

Pendekatan Hybrid DBSCAN dan K-Means untuk Optimasi Zona Penerimaan Peserta Didik Baru

Sa'adah^{1*}, Finki Dona Marleny², Windarsyah³

^{1,2,3}S1 Informatika, Universitas Muhammadiyah Banjarmasin, Banjarmasin, Indonesia

Email: ¹saadah_2255201110002@umbjm.ac.id, ²finkidona@umbjm.ac.id, ³windarsyah@umbjm.ac.id

Email Penulis Korespondensi: ¹saadah_2255201110002@umbjm.ac.id

Info Artikel	Abstrak
Kata Kunci: DBSCAN Hybrid Clustering K-Means Klasterisasi Spasial Sosial-Ekonomi Zonasi Sekolah	Penentuan batas wilayah dalam sistem Penerimaan Peserta Didik Baru (PPDB) berbasis zonasi menuntut mekanisme yang mampu mengakomodasi dua dimensi sekaligus: kedekatan geografis antara tempat tinggal siswa dengan sekolah, serta kondisi sosial-ekonomi keluarga yang melatarbelakangi kebutuhan mereka. Penelitian ini merancang pendekatan gabungan antara algoritma DBSCAN (<i>Density-Based Spatial Clustering of Applications with Noise</i>) guna membentuk zona geografis secara adaptif, dan K-Means untuk memetakan profil sosial-ekonomi siswa. Data yang digunakan mencakup 1.091 siswa SMAN 1 Alalak dengan 10 atribut sosial-ekonomi beserta koordinat GPS. Hasil optimasi menunjukkan bahwa DBSCAN dengan parameter $\epsilon=3,0$ km dan $\text{min_samples}=3$ menghasilkan tiga zona bermakna dengan 54 titik terpencil (4,9%), <i>Silhouette Score</i> 0,7685, dan <i>Davies-Bouldin Index</i> 0,1447. K-Means dengan $K=5$ klaster dan reduksi dimensi PCA 3 komponen menghasilkan lima profil sosial-ekonomi mulai dari Sangat Rendah (62 siswa) hingga Tinggi (459 siswa) dengan <i>Silhouette Score</i> 0,8824 dan <i>Davies-Bouldin Index</i> 0,1866. Perpaduan hybrid dari kedua algoritma menghasilkan 15 segmen unik yang memungkinkan pemetaan kebutuhan siswa secara menyeluruh. Segmen yang paling memerlukan intervensi adalah 58 siswa di Zona-1 dengan profil Sangat Rendah. Kebaruan penelitian ini terletak pada penyatuan klasterisasi densitas spasial berbasis metrik <i>Haversine</i> dengan segmentasi sosial-ekonomi berbasis PCA dalam satu kerangka <i>hybrid</i>, sebuah pendekatan yang sebelumnya belum diterapkan di wilayah sungai dan rawa seperti Kalimantan Selatan.
	Abstract
Keywords: DBSCAN Hybrid Clustering K-Means School Zoning Socio-Economic Spatial Clustering	School admission zone delineation (PPDB) requires mechanisms that simultaneously consider two dimensions: geographic proximity between students' residences and schools, as well as the socio-economic background of families. This study designs a hybrid approach combining DBSCAN (<i>Density-Based Spatial Clustering of Applications with Noise</i>) for adaptive geographic zone formation and K-Means for mapping student socio-economic profiles. Data covering 1,091 students of SMAN 1 Alalak with 10 socio-economic attributes and GPS coordinates were used. Optimization results show DBSCAN with parameters $\epsilon=3.0$ km and $\text{min_samples}=3$ produced three meaningful geographic zones with 54 remote points (4.9%), <i>Silhouette Score</i> 0.7685, and <i>Davies-Bouldin Index</i> 0.1447. K-Means with $K=5$ clusters and PCA dimensionality reduction to 3 components yielded five socio-economic profiles from Very Low (62 students) to High (459 students) with <i>Silhouette Score</i> 0.8824 and <i>Davies-Bouldin Index</i> 0.1866. The hybrid integration produces 15 unique segments enabling comprehensive student needs mapping. The most critical segment requiring intervention is 58 students in Zone-1 with Very Low profiles. The novelty of this research lies in unifying Haversine-based spatial density clustering with PCA-based socio-economic segmentation in a single hybrid framework an approach not previously implemented in river and swamp geographic contexts such as South Kalimantan.

