antara 1 hingga 6 kali. Sementara itu, kluster tinggi mencakup 76 wilayah yang mengalami banjir dengan frekuensi antara 7 hingga 33 kali. Temuan ini diharapkan dapat menjadi dasar dalam perencanaan mitigasi banjir, penyusunan tata ruang, pembangunan infrastruktur pengendali banjir, serta membantu masyarakat dalam mempertimbangkan tingkat keamanan wilayah tempat tinggal.

Kata Kunci: Banjir; K-Means Clustering; Klasterisasi; Mitigasi Bencana; Jawa Barat

Abstract—Floods are among the most frequent natural disasters occurring in West Java Province and have significant impacts on social and economic conditions. Although the government provides flood incident frequency data down to the village and sub-district levels, its utilization for detailed vulnerability analysis remains limited. This study aims to classify regions based on the frequency of flood events using the K-Means Clustering method as an analytical approach to produce a more comprehensive risk mapping. The dataset consists of flood incident records from 2022 to 2024 obtained from official government sources. The analytical process follows the stages of Knowledge Discovery in Database with a primary focus on the implementation of the K-Means algorithm, while model evaluation is conducted using the Elbow Method and Silhouette Score to determine the optimal number of clusters. The results indicate that three clusters provide the most structured grouping of flood risk. The low-risk cluster consists of 12,454 regions that experienced zero flood events. The medium-risk cluster includes 3,404 regions with flood frequencies ranging from 1 to 6 events. Meanwhile, the high-risk cluster comprises 76 regions with flood occurrences between 7 and 33 events. These findings are expected to support flood mitigation planning, spatial planning strategies, the development of flood-control infrastructure, and to assist communities in evaluating the safety of potential residential areas.

Keywords: Flood; K-Means Clustering; Clustering; Disaster Mitigation; West Java

1. PENDAHULUAN

Banjir merupakan salah satu bencana hidrometeorologi yang paling sering terjadi di Indonesia dan mempengaruhi berbagai aspek kehidupan masyarakat, mulai dari kerusakan infrastruktur, gangguan aktivitas ekonomi, hingga meningkatnya kerentanan sosial penduduk. Provinsi Jawa Barat menjadi salah satu wilayah dengan jumlah kejadian banjir yang tinggi karena dipengaruhi oleh kondisi topografi, intensitas curah hujan, serta tekanan urbanisasi yang menyebabkan perubahan tata guna lahan secara masif. Walaupun Pemerintah Provinsi Jawa Barat telah menyediakan data frekuensi kejadian banjir hingga tingkat desa/kelurahan melalui platform *open data*, data tersebut belum dimanfaatkan secara optimal untuk analisis kerawanan secara mendalam. Hal ini menyebabkan belum tersedianya pemetaan objektif yang mampu menggambarkan tingkat risiko banjir secara terukur dan berbasis data aktual, sehingga pemangku kebijakan dan masyarakat masih kesulitan menentukan prioritas mitigasi maupun pemilihan lokasi tempat tinggal. Kondisi ini mendorong kebutuhan akan suatu pendekatan analitik berbasis data yang mampu menghasilkan segmentasi wilayah berdasarkan pola kejadian banjir secara sistematis.

Dalam konteks analisis data kebencanaan, metode *data mining* menjadi pendekatan yang banyak digunakan karena kemampuannya dalam mengolah data berukuran besar dan mengidentifikasi pola tersembunyi yang tidak dapat dilihat melalui analisis konvensional [5]. Salah satu algoritma *data mining* yang paling umum digunakan adalah *K-Means Clustering*, yaitu metode pengelompokan data berdasarkan kedekatan jarak antar objek dalam ruang multidimensi [6]. *K-Means* bekerja dengan menentukan sejumlah kluster (k), kemudian mengelompokkan setiap data ke dalam kluster yang memiliki pusat (*centroid*) terdekat, sehingga data dengan karakteristik yang serupa akan terkumpul dalam satu kluster, sementara data dengan perbedaan signifikan akan dipisahkan pada kluster yang berbeda [6][7][8]. Keunggulan *K-Means* yang paling signifikan adalah kesederhanaan, efisiensi perhitungan, dan kemampuan adaptasinya yang efektif terhadap pemrosesan data bervolume besar, seperti yang sering ditemukan dalam analisis kebencanaan [9]. Oleh karena itu, algoritma ini banyak digunakan dalam penelitian kebencanaan untuk mengidentifikasi pola risiko dan mengelompokkan wilayah berdasarkan tingkat kerawanan.



Berbagai penelitian sebelumnya menunjukkan bahwa *K-Means* mampu memberikan hasil signifikan dalam pengelompokan risiko bencana. Oktaviany dkk. (2025) melakukan pengelompokan wilayah bencana banjir di Indonesia dan menunjukkan bahwa *K-Means* mampu memetakan kerentanan banjir secara terstruktur berdasarkan variabel dampak dan karakteristik wilayah [9]. Rohman dkk. (2024) menerapkan *K-Means* untuk melakukan klusterisasi bencana alam di wilayah Jawa Barat dengan menekankan pada identifikasi pola persebaran bencana di tingkat kabupaten/kota [10]. Penelitian lain oleh Khomsiyah dkk. (2021) juga berhasil menggunakan *K-Means* untuk mengelompokkan wilayah rawan banjir dengan memanfaatkan data kejadian banjir dan variabel geografis [11]. Rosaliyah dan Nurhakim (2023) melakukan pengelompokan kejadian bencana alam di Jawa Barat berdasarkan jenis bencana menggunakan algoritma *K-Means* untuk melihat distribusi kejadian pada berbagai kategori bencana [12]. Mariam (2023) memanfaatkan algoritma *K-Means* untuk menentukan prioritas penerima bantuan perumahan pasca bencana, yang menunjukkan kemampuan *K-Means* dalam membantu pengambilan keputusan berbasis data pada konteks kebencanaan [13].

Meskipun seluruh penelitian terdahulu tersebut menunjukkan bahwa algoritma *K-Means* efektif dalam proses pengelompokan data kebencanaan, masih terdapat beberapa celah yang belum terjawab. Sebagian penelitian berfokus pada pengelompokan bencana dalam cakupan luas seperti tingkat nasional maupun provinsi, sehingga belum menyediakan analisis kerawanan pada skala mikro seperti desa atau kelurahan, padahal skala tersebut lebih relevan untuk mitigasi berbasis komunitas [9], [10]. Selain itu, sebagian penelitian sebelumnya lebih banyak menggunakan variabel kategori bencana, intensitas dampak, atau karakteristik geografis, namun belum secara khusus memanfaatkan variabel frekuensi kejadian banjir aktual sebagai indikator empiris tingkat kerawanan wilayah [11], [12]. Beberapa penelitian memang telah menerapkan algoritma *K-Means*, tetapi tidak seluruhnya mengikuti tahapan *Knowledge Discovery in Database (KDD)* secara lengkap, sehingga proses penghasilannya belum sepenuhnya komprehensif [13]. Lebih jauh lagi, belum ada penelitian yang secara spesifik melakukan klusterisasi seluruh desa/kelurahan di Provinsi Jawa Barat berdasarkan data resmi pemerintah yang bersifat aktual dan mencakup keseluruhan entitas wilayah, sehingga hasil yang tersedia masih bersifat makro dan kurang aplikatif untuk kebutuhan pengambilan keputusan tingkat lokal.

Oleh karena itu, penelitian ini hadir untuk mengisi celah tersebut dengan memanfaatkan algoritma *K-Means Clustering* untuk mengelompokkan seluruh desa/kelurahan di Provinsi Jawa Barat berdasarkan frekuensi kejadian banjir periode 2022–2024. Pendekatan ini memanfaatkan data empiris kejadian banjir aktual sebagai dasar segmentasi wilayah sehingga menghasilkan pemetaan risiko banjir yang lebih realistis. Dengan menerapkan *K-Means Clustering*, proses analisis dilakukan secara terstruktur untuk menghasilkan pengelompokan wilayah berdasarkan kemiripan karakteristik banjir [14]. Diharapkan penelitian ini menghasilkan tiga kluster risiko yaitu kluster rendah, sedang, dan tinggi yang dapat digunakan oleh berbagai pihak. Pemerintah dapat menggunakan informasi ini untuk menentukan prioritas pembangunan infrastruktur pengendali banjir, memperkuat sistem peringatan dini, serta menyusun tata ruang yang selaras dengan karakteristik kerawanan wilayah. Sementara itu, masyarakat dapat memanfaatkan hasil klusterisasi ini sebagai pertimbangan dalam memilih lokasi tempat tinggal yang lebih aman.

Dengan demikian, tujuan utama penelitian ini adalah menghasilkan model klusterisasi kerawanan banjir menggunakan *K-Means* berbasis data frekuensi kejadian banjir yang komprehensif dan aplikatif. Penelitian ini diharapkan memberikan kontribusi signifikan baik secara ilmiah maupun praktis dalam mendukung mitigasi banjir di Provinsi Jawa Barat.

2. METODOLOGI PENELITIAN

2.1 Sumber dan Jenis Data

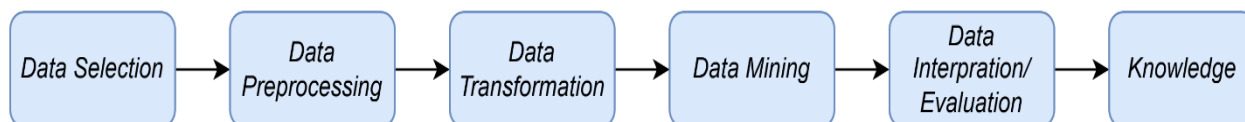
Data yang digunakan merupakan data sekunder yang diperoleh dari *website* resmi pemerintah yaitu Open Data Jabar <https://opendata.jabarprov.go.id/>, mengenai jumlah kejadian banjir pada tingkat desa/kelurahan di Provinsi Jawa Barat.

2.2 Teknik Pengumpulan Data

Data dikumpulkan melalui metode dokumentasi dengan mengunduh dataset resmi dari *website* Open Data Jabar, yang menampilkan 16 atribut dan 31.870 *records* dari tahun 2019-2024.

2.3 Tahapan Penelitian

Penelitian ini menggunakan metode *Knowledge Discovery in Database (KDD)* dalam prosesnya. KDD merupakan rangkaian tahapan untuk mengumpulkan dan memanfaatkan data guna menemukan pola atau pengetahuan tertentu dari himpunan data yang berukuran sangat besar [17]. Berikut merupakan langkah dari proses KDD, Tahapan proses tersebut dapat ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Proses KDD



2.3.1 Data Selection

Data selection adalah proses pemilihan data yang relevan dari kumpulan data yang lebih besar untuk digunakan dalam penelitian [7]. Tujuan dari *data selection* adalah memastikan bahwa hanya data yang berkualitas sehingga hasil analisis menjadi lebih efektif dan akurat. Pada penelitian ini, seleksi data difokuskan pada jumlah kejadian bencana banjir tahun 2022-2024 berdasarkan desa/kelurahan di Jawa Barat.

2.3.2 Data Preprocessing

Tahap *Data Preprocessing* dilakukan untuk memastikan bahwa data berada dalam kondisi bersih, konsisten, dan siap digunakan pada proses klusterisasi. Dataset awal terdiri dari 16 atribut yang mencakup informasi administratif serta data jumlah kejadian banjir per tahun. Pada tahap ini, dataset yang mengandung nilai kosong (*missing*) akan dihapus. Sehingga, dataset yang dihasilkan akan lebih bersih dan siap untuk diproses pada tahap berikutnya [12]. Selain itu, dilakukan seleksi atribut untuk memilih variabel yang relevan dengan tujuan penelitian. Dari keseluruhan atribut, hanya variabel numerik terkait kejadian banjir per tahun serta informasi identitas wilayah seperti nama desa/kelurahan yang dipertahankan, sementara atribut lain yang tidak berkontribusi terhadap proses klusterisasi dieliminasi. Selanjutnya, data numerik multivariabel yang memuat jumlah kejadian banjir pada beberapa tahun digabungkan menjadi satu nilai representatif berupa total frekuensi kejadian banjir di tiap wilayah. Transformasi ini diperlukan untuk menyederhanakan struktur data dan memfokuskan analisis pada tingkat kerawanan banjir secara keseluruhan. Dengan langkah-langkah tersebut, dataset dipersiapkan agar berada dalam kondisi optimal sebelum memasuki tahap transformasi dan pemodelan.