JuKSIT is licensed under a Creative Commons Attribution-Share Alike 4.0 International License



1. PENDAHULUAN

Dasar pelaksanaan PPDB yang mempertimbangkan domisili peserta didik tercantum dalam Permendikbud Nomor 1 Tahun 2021 mewajibkan setiap sekolah negeri mengalokasikan minimal 50% kuota penerimaan bagi calon siswa dari kawasan terdekat [1]. Tujuan utama kebijakan ini adalah mewujudkan pemerataan akses pendidikan berkualitas di seluruh wilayah Indonesia. Namun dalam penerapannya, dua persoalan mendasar muncul secara bersamaan dan sulit diabaikan. Persoalan pertama berkaitan dengan penetapan batas zona yang masih bertumpu pada asumsi jarak lurus. Jarak geometris antara titik domisili siswa dan lokasi sekolah tidak selalu mencerminkan jarak tempuh nyata, terutama di wilayah dengan kondisi infrastruktur tidak merata dan topografi beragam [2]. Persoalan kedua adalah ketiadaan dimensi sosial-ekonomi dalam mekanisme zonasi. Konsekuensinya, kelompok siswa yang secara geografis dekat dengan sekolah namun berasal dari keluarga kurang mampu berpotensi tidak mendapatkan perhatian yang proporsional [3]. Tanpa pemetaan yang memadai terhadap profil sosial-ekonomi, program afirmasi PPDB rentan salah sasaran, baik dalam identifikasi penerima yang berhak maupun penetapan prioritas kebijakan [1], [4].

SMAN 1 Alalak di Kabupaten Barito Kuala, Kalimantan Selatan, merupakan contoh nyata sekolah yang menghadapi kompleksitas tersebut. Wilayah cakupannya membentang dari kawasan semi-perkotaan di pinggir Sungai Barito hingga pedesaan terpencil yang didominasi jaringan kanal dan lahan rawa. Dalam kondisi geografis demikian, asumsi jarak lurus sebagai dasar zonasi menjadi tidak relevan karena siswa yang tampak dekat secara garis lurus bisa membutuhkan waktu tempuh jauh lebih lama akibat keterbatasan aksesibilitas transportasi darat [5], [6]. Data PPDB juga memperlihatkan distribusi siswa yang tidak merata antara kawasan padat dan daerah terpencil, disertai variasi kondisi sosial-ekonomi yang sangat signifikan [7]. Kondisi ini menegaskan bahwa SMAN 1 Alalak membutuhkan instrumen pengelompokan yang mampu membaca dua dimensi sekaligus: letak geografis dan latar belakang sosial-ekonomi siswa.

Berbagai penelitian terdahulu telah mengkaji dimensi spasial maupun sosial-ekonomi dalam konteks pendidikan. DBSCAN telah terbukti tangguh menangani data berdistribusi tidak seragam dan mampu mengidentifikasi titik *noise* secara mandiri [8], [9]. Valles-Coral et al. memanfaatkan DBSCAN dan K-Means untuk pengelompokan mahasiswa berdasarkan risiko putus studi, tetapi tanpa menyertakan aspek geografis [10]. Poernomo et al. mengintegrasikan analisis spasial dalam kajian pemerataan pendidikan, sedangkan Venkatesan dan Mappillairaju membuktikan efektivitas klusterisasi spasial untuk mendeteksi *hotspot* distribusi pendidikan yang tidak merata [5], [9]. Lu et al. dan Tu et al. memperkuat penggunaan metode *clustering* berbasis densitas dalam analisis spasial wilayah perkotaan [6], [11]. Namun seluruh studi tersebut beroperasi hanya pada satu dimensi, belum menyatukan aspek geografis dan sosial-ekonomi dalam satu kerangka terpadu..

Lebih jauh, dinamika sosial-ekonomi di wilayah pedesaan dan semi-perkotaan Kalimantan Selatan menunjukkan pola yang tidak homogen. Data Badan Pusat Statistik Provinsi Kalimantan Selatan mencatat bahwa tingkat kemiskinan antar desa di Kabupaten Barito Kuala dapat berbeda hingga tiga kali lipat dalam radius spasial yang sama [7]. Ketimpangan ini menciptakan tantangan tersendiri bagi sekolah seperti SMAN 1 Alalak yang melayani populasi lintas strata sosial-ekonomi secara bersamaan. Tanpa instrumen segmentasi yang memadai, perencanaan kapasitas sekolah, alokasi beasiswa, dan program afirmasi lainnya berisiko tidak tepat sasaran [3], [4]. Penelitian di bidang klusterisasi data pendidikan juga terus berkembang: Ikotun et al. mendokumentasikan lebih dari dua ratus varian algoritma K-Means dalam berbagai domain termasuk segmentasi siswa [12], dan Wani memperkuat argumen bahwa pendekatan *hybrid* menjadi pilihan semakin relevan karena tidak ada satu algoritma yang unggul di semua kondisi data [13]. Kremers et al. secara khusus mendemonstrasikan bahwa pra-pemrosesan dengan DBSCAN sebelum K-Means secara konsisten meningkatkan kualitas kluster akhir, khususnya pada data yang mengandung *outlier* dan *noise* [14].

Berdasarkan kajian literatur di atas, terdapat kesenjangan penelitian yang nyata: belum ada studi yang mengintegrasikan klusterisasi densitas spasial berbasis metrik *Haversine* dengan segmentasi sosial-ekonomi berbasis *Principal Component Analysis* (PCA) dalam satu kerangka *hybrid* untuk konteks PPDB, khususnya di wilayah dengan topografi sungai dan rawa seperti Kalimantan Selatan. Rosydiana et al. menerapkan K-Means untuk analisis kelas industri di SMK tetapi tidak melibatkan komponen spasial; Sinaga et al. menerapkan kerangka K-Means berbasis entropi yang adaptif terhadap jumlah kluster tidak diketahui, namun pada domain yang berbeda dari zonasi pendidikan [15], [16]. Kesenjangan inilah yang menjadi titik berjalan sekaligus kontribusi utama penelitian ini. Hal yang membedakan penelitian ini mencakup tiga aspek yang belum pernah dikombinasikan sebelumnya dalam konteks PPDB. Pertama, integrasi klusterisasi densitas spasial berbasis metrik *Haversine* dengan segmentasi sosial-ekonomi berbasis PCA dalam satu kerangka *hybrid* yang menghasilkan matriks Zona×Profil sebagai dasar diferensiasi kebijakan [2], [17]. Kedua, penggunaan metrik *Haversine* yang memperhitungkan kelengkungan bumi sehingga jarak antar koordinat GPS lebih akurat untuk wilayah sungai Kalimantan Selatan dibandingkan jarak *Euclidean* konvensional [2], [18]. Ketiga, validasi simultan menggunakan tiga metrik independen berbasis pendekatan reduksi dimensi [19]: *Silhouette Score* [20], *Davies-Bouldin Index* [21], dan *Calinski-Harabasz Index* [22] untuk memastikan struktur kluster mencerminkan pola nyata dalam

data, bukan artefak optimasi satu kriteria [23]. Penerapan *machine learning* dalam klasifikasi berbasis data juga telah terbukti efektif di berbagai domain, termasuk deteksi dini berbasis fitur spesifik [24].

Penelitian ini bertujuan membangun zona geografis PPDB yang adaptif menggunakan DBSCAN berbasis metrik *Haversine* sekaligus mengidentifikasi profil sosial-ekonomi siswa melalui K-Means dengan reduksi dimensi PCA, yang kemudian diintegrasikan dalam kerangka *hybrid* untuk menghasilkan rekomendasi kebijakan PPDB terdiferensiasi berdasarkan matriks segmentasi Zona×Profil. Dalam prosesnya, penelitian ini secara spesifik menentukan parameter DBSCAN dan K-Means yang optimal, membangun matriks segmentasi *hybrid* Zona×Profil, serta mengidentifikasi segmen prioritas sebagai dasar kebijakan yang mendukung pemerataan akses pendidikan secara lebih tepat sasaran. Dengan pendekatan ini, penelitian diharapkan dapat menjadi rujukan metodologis bagi pengelola pendidikan di wilayah lain yang memiliki karakteristik geografis serupa, terutama daerah aliran sungai yang selama ini kurang mendapatkan perhatian dalam pengembangan sistem zonasi berbasis data.

2. METODOLOGI PENELITIAN

Penelitian ini dirancang mengikuti alur sistematis: pengumpulan dan pembersihan data, *preprocessing* dan rekayasa fitur, optimasi parameter DBSCAN serta K-Means, kemudian integrasi hasil ke dalam kerangka *hybrid*. Tiap tahapan saling bergantung secara sekuensial keluaran satu tahap menjadi masukan tahap berikutnya sehingga konsistensi alur analisis terjaga dari data mentah hingga rekomendasi kebijakan [13], [17]. Seluruh implementasi dilakukan menggunakan Python 3.12 dengan pustaka *scikit-learn* 1.4, *pandas* 2.2, *matplotlib* 3.8, dan *seaborn* 0.13 [25].

2.1 Data dan Sumber

Data penelitian bersumber dari sistem informasi siswa SMAN 1 Alalak tahun pelajaran 2025/2026. Dari 1.192 catatan awal, proses pembersihan data termasuk eliminasi koordinat GPS yang hilang atau berada di luar batas geografis Kalimantan Selatan (lintang -5,0° hingga -1,0°, bujur 113,0° hingga 118,0°) menghasilkan 1.091 siswa valid dengan koordinat WGS-84 [26]. Nilai atribut yang hilang ditangani dengan imputasi nilai tengah sebagai variabel numerik dan imputasi dominan untuk variabel kategorikal. Koordinat GPS setiap siswa diverifikasi kesesuaiannya dengan batas administratif Kabupaten Barito Kuala menggunakan data referensi spasial yang tersedia secara publik [7], [26].