2.3.3 Data Transformation

Pada tahap ini adalah proses mengubah atau menggabungkan data ke dalam format yang sesuai dengan kebutuhan algoritma *data mining* yang akan diterapkan. Setiap algoritma *data mining* memiliki persyaratan struktur data tertentu, sehingga proses transformasi diperlukan agar data dapat diolah secara optimal [18].

2.3.4 Data Mining

Tahap data mining merupakan inti dari proses pencarian informasi dari dataset yang digunakan [7]. Pada penelitian ini, proses klusterisasi dilakukan dengan menerapkan algoritma *K-Means* menggunakan bahasa pemrograman *Python* dengan bantuan pustaka *Scikit-learn*. Variabel utama yang menjadi input klusterisasi adalah frekuensi kejadian banjir pada setiap desa/kelurahan. Nilai frekuensi ini digunakan sebagai dasar untuk mengelompokkan wilayah berdasarkan tingkat intensitas kejadian banjir. Proses ini dimulai dengan menentukan jumlah kluster (k), kemudian algoritma akan memilih pusat kluster (*centroid*) secara awal, menghitung jarak setiap data terhadap *centroid*, dan menempatkan data pada kluster terdekat [5][10]. Selanjutnya, *centroid* diperbarui berdasarkan rata-rata anggota kluster, dan proses tersebut diulang hingga posisi kluster mencapai kondisi yang stabil. Dengan demikian, *K-Means* mampu menghasilkan pengelompokan yang menunjukkan pola, kecenderungan, serta tingkat kerawanan banjir antar wilayah secara lebih jelas dan terstruktur.

2.3.5 Data Interpretation/Evaluation

Pada tahap *evaluation* atau *interpretation*, hasil klusterisasi yang diperoleh dari proses *data mining* dievaluasi untuk memastikan bahwa pola atau informasi yang ditemukan benar-benar relevan, valid, dan sesuai dengan tujuan penelitian. Pada penelitian ini, kualitas hasil klusterisasi dinilai menggunakan *Elbow Method* dan *Silhouette Score* untuk memastikan bahwa jumlah kluster yang dipilih benar-benar optimal. *Elbow Method* digunakan untuk mengamati nilai *Within-Cluster Sum of Squares* (WCSS) dan menentukan titik siku yang menunjukkan jumlah kluster dengan tingkat efisiensi pemisahan terbaik [19] [20]. Setelah itu, *Silhouette Score* digunakan untuk mengevaluasi seberapa baik setiap data berada pada klasternya masing-masing, sehingga dapat terlihat apakah pemisahan antar kluster cukup jelas dan tidak tumpang tindih [19] [20]. Melalui proses evaluasi ini, hasil pengelompokan tidak hanya akurat secara komputasional, tetapi juga logis ketika dikaitkan dengan distribusi frekuensi banjir tahun 2022–2024 di Jawa Barat.

2.3.6 Knowledge

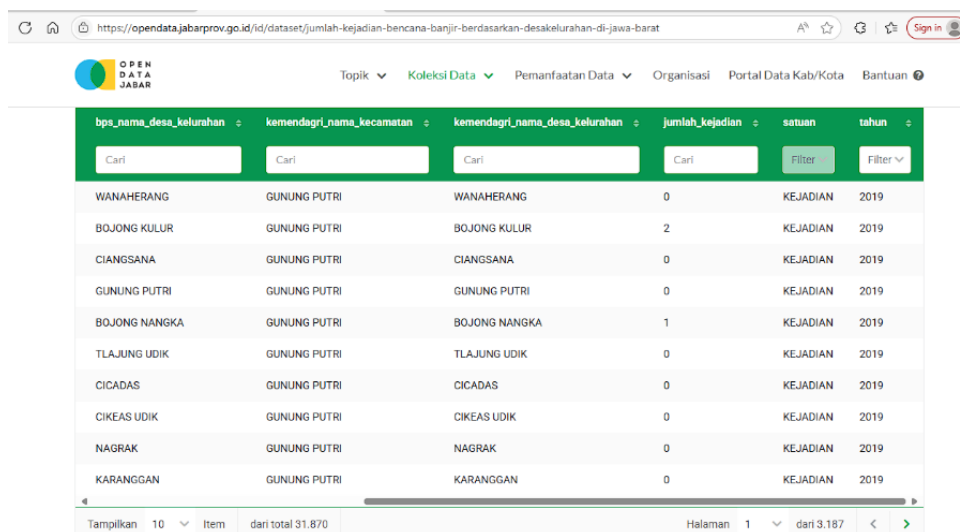
Tahap *knowledge* merupakan proses akhir di mana hasil evaluasi yang telah divalidasi diterjemahkan menjadi pengetahuan yang dapat dimanfaatkan secara praktis dan dapat dijadikan dasar kuat untuk mengambil keputusan [21][22]. Pada tahap ini, pola atau kluster yang ditemukan diinterpretasikan menjadi informasi strategis yang dapat dimanfaatkan oleh pemangku kepentingan, seperti identifikasi wilayah berisiko tinggi, rekomendasi prioritas mitigasi, atau arahan perencanaan tata ruang.

3. HASIL DAN PEMBAHASAN

Hasil pada penelitian ini didapat dengan memanfaatkan data jumlah kejadian bencana banjir berdasarkan desa/kelurahan di Jawa Barat yang bersumber dari Open Data Provinsi Jawa Barat dengan menerapkan algoritma *K-Means Clustering* menggunakan tools *Visual Studio Code*. Metode yang digunakan dalam penelitian ini menggunakan *Knowledge Discovery in Database*, langkah-langkah tersebut meliputi:

3.1 Data Selection

Penelitian ini menggunakan *dataset* terbuka mengenai jumlah kejadian bencana banjir di Provinsi Jawa Barat. Data diperoleh dari sumber resmi Pemerintah Provinsi Jawa Barat (*Open Data Jawa Barat*). Berikut adalah tampilan data yang didapatkan dari *website* tersebut, seperti yang dapat dilihat pada Gambar 2.