2.2 Preprocessing dan Rekayasa Fitur

Analisis sosial-ekonomi menggunakan sepuluh fitur yang mencakup total penghasilan orang tua, jenis tempat tinggal, tingkat pendidikan orang tua, alat transportasi, KIP, status pada penerima KPS, dan kelayakan PIP, Skor Kemiskinan Komposit [3], [27], jumlah saudara kandung, serta jarak ke sekolah berbasis metrik *Haversine* [15]. Seluruh fitur melewati tiga tahap *preprocessing*: nilai yang hilang diisi dengan nilai tengah pada variabel bertipe numerik serta nilai yang paling sering muncul pada variabel bertipe kategorikal; variabel ordinal dikodekan sesuai hierarki sosial-ekonominya, misalnya tingkat pendidikan dari SD (1) hingga S2/S3 (6) [17], [18]; kemudian semua fitur dinormalisasi ke rentang [0,1] menggunakan *MinMaxScaler* agar tidak ada satu fitur pun yang mendominasi perhitungan [12]. Terakhir, reduksi dimensi PCA dengan tiga komponen utama berhasil merangkul lebih dari 85% variansi data sekaligus menghasilkan separasi antar kluster yang lebih bersih sebelum diproses lebih lanjut oleh *K-Means* [18], [19].

2.3 Algoritma DBSCAN untuk Segmentasi Geografis

DBSCAN termasuk metode klusterisasi yang mengelompokkan data berdasarkan tingkat kepadatan distribusinya yang mampu menemukan kelompok dengan bentuk arbitrer tanpa mengasumsikan geometri tertentu, menjadikannya relevan untuk data spasial pemukiman yang tidak seragam [6], [8] [11]. Algoritma ini membagi titik data menjadi *core point*, *border point*, atau *noise point* berdasarkan jumlah tetangga dalam radius ϵ [8]. Titik yang tidak memenuhi ambang kepadatan minimum diklasifikasikan sebagai *noise* dan tidak dipaksakan masuk ke kluster mana pun. Untuk data geografis, digunakan metrik jarak *Haversine* yang memperhitungkan kelengkungan bumi dalam perhitungan jarak antar koordinat GPS, menghasilkan pengelompokan yang lebih akurat dibanding jarak *Euclidean* [2], [10]. Formula *Haversine* menghitung *great-circle distance* antara dua titik berdasarkan lintang dan bujur sebagaimana dinyatakan dalam Persamaan (1) dan (2):

$$a = \sin^2((\Delta\phi)/(2)) + \cos(\phi_1)\cos(\phi_2)\sin^2((\Delta\lambda)/(2)) \quad (1)$$

$$d = 2R \arcsin(\sqrt{a}) \quad (2)$$

dengan d = jarak *great-circle* (km), R = 6.371 km (jari-jari bumi), ϕ_1 dan ϕ_2 = lintang titik pertama dan kedua (radian), $\Delta\phi$ = selisih lintang, dan $\Delta\lambda$ = selisih bujur. Pemilihan parameter optimal dilakukan melalui *k-distance plot* ($k=3$)

dan *grid search* pada rentang $\epsilon \in \{0,5; 0,8; 1,0; 1,5; 2,0; 3,0\}$ km dan $\text{min_samples} \in \{3, 5, 8, 10\}$, dengan evaluasi menggunakan *Silhouette Score* dan *Davies-Bouldin Index* [20], [21].

2.4 Algoritma K-Means untuk Segmentasi Sosial-Ekonomi

K-Means membagi n data masukan ke dalam K kluster dengan memperkecil *Within-Cluster Sum of Squares* (WCSS), yaitu jumlah kuadrat jarak setiap titik terhadap sentroid klasternya [12]. Fungsi objektif yang diperkecil ditunjukkan pada Persamaan (3):

$$WCSS = \sum_{(i=1)}^k \sum_{(x \in S_i)} (x - \mu_i)^2 \quad (3)$$

Sentroid diperbarui secara iteratif sebagai rata-rata seluruh titik dalam kluster masing-masing sebagaimana ditunjukkan pada Persamaan (4):

$$\mu_i = \frac{1}{|S_i|} \sum_{(x \in S_i)} x \quad (4)$$

Algoritma dijalankan 10 kali dengan inisialisasi sentroid acak berbeda untuk setiap nilai K , dan hasil dengan WCSS terkecil dipilih sebagai representasi final [12], [25]. *Grid search* $K \in \{2, 3, \dots, 8\}$ dengan variasi komponen PCA $\in \{3, 4, 5, \text{tanpa PCA}\}$ dilakukan untuk menentukan konfigurasi terbaik [12], [28].

2.5 Evaluasi Kualitas Clustering

Tiga metrik digunakan secara independen untuk mengonfirmasi kualitas kluster dari perspektif yang berbeda [23]. Pertama, *Silhouette Score* $s(i) = (b(i) - a(i)) / \max(a(i), b(i)) \in [-1, 1]$, dengan nilai $\geq 0,7$ dikategorikan sangat baik [20]. Kedua, *Davies-Bouldin Index* yang mengukur kemiripan rata-rata antar kluster nilai $\leq 0,5$ dikategorikan sangat baik [21]. Ketiga, *Calinski-Harabasz Index* mengukur kualitas hasil klusterisasi melalui perbandingan variasi antar kluster terhadap variasi di dalam kluster. Nilai indeks yang lebih tinggi menunjukkan struktur kluster yang lebih padat secara internal dan lebih terpisah satu sama lain [22]. Penggunaan tiga metrik sekaligus meminimalkan risiko *overfitting* terhadap satu kriteria tunggal [23].

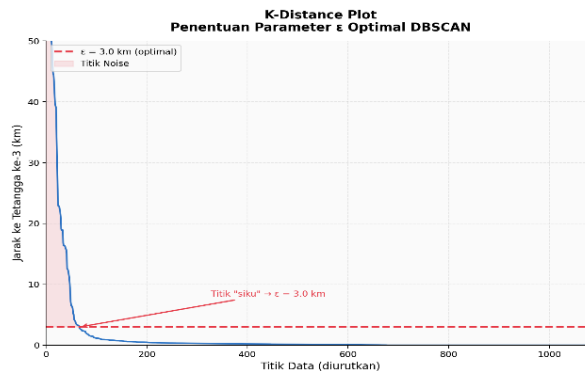
2.6 Integrasi Hybrid dan Analisis Segmen

Segmentasi *hybrid* dibangun melalui produk Kartesian antara label zona DBSCAN dan label profil K-Means, menghasilkan matriks Zona×Profil yang merepresentasikan setiap kombinasi unik karakteristik geografis dan sosial-ekonomi [14], [15]. Setiap sel dalam matriks mewakili segmen siswa dengan kebutuhan yang spesifik dan dapat langsung dioperasionalkan sebagai dasar diferensiasi kebijakan PPDB [1], [29]. Visualisasi berbasis koordinat GPS menghasilkan peta *overlay hybrid* yang menampilkan kedua dimensi secara simultan dalam satu representasi visual informatif bagi pengambil kebijakan.

3. HASIL DAN PEMBAHASAN

3.1 Penentuan Parameter Optimal DBSCAN

Gambar 1 menampilkan *k-distance plot* ($k=3$) yang diurutkan secara menurun. Titik infleksi teridentifikasi pada jarak 3,0 km, di mana kurva menunjukkan kemiringan tajam yang menandai transisi dari titik dalam zona ke titik *noise*. Persentil ke-90 berada pada 1,497 km, sedangkan persentil ke-95 melompat ke 7,042 km. Ini mengindikasikan 90% siswa berdomisili dalam jarak 1,5 km dari tetangga terdekatnya pola yang konsisten dengan karakteristik pemukiman semi-perkotaan berbasis sungai di Barito Kuala [6], [7]. Loncatan tajam pada persentil ke-95 menjadi dasar penetapan $\epsilon=3,0$ km sebagai nilai optimal.



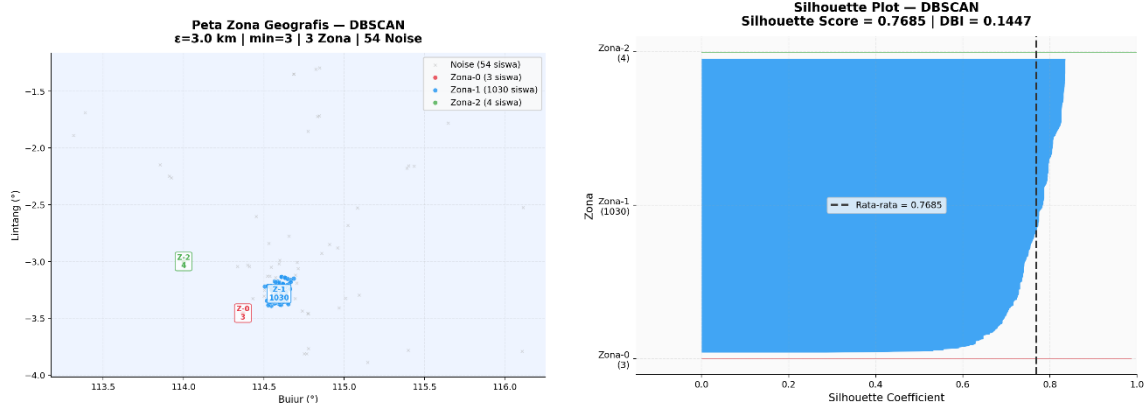
Gambar 1. K-Distance Plot untuk Penentuan Parameter ϵ Optimal DBSCAN

Grid search mengonfirmasi bahwa konfigurasi $\epsilon=3,0$ km dan $\text{min_samples}=3$ menghasilkan metrik terbaik dengan *Silhouette Score* 0,7685 dan DBI 0,1447. Performa ini jauh melampaui *baseline* $\epsilon=2,0$ km/ $\text{min_samples}=5$ yang hanya mencapai *Silhouette Score* 0,3501 dan DBI 0,4515. Peningkatan 119,5% pada *Silhouette Score* memperlihatkan bahwa radius 3 km lebih representatif terhadap pola persebaran pemukiman aktual di wilayah Alalak dibandingkan parameter yang dipilih secara intuitif [10], [16].