bps_nama_desa_kelurahan	kemendagri_nama_kecamatan	kemendagri_nama_desa_kelurahan	jumlah_kejadian	satuan	tahun
WANAHERANG	GUNUNG PUTRI	WANAHERANG	0	KEJADIAN	2019
BOJONG KULUR	GUNUNG PUTRI	BOJONG KULUR	2	KEJADIAN	2019
CIANGSANA	GUNUNG PUTRI	CIANGSANA	0	KEJADIAN	2019
GUNUNG PUTRI	GUNUNG PUTRI	GUNUNG PUTRI	0	KEJADIAN	2019
BOJONG NANGKA	GUNUNG PUTRI	BOJONG NANGKA	1	KEJADIAN	2019
TLAJUNG UDIK	GUNUNG PUTRI	TLAJUNG UDIK	0	KEJADIAN	2019
CICADAS	GUNUNG PUTRI	CICADAS	0	KEJADIAN	2019
CIKEAS UDIK	GUNUNG PUTRI	CIKEAS UDIK	0	KEJADIAN	2019
NAGRAK	GUNUNG PUTRI	NAGRAK	0	KEJADIAN	2019
KARANGGAN	GUNUNG PUTRI	KARANGGAN	0	KEJADIAN	2019

Gambar 2. Data Jumlah Kejadian Bencana Banjir Provinsi Jawa Barat dari Website Open Data Jabar

Pada tahap ini, data awal yang berjumlah sebanyak 31.870 akan disaring menjadi 15.934 data berdasarkan frekuensi kejadian banjir pada periode 2022–2024. Fokus seleksi data adalah pada variabel tunggal (*jumlah_kejadian*) sebagai fitur utama untuk *clustering*, sementara variabel identifikasi spasial (kode dan nama Desa/Kelurahan/Kabupaten/Kota) dipertahankan sebagai atribut deskriptif.

3.2 Data Preprocessing

Tahap ini berfokus pada pemilihan atribut dan pembersihan data. Atribut yang digunakan yaitu *jumlah_kejadian* sebagai fitur numerik utama untuk *clustering*, dan nama desa/kelurahan sebagai label identifikasi spasial. Atribut yang tidak relevan (seperti kode administrasi, nama provinsi, dan satuan) dibuang untuk menyederhanakan model. Selain itu, berdasarkan pemeriksaan, mengkonfirmasi bahwa data tidak mengandung *missing values* (nilai yang hilang) pada variabel *jumlah_kejadian*, sehingga data siap untuk proses transformasi.

3.3 Data Transformation

Tahap ini, yang bertujuan mempersiapkan data untuk pemodelan, dilakukan proses transformasi data untuk memastikan dataset berada dalam kondisi yang optimal sebelum dilakukan pemodelan menggunakan algoritma *K-Means*. Meskipun variabel fitur utama, yaitu *jumlah_kejadian*, telah bertipe numerik dan tidak memiliki *missing values*, proses normalisasi tetap diperlukan. Hal ini dikarenakan nilai frekuensi kejadian banjir memiliki rentang yang bervariasi antar wilayah, sehingga berpotensi memengaruhi perhitungan jarak (*Euclidean distance*) dalam *K-Means*. Oleh karena itu, pada tahap ini diterapkan normalisasi menggunakan *StandardScaler* dari pustaka *Scikit-learn* untuk menyeragamkan skala variabel agar setiap titik data memiliki kontribusi yang seimbang dalam proses klusterisasi. Dengan demikian, hasil transformasi data pada tahap ini menghasilkan dataset yang telah dinormalisasi dan siap digunakan pada tahap Data Mining.

3.4 Pemodelan K-Means Clustering

Pada tahap ini dilakukan proses *K-Means Clustering* terhadap data kejadian banjir yang telah melalui proses pembersihan serta penyiapan data sebelumnya. Data yang digunakan merupakan catatan jumlah kejadian banjir pada tingkat desa/kelurahan di Provinsi Jawa Barat untuk periode tahun 2022 hingga 2024. Untuk menyederhanakan dimensi temporal pada data mentah, seluruh nilai kejadian per tahun digabungkan dan disesuaikan sehingga menghasilkan satu nilai representatif berupa total frekuensi kejadian banjir di setiap wilayah. Pendekatan ini dilakukan agar analisis berfokus pada tingkat intensitas kejadian secara keseluruhan, bukan pada variasi antar tahun, sehingga karakter kerawanan banjir setiap desa/kelurahan dapat teridentifikasi secara lebih jelas. Proses klusterisasi dilakukan menggunakan bahasa pemrograman *Python* dan pustaka *Scikit-learn*, yang menyediakan fungsi *K-Means* untuk mengelompokkan wilayah berdasarkan kesamaan frekuensi kejadian banjir. Melalui proses inilah terbentuk kluster-kluster yang dapat merepresentasikan tingkat risiko banjir secara terstruktur.

Atribut utama yang digunakan dalam proses klusterisasi adalah frekuensi kejadian banjir sebagai indikator tingkat kerentanan wilayah. Sementara itu, informasi administratif seperti nama desa/kelurahan dan kecamatan hanya berfungsi sebagai identitas wilayah untuk kebutuhan interpretasi dan visualisasi hasil klusterisasi. Setelah data diformat ulang, dilakukan proses normalisasi menggunakan *StandardScaler* untuk memastikan seluruh atribut berada pada skala yang



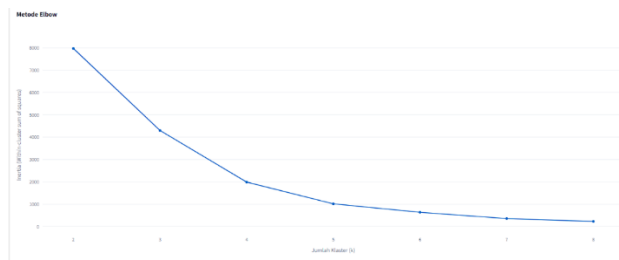
sebanding. Normalisasi ini penting karena frekuensi kejadian banjir memiliki rentang nilai yang cukup lebar dan berpotensi mendominasi proses pengelompokan jika tidak distandardisasi. Mengingat *K-Means* sangat bergantung pada perhitungan jarak antar titik, penyetaraan skala menjadi langkah krusial untuk menghasilkan kluster yang lebih akurat.

Tahap selanjutnya adalah penentuan jumlah kluster optimal dengan memanfaatkan pendekatan *Elbow Method* yang dihasilkan dari nilai inerti pada berbagai variasi jumlah kluster. Setelah jumlah kluster optimal ditetapkan, algoritma *K-Means* diterapkan untuk membentuk kelompok wilayah berdasarkan pola kemiripan frekuensi kejadian banjir. Hasil klusterisasi kemudian dievaluasi menggunakan *Silhouette Score* untuk menilai kualitas pemisahan antar kluster serta konsistensi struktur kluster yang terbentuk. Melalui rangkaian proses tersebut, diperoleh pengelompokan wilayah yang merepresentasikan tingkat kerawanan banjir secara lebih sistematis dan komprehensif.

3.5 Data Interpration/Evaluation

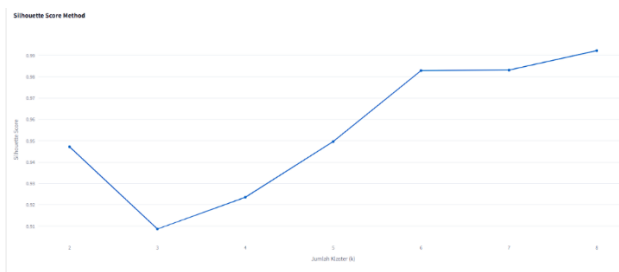
Evaluasi model *K-Means* dilakukan menggunakan dua pendekatan, yaitu *Elbow Method* dan *Silhouette Score*, untuk memastikan pemilihan jumlah kluster yang paling representatif terhadap pola distribusi frekuensi kejadian banjir. Kedua metode ini memberikan gambaran yang saling melengkapi terkait kualitas pemisahan antar kluster serta stabilitas struktur kluster yang terbentuk.

Pada Gambar 3 terlihat grafik *Elbow Method*, nilai inerti menunjukkan penurunan yang tajam pada $k = 2$ hingga $k = 4$, kemudian mengalami perlambatan penurunan pada k di atas 4. Titik tekuk (*elbow point*) terlihat berada pada kisaran $k = 3$, yang menandakan bahwa penambahan jumlah kluster setelah nilai tersebut tidak lagi memberikan pengurangan inerti yang signifikan. Dengan demikian, tiga kluster dianggap sebagai pilihan yang efisien karena mampu menyeimbangkan antara tingkat pemisahan dan kompleksitas model.



Gambar 3. Grafik Elbow Method untuk Penentuan Jumlah Kluster Optimal pada Data Kejadian Banjir

Sementara itu, evaluasi menggunakan *Silhouette Score* pada Gambar 4 menunjukkan pola yang bertahap meningkat seiring dengan bertambahnya jumlah kluster, dengan nilai tertinggi berada pada $k = 8$. Namun demikian, meskipun skor di k yang lebih tinggi tampak lebih baik, interpretasi substantif terhadap konteks data harus menjadi pertimbangan utama. Dalam kasus ini, frekuensi kejadian banjir di tingkat desa/kelurahan memiliki distribusi yang sangat timpang, di mana sebagian besar wilayah tidak mengalami banjir dan hanya sebagian kecil wilayah memiliki frekuensi yang tinggi. Kondisi tersebut menjadikan kluster dengan jumlah terlalu besar ($k > 4$) cenderung menghasilkan pemisahan yang tidak bermakna secara kebijakan, serta berpotensi menimbulkan *overfitting*.



Gambar 4. Grafik Silhouette Score terhadap Variasi Jumlah Kluster pada Data Kejadian Banjir

Tabel 1. Skor Elbow dan Silhouette

	k	Inertia	Silhouette Score
0	2	7968.2653	0.9472
1	3	4297.6503	0.9087
2	4	1988.921	0.9236
3	5	1013.8437	0.9497
4	6	638.4714	0.9829
5	7	356.665	0.9831
6	8	230.2917	0.9923

Dengan mempertimbangkan kedua indikator tersebut, serta kebutuhan untuk memperoleh segmentasi yang sederhana, stabil, dan mudah diinterpretasikan, maka nilai $k = 3$ dipilih sebagai jumlah kluster optimal. Pemilihan ini juga

konsisten dengan tabel ringkasan klaster yang menunjukkan terbentuknya tiga kelompok wilayah dengan karakteristik risiko banjir yang berbeda secara jelas. Berdasarkan Tabel 2, klaster 0 terdiri dari wilayah dengan frekuensi kejadian banjir sangat rendah hingga menengah, klaster 1 mencakup wilayah yang tidak mengalami kejadian banjir, sedangkan klaster 2 merepresentasikan wilayah dengan frekuensi kejadian banjir relatif tinggi. Perbedaan karakteristik antar klaster tersebut menunjukkan bahwa pemilihan $k = 3$ mampu menangkap variasi tingkat risiko banjir secara lebih representatif dibandingkan jumlah klaster lainnya.

Pemilihan $k = 3$ tidak hanya didukung oleh keseimbangan antara nilai *inertia* dan *silhouette score*, tetapi juga memberikan segmentasi yang lebih relevan secara substantif, khususnya dalam konteks analisis risiko dan penentuan prioritas mitigasi bencana. Hasil evaluasi ini menunjukkan bahwa model *K-Means* mampu mengidentifikasi struktur alami dalam data kejadian banjir secara efektif, sehingga menghasilkan pengelompokan wilayah yang bermakna dan mudah dipahami. Dengan demikian, hasil klasterisasi ini dapat dimanfaatkan sebagai dasar pendukung pengambilan keputusan, terutama dalam perencanaan strategi penanggulangan banjir dan pengalokasian sumber daya yang lebih tepat sasaran.

Tabel 2. Ringkasan Klaster ($K=3$)

Cluster	Rata-rata Frekuensi	Min Frekuensi	Max Frekuensi	Jumlah Wilayah
0	1.3800	1	6	3404
1	0.0000	0	0	12454
2	12.3200	7	33	76

3.6 Knowledge

Hasil penerapan algoritma *K-Means* dengan jumlah klaster optimal $k = 3$ menghasilkan tiga kelompok wilayah dengan karakteristik frekuensi kejadian banjir yang berbeda. Visualisasi melalui bubble chart, grafik distribusi jumlah wilayah per klaster, serta contoh daftar wilayah dari masing-masing klaster memberikan gambaran yang lebih komprehensif mengenai pola persebaran risiko banjir pada tingkat desa/kelurahan.