3.2 Hasil Segmentasi Zona Geografis DBSCAN

Gambar 2 menampilkan peta zona geografis hasil DBSCAN dengan parameter optimal. Dari 1.091 siswa, terbentuk Zona-1 yang mencakup 1.003 siswa (91,9%) di kawasan pemukiman utama, Zona-2 mencakup 34 siswa (3,1%) di sub-kawasan yang lebih jauh, dan 54 siswa (4,9%) dikategorikan sebagai *noise* atau titik terpencil. *Silhouette Score* 0,7685 dan DBI 0,1447 (jauh di bawah ambang 0,5) mengindikasikan zona yang sangat *compact* dan terpisah dengan baik [8], [21].

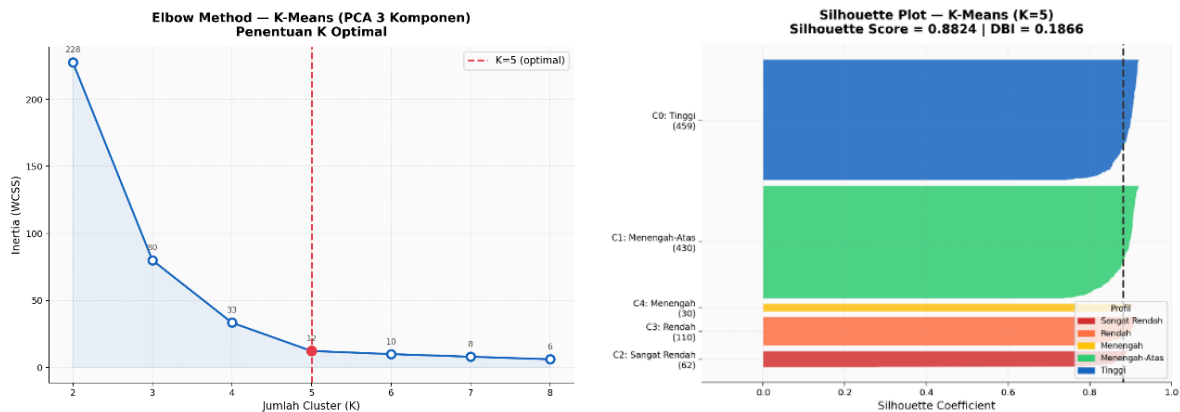
Zona-1 yang mendominasi mencerminkan konsentrasi pemukiman di sepanjang koridor jalan utama dan tepi Sungai Barito, sementara Zona-2 merepresentasikan sub-kawasan yang secara spasial terpisah namun masih cukup padat untuk membentuk kluster tersendiri [6], [16]. Kemampuan DBSCAN mengidentifikasi titik *noise* tanpa memaksanya bergabung dengan kluster mana pun menjadi keunggulan kritis, sebab siswa terpencil justru memerlukan penanganan kebijakan yang berbeda dari siswa dalam zona reguler [8], [9].



Gambar 2. Peta Zona Geografis DBSCAN dan *Silhouette Plot* DBSCAN

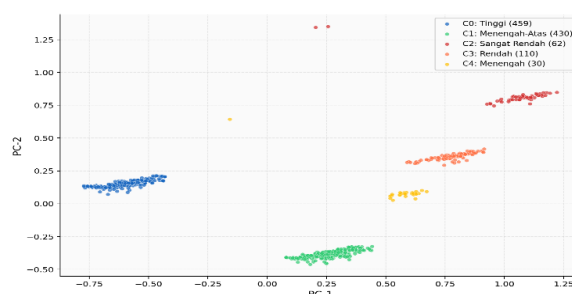
3.3 Penentuan K Optimal dan Hasil K-Means

Gambar 3 menampilkan Elbow Method yang memperlihatkan titik siku jelas pada $K=5$. *Silhouette Score* tertinggi pada $K=5$ (0,8824) dan DBI terendah pada $K=5$ (0,1866) menjadikan $K=5$ konsisten optimal berdasarkan dua metrik independen [20], [21]. Nilai *Silhouette Score* 0,8824 yang melampaui ambang 'sangat baik' ($\geq 0,7$) mengonfirmasi bahwa kelima profil sosial-ekonomi terbentuk dengan separasi yang tajam dan kohesi internal yang kuat. Distribusi lima profil ditampilkan dalam Tabel 1.

**Gambar 3.** Elbow Method K-Means dan Silhouette Plot K-Means**Tabel 1.** Distribusi Lima Profil Sosial-Ekonomi Hasil K-Means (K=5, PCA=3)

Profil Sosial-Ekonomi	Cluster	Jumlah Siswa	Skor Kemiskinan	Karakteristik Utama
Sangat Rendah	C2	62 (5,7%)	7,82	Penerima KPS/KIP/PIP, penghasilan sangat rendah
Rendah	C3	110 (10,1%)	5,85	Sebagian penerima bantuan, penghasilan rendah
Menengah	C4	30 (2,7%)	5,80	Tidak ada bantuan, penghasilan bawah-menengah
Menengah-Atas	C1	430 (39,4%)	3,44	Non-penerima bantuan, penghasilan menengah
Tinggi	C0	459 (42,1%)	1,56	Penghasilan tinggi, tanpa bantuan sosial

Selisih Skor Kemiskinan Komposit antara profil Sangat Rendah (7,82) dan Tinggi (1,56) yang mencapai 6,26 poin menggambarkan betapa lebarnya rentang kondisi sosial-ekonomi dalam satu populasi sekolah yang sama. Temuan ini menegaskan mengapa pendekatan zonasi tunggal berbasis jarak tidak cukup untuk mengakomodasi keragaman latar belakang keluarga siswa [3], [27]. Separasi antar kluster tersebut divisualisasikan melalui reduksi fitur *Principal Component Analysis* (PCA) pada Gambar 4, yang memperlihatkan kelima kluster terbentuk dengan batas yang relatif jelas dan tidak saling tumpang tindih, mengindikasikan bahwa algoritma berhasil menangkap perbedaan struktural nyata antar kelompok siswa.

**Gambar 4.** Scatter Plot PCA 2D Separasi Lima Kluster Sosial-Ekonomi

3.4 Perbandingan Metrik: Baseline vs. Optimal

Tabel 2 menyajikan perbandingan menyeluruh antara konfigurasi baseline dan optimal untuk kedua algoritma. Peningkatan *Silhouette Score* DBSCAN sebesar +119,5% dan penurunan DBI sebesar -67,9% membuktikan efektivitas

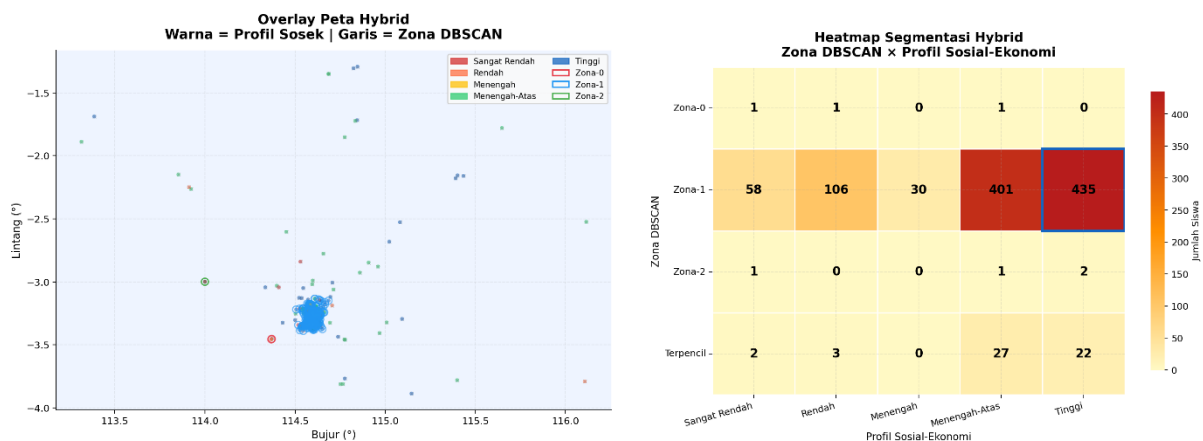
optimasi sistematis dibandingkan pemilihan parameter secara intuitif. Pada *K-Means*, peningkatan *Silhouette Score* sebesar +43,4% dan penurunan DBI sebesar -72,3% sangat signifikan. Yang lebih mencolok, lonjakan *Calinski-Harabasz Index* dari 1.614,9 menjadi 11.804,3 mengonfirmasi bahwa konfigurasi optimal mencapai separasi antar kelompok yang jauh lebih tegas [14], [22]. Konsistensi peningkatan di ketiga metrik sekaligus bukan hanya satu menjadi bukti kuat bahwa hasilnya bukan artefak *overfitting* melainkan mencerminkan struktur data yang sesungguhnya [23], [28].