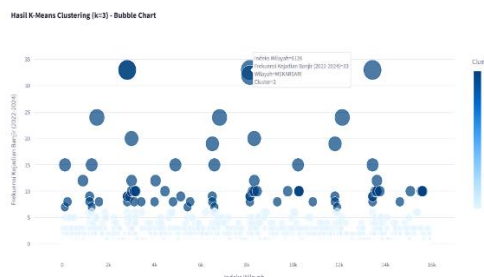
Berdasarkan Gambar 5, hasil visualisasi bubble chart ini menunjukkan adanya pola pengelompokan yang jelas antara ketiga klaster. Sumbu horizontal menampilkan indeks wilayah sebagai representasi posisi atau identitas masing-masing desa/kelurahan, sedangkan sumbu vertikal menunjukkan frekuensi kejadian banjir selama periode 2022–2024. Ukuran gelembung mencerminkan besarnya frekuensi banjir, dan warna digunakan untuk membedakan klaster.

Cluster 1 (warna paling terang) didominasi oleh gelembung kecil yang berkumpul pada bagian bawah grafik, tepat pada frekuensi 0, menandakan wilayah yang tidak mengalami banjir sama sekali. Penyebaran gelembung cluster ini sangat luas, menunjukkan bahwa wilayah tanpa banjir tersebar merata di seluruh indeks wilayah.

Cluster 0 ditandai dengan gelembung berukuran sedang yang berada pada rentang frekuensi 1 hingga 6 kali. Gelembung cluster ini cenderung tersebar di berbagai posisi indeks wilayah, menunjukkan bahwa wilayah dengan risiko banjir menengah tidak terpusat pada lokasi tertentu. Pola ini memperlihatkan bahwa wilayah yang sesekali mengalami banjir muncul secara acak dan tidak membentuk konsentrasi geografis yang kuat.

Sementara itu, *Cluster 2* (warna paling gelap) dengan ukuran gelembung terbesar terlihat jelas menonjol pada bagian atas grafik, yaitu frekuensi 7 hingga 33 kejadian. Meskipun jumlah wilayah dalam klaster ini relatif sedikit, gelembung-gelembung tersebut membentuk kelompok yang terpisah dari dua klaster lainnya. Hal ini mengindikasikan adanya wilayah-wilayah spesifik yang secara konsisten mengalami kejadian banjir dengan frekuensi yang jauh lebih tinggi dibandingkan wilayah lain.

Selain itu, pola persebaran ini menguatkan temuan bahwa karakteristik banjir di Provinsi Jawa Barat tidak hanya dipengaruhi oleh faktor geografis, tetapi juga oleh kondisi lingkungan dan tingkat kerentanan infrastruktur di masing-masing wilayah. Perbedaan antara *Cluster 1*, *Cluster 0*, dan *Cluster 2* memperlihatkan adanya gap risiko yang penting untuk diperhatikan dalam perencanaan mitigasi. Wilayah yang termasuk dalam *Cluster 2* terutama membutuhkan perhatian khusus dari pemerintah daerah, baik melalui peningkatan sistem drainase, penguatan program normalisasi sungai, maupun edukasi kesiapsiagaan masyarakat. Sebaliknya, wilayah dalam *Cluster 1* dapat dijadikan acuan untuk memahami faktor-faktor yang mampu menekan risiko banjir sehingga dapat direplikasi ke wilayah lain. Visualisasi ini secara keseluruhan memberikan dasar yang kuat untuk analisis lanjutan dalam perumusan strategi penanggulangan banjir yang lebih tepat sasaran.



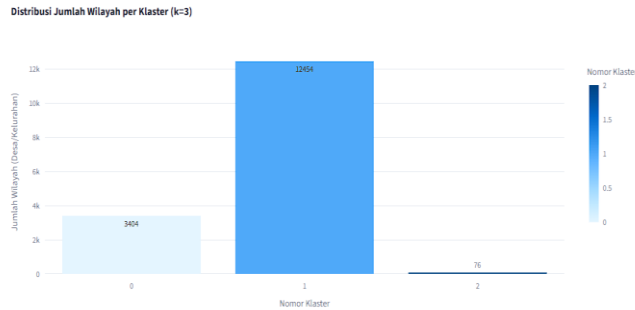
Gambar 5. Visualisasi Hasil Clustering Menggunakan Bubble Chart



Grafik distribusi jumlah wilayah per kluster menunjukkan proporsi jumlah desa/kelurahan yang masuk ke masing-masing kelompok. Pada Gambar 6, tampak bahwa *Cluster 1* mendominasi dengan jumlah wilayah terbanyak, yaitu 12.454 wilayah, atau mayoritas dari keseluruhan data. Hal ini menunjukkan bahwa sebagian besar wilayah di Jawa Barat tidak mengalami banjir selama periode 2022–2024.

Selanjutnya, *Cluster 0* terdiri dari 3.404 wilayah, yang mencerminkan kelompok daerah yang memiliki frekuensi banjir rendah hingga menengah. Jumlah wilayah di cluster ini cukup besar, menandakan bahwa kejadian banjir sporadis merupakan fenomena yang lazim namun tidak ekstrem.

Cluster 2, dengan 76 wilayah, merupakan kelompok terkecil tetapi paling signifikan dari perspektif mitigasi risiko. Grafik menunjukkan bar yang jauh lebih kecil dibanding dua kluster lain, namun kelompok ini merupakan prioritas utama karena wilayah-wilayah tersebut mengalami frekuensi banjir yang tinggi dan berulang.



Gambar 6. Distribusi Jumlah Wilayah dalam Setiap Cluster

Tabel 3 menampilkan beberapa contoh wilayah yang termasuk dalam *Cluster 0*, yaitu kelompok wilayah dengan frekuensi kejadian banjir sedang, berada pada rentang 1 hingga 6 kejadian selama periode 2022–2024. Wilayah-wilayah seperti Cileunyi, Wetan, Jatiseng, Ciasem Hilir, dan Mandalasari menunjukkan pola kejadian banjir yang muncul secara sporadis, namun tidak cukup sering untuk dianggap sebagai daerah rawan. Karakteristik cluster ini menggambarkan wilayah yang memiliki tingkat paparan banjir yang cukup signifikan sehingga tetap memerlukan perhatian, tetapi belum termasuk kategori ancaman tinggi. Jumlah wilayah dalam cluster ini cukup besar (3.404 wilayah), menandakan bahwa banjir dengan intensitas menengah merupakan fenomena yang relatif umum terjadi di banyak wilayah.

Tabel 3. Contoh Wilayah Hasil Cluster 0

Wilayah	Frekuensi Kejadian	Cluster
Cileunyi	6	0
Wetan	5	0
Jatiseng	4	0
Ciasem Hilir	3	0
Jatiserang	2	0
Mandalasari	2	0

Tabel 4 menyajikan contoh wilayah yang masuk ke dalam *Cluster 1*, yaitu kelompok wilayah dengan frekuensi kejadian banjir nol (0). Daerah seperti Karangmulya, Putrapinggan, Emplak, Bagolo, dan Ciparakan tidak menunjukkan adanya catatan kejadian banjir selama periode analisis. *Cluster* ini merupakan kelompok terbesar, yaitu mencakup 12.454 wilayah, yang menunjukkan bahwa sebagian besar wilayah di dataset relatif aman dari kejadian banjir. Keberadaan *cluster* ini sangat penting sebagai pembandingan terhadap wilayah rawan, dan mengindikasikan bahwa kondisi topografi atau sistem drainase di sebagian besar wilayah berada dalam kondisi yang cukup baik atau berada di area yang memang tidak rentan banjir.

Tabel 4. Contoh Wilayah Hasil Cluster 1

Wilayah	Frekuensi Kejadian	Cluster
Karangmulya	0	1
Putrapinggan	0	1
Emplak	0	1
Bagolo	0	1
Ciparakan	0	1

Tabel 5 menampilkan contoh wilayah dari *Cluster 2*, yaitu kelompok wilayah dengan frekuensi kejadian banjir tinggi, berada pada rentang 7 hingga 33 kali. Wilayah seperti Mekarsari, Gumegarsari, Leles, Karangasem, dan Dayeuhkolot menunjukkan intensitas kejadian banjir yang jauh lebih tinggi daripada dua *cluster* lainnya. *Cluster* ini hanya terdiri dari 76 wilayah, namun memiliki signifikansi strategis karena wilayah-wilayah ini tergolong sangat rawan terhadap



banjir. Kejadian yang berulang dengan frekuensi tinggi mengindikasikan kemungkinan adanya permasalahan struktural seperti kedekatan dengan aliran sungai, kapasitas drainase yang rendah, atau faktor lingkungan lain yang memperkuat risiko banjir.

Tabel 5. Contoh Wilayah Hasil Cluster 2

Wilayah	Frekuensi Kejadian	Cluster
Mekarsari	33	2
Gunungsari	32	2
Leles	24	2
Karangasem	20	2
Dayeuhkolot	19	2

4. KESIMPULAN

Penelitian ini berhasil memetakan tingkat kerawanan banjir pada tingkat desa/kelurahan di Provinsi Jawa Barat dengan menerapkan metode *K-Means Clustering* terhadap data frekuensi kejadian banjir periode 2022–2024. Hasil klusterisasi menghasilkan tiga kelompok utama yang menunjukkan perbedaan tingkat risiko banjir secara jelas. *Cluster 1* diidentifikasi sebagai kategori risiko rendah, terdiri dari 12.454 wilayah dengan minimal frekuensi 0 dan maksimal 0, menandakan wilayah yang tidak mengalami kejadian banjir sama sekali selama periode penelitian. *Cluster 0* termasuk kategori risiko sedang, mencakup 3.404 wilayah dengan minimal frekuensi 1 dan maksimal 6, yang menunjukkan bahwa wilayah-wilayah tersebut mengalami banjir secara sporadis namun tidak dalam intensitas tinggi. Sementara itu, *Cluster 2* merupakan kategori risiko tinggi, terdiri dari 76 wilayah dengan minimal frekuensi 7 dan maksimal 33, menandai wilayah yang secara konsisten mengalami kejadian banjir berulang dan memiliki tingkat kerawanan tertinggi. Hasil klusterisasi ini memberikan kontribusi penting bagi perencanaan kebijakan pengurangan risiko bencana di Jawa Barat. Informasi klaster dapat dimanfaatkan untuk menentukan prioritas pembangunan infrastruktur pengendali banjir, memperkuat sistem peringatan dini, serta menyusun kebijakan tata ruang yang selaras dengan karakteristik kerawanan wilayah. Selain itu, temuan penelitian ini juga bermanfaat bagi masyarakat umum sebagai acuan dalam mempertimbangkan lokasi tempat tinggal berdasarkan tingkat risiko banjir di wilayah mereka. Secara keseluruhan, penelitian ini diharapkan memberikan landasan ilmiah yang kuat bagi pemerintah, pemangku kebijakan, dan masyarakat dalam memahami pola kerawanan banjir serta mengambil langkah mitigasi yang lebih efektif dan berkelanjutan. Untuk penelitian selanjutnya, disarankan agar model klusterisasi dikembangkan dengan memasukkan variabel tambahan seperti curah hujan, ketinggian wilayah, kepadatan penduduk, atau kondisi tata guna lahan. Variabel-variabel tersebut berpotensi memberikan gambaran yang lebih komprehensif mengenai tingkat kerawanan banjir, namun belum dapat digunakan dalam penelitian ini karena tidak tersedia pada sumber data resmi yang digunakan. Penggunaan variabel yang lebih beragam di masa mendatang diharapkan dapat meningkatkan akurasi pemetaan risiko serta menghasilkan model klasifikasi yang lebih andal.

REFERENCES

- [1] S. Supriyadi and A. G. Malau, "Clustering dengan Menggunakan K-Means untuk Analisa Dampak Banjir," *J. Impresi Indones.*, vol. 4, no. 9, pp. 3428–3436, 2025, doi: 10.58344/jii.v4i9.6998.
- [2] A. Alamsyah and E. Wahyudi, "Transformasi Digital untuk Mitigasi Banjir: Optimalisasi Sistem Informasi di Jawa Barat," *J. Perlindungan Masy. Best. Praesidium*, vol. 1, no. 2, pp. 50–62, 2024, [Online]. Available: <https://ejournal.ipdn.ac.id/jpa/article/view/5102>
- [3] M. F. Al Halik and L. Septiana, "Analisa Data Untuk Prediksi Daerah Rawan Bencana Alam Di Jawa Barat Menggunakan Algoritma K-Means Clustering," *J. Inf. Syst. Applied, Manag. Account. Res.*, vol. 6, no. 4, pp. 856–870, 2022, doi: 10.52362/jisamar.v6i4.939.
- [4] M. Murdiaty, A. Angela, and C. Sylvia, "Pengelompokan Data Bencana Alam Berdasarkan Wilayah, Waktu, Jumlah Korban dan Kerusakan Fasilitas Dengan Algoritma K-Means," *J. Media Inform. Budidarma*, vol. 4, no. 3, p. 744, 2020, doi: 10.30865/mib.v4i3.2213.
- [5] Rossa Amelia Manik and Atik Ariesta, "Data Mining Untuk Klusterisasi Provinsi Di Indonesia Berdasarkan Data Pokok Pendidikan Nasional," *J. Ticom Technol. Inf. Commun.*, vol. 11, no. 3, pp. 159–164, 2023, doi: 10.70309/ticom.v11i3.115.
- [6] T. Asy Aria, M. Julkarnain, and F. Hamdani, "Penerapan Algoritma K-Means Clustering Untuk Data Obat," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 1, pp. 649–657, 2023, doi: 10.30865/klik.v4i1.1117.
- [7] M. A. Aldi, Z. Fatah., "IMPLEMENTASI K-MEANS CLUSTERING DALAM PENGELOMPOKAN DATA KUNJUNGAN WISATAWAN ASING DI INDONESIA," *Jurnal JIMI: Jurnal Ilmiah Multidisiplin Ilmu*, vol. 2, no. 1, pp. 13–19, 2025, doi: <https://doi.org/10.69714/3hhfj353>.
- [8] Y. A. Ilsa Hidayat, Eva Damila, "Clustering Zonasi Daerah Rawan Bencana Alam di Kabupaten Mandailing Natal menggunakan Algoritma K-Means," *J. Teknol. Terap.*, vol. 7, no. 3, p. 9, 2023, doi: <https://doi.org/10.33379/gtech.v7i3.2880>.
- [9] 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, 2025, doi: 10.47065/bulletincsr.v5i4.608.
- [10] D. Rohman, R. Annisa, D. Indriyana Efendi, and D. Solahudin, "Clustering Bencana Alam Menggunakan K-Means Pada Wilayah Jawa Barat," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 8, no. 1, pp. 493–500, 2024, doi: 10.36040/jati.v8i1.8409.
- [11] J. Khomsiyah, A. Ramdhani, A. Damayanti, and D. Rohman, "PENERAPAN ALGORITMA K-MEANS CLUSTERING UNTUK PENGELOMPOKAN WILAYAH BANJIR," *J. Ilm. Betrik*, vol. 12, No.03, no. 03, pp. 249–253, 2021.



- [12] I. Rosaliyah and B. Nurhakim, "CLUSTERING KEJADIAN BENCANA ALAM di JAWA BARAT BERDASARKAN JENIS BENCANA MENGGUNAKAN K-MEANS," *E-Link J. Tek. Elektro dan Inform.*, vol. 18, no. 1, p. 10, 2023, doi: 10.30587/e-link.v18i1.5318.
- [13] C. J. Siti Mariam, Fitri Handayani, "Penerapan Algoritma Clustering K-Means Untuk Menentukan Prioritas Penerima Bantuan Rumah Akibat Bencana Alam," *J. Tek. Inform. dan Sist. Inf.*, vol. 10, no. 2, pp. 231–240, 2023.
- [14] D. Kartika Widyawati and A. Fauzy, "Pengelompokan Daerah Rawan Bencana Tanah Longsor di Provinsi Daerah Istimewa Yogyakarta Menggunakan Metode K-Means Clustering," *Emerg. Stat. Data Sci. J.*, vol. 2, no. 2, pp. 212–221, 2024, doi: 10.20885/esds.vol2.iss.2.art14.
- [15] L. Tsiqoyati, R. Ferdiansyah, and M. Rizki, "Analisis Kerentanan Banjir di Kota Tangerang : Pendekatan Berbasis Data Geospasial untuk Pengambilan Keputusan dalam Perencanaan Tata Ruang Berkelanjutan," *Sainstech J. Penelit. Dan Pengkaj. Sains Dan Teknol.*, vol. 35, no. 2, p. 11, 2025, doi: <https://doi.org/10.37277/stch.v35i2.2356>.
- [16] F. X. Moruk, V. D. Boboy, W. J. Tahuk, Y. P. Kamirsa, and Y. R. Kaesmetan, "Penentuan Titik Lokasi Daerah Rawan Banjir Di Kabupaten Malaka Menggunakan Metode K-Means Clustering," *Simpatik J. Sist. Inf. dan Inform.*, vol. 3, no. 2, pp. 67–76, 2024, doi: 10.31294/simpatik.v3i2.2948.
- [17] E. D. N. Lestari, S. F. Rezky, and A. Alhafiz, "Penerapan Data Mining Untuk Pengelompokan Data Stok Barang Dengan Metode K-Means Clustering," *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, vol. 4, no. 3, pp. 528–539, 2025, doi: 10.53513/jursi.v4i3.8515.
- [18] H. Syahputra, "Clustering Tingkat Penjualan Menu (Food and Beverage) Menggunakan Algoritma K-Means," *J. KomtekInfo*, vol. 9, no. 1, pp. 29–33, 2022, doi: 10.35134/komtekinfo.v9i1.274.
- [19] P. Khairani Ritonga and M. S. Hasibuan, "Analisis Perbandingan Silhouette dengan Elbow pada Algoritma K-Means dan DBSCAN," *Metik J.*, vol. 9, no. 1, p. 2025, 2025, doi: 10.47002/metik.v9i1.1027.
- [20] M. Sholeh and K. Aeni, "Perbandingan Evaluasi Metode Davies Bouldin, Elbow dan Silhouette pada Model Clustering dengan Menggunakan Algoritma K-Means," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 8, no. 1, p. 56, 2023, doi: 10.30998/string.v8i1.16388.
- [21] S. Difa Pratama Hakim, "Segmentasi Pasar Pariwisata Provinsi Jawa Barat Menggunakan K-Means Berdasarkan Jumlah Kunjungan, Hotel, dan Restoran," *J. Inform. dan Teknologi Komput.*, vol. 5, no. 2, pp. 10–13, 2025, doi: <https://doi.org/10.55606/jitek.v5i2.6170>.
- [22] A. Asmana, Y. Arie Wijaya, and M. Martanto, "Clustering Data Calon Siswa Baru Menggunakan Metode K-Means Di Sekolah Menengah Kejuruan Wahidin Kota Cirebon," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, pp. 552–559, 2022, doi: 10.36040/jati.v6i2.5236.