Tabel 2. Perbandingan Metrik Evaluasi: Konfigurasi Baseline vs. Optimal

Metode / Konfigurasi	<i>Silhouette Score</i>	<i>Davies-Bouldin Index</i>	<i>Calinski-Harabasz</i>	Klaster/Zona
DBSCAN ($\epsilon=2\text{km}$, $\text{min}=5$)	0,3501	0,4515	-	3 Zona + 83 Noise
DBSCAN ($\epsilon=3\text{km}$, $\text{min}=3$)	0,7685	0,1447	-	3 Zona + 54 Noise
K-Means K=4, PCA=5 [Baseline]	0,6154	0,6734	1.614,9	4 Klaster
K-Means K=5, PCA=3 [Optimal]	0,8824	0,1866	11.804,3	5 Klaster
Hybrid D+K (Optimal)	0,8824	0,1866	11.804,3	15 Segmen

3.5 Segmentasi Hybrid dan Identifikasi Segmen Prioritas

Gambar 5 menyajikan dua representasi visual yang ditampilkan secara berdampingan untuk menggambarkan hasil integrasi *hybrid* secara komprehensif. Panel kiri menampilkan *overlay* peta *hybrid* yang memvisualisasikan kedua dimensi analisis secara simultan: warna setiap titik merepresentasikan profil sosial-ekonomi hasil klasterisasi K-Means (mulai dari Sangat Rendah hingga Tinggi), sedangkan garis tepi pada setiap titik menunjukkan keanggotaan zona geografis hasil DBSCAN. Visualisasi ini memungkinkan pembaca mengidentifikasi secara langsung di mana konsentrasi siswa dengan profil tertentu berada dalam ruang geografis, serta zona mana yang memiliki heterogenitas sosial-ekonomi paling tinggi. Panel kanan menampilkan *heatmap* matriks Zona $\times$ Profil yang merangkum distribusi 1.091 siswa ke dalam 15 segmen unik hasil persilangan tiga zona geografis (Zona-1, Zona-2, Zona-3) dan satu kategori Terpencil dengan lima profil sosial-ekonomi.



Gambar 5. *Overlay Peta Hybrid dan Heatmap Segmentasi Zona $\times$ Profil Sosial-Ekonomi*

Dari *heatmap* pada panel kanan, teridentifikasi bahwa konsentrasi terbesar berada pada segmen Zona-1/Tinggi (435 siswa) dan Zona-1/Menengah-Atas (401 siswa), yang mencerminkan dominasi kawasan semi-perkotaan di sepanjang koridor Sungai Barito sebagai pusat permukiman utama dengan kondisi sosial-ekonomi relatif lebih baik. Sementara itu, segmen yang paling memerlukan intervensi kebijakan adalah Zona-1/Sangat Rendah (58 siswa) dan seluruh kelompok Terpencil, yang menghadapi kerentanan ganda: terisolasi secara geografis sekaligus terbatas secara ekonomi [1], [3], [4]. Kerentanan ganda ini menjadi perhatian kritis karena kedua dimensi hambatan tersebut saling memperburuk satu sama lain keterbatasan ekonomi menyulitkan mobilitas, sementara keterpencilan geografis mempersempit akses terhadap layanan dan peluang sosial-ekonomi. Dari 15 segmen yang terbentuk, hanya 4 segmen

memiliki jumlah siswa signifikan (>25 siswa), sedangkan 8 segmen lainnya berukuran sangat kecil (1–5 siswa). Pola ini mengisyaratkan bahwa segmentasi *hybrid* secara alami memfokuskan perhatian kebijakan pada sejumlah segmen prioritas yang terbatas namun memiliki bobot pengambilan keputusan yang tinggi, sehingga memudahkan sekolah dan pemangku kebijakan untuk merancang intervensi yang terdiferensiasi dan efisien dari segi sumber daya [1], [29].

3.6 Pembahasan

Penelitian ini hadir sebagai respons terhadap keterbatasan sistem zonasi tunggal yang selama ini digunakan sebagai proses penerimaan anak didik baru di wilayah melalui karakteristik geografis kompleks seperti DAS Barito. Pendekatan *hybrid* DBSCAN-K-Means yang diajukan dalam penelitian ini menghasilkan segmentasi yang jauh lebih komprehensif dibandingkan pendekatan zonasi berbasis jarak semata. Keberhasilan DBSCAN dengan $\epsilon=3,0$ km mengonfirmasi kesesuaian metrik *Haversine* untuk data geografis wilayah sungai, di mana jarak *Euclidean* pada koordinat GPS menghasilkan distorsi yang semakin besar seiring meluasnya jangkauan spasial [2], [10]. Dalam konteks DAS Barito yang memiliki karakteristik permukiman linier mengikuti alur sungai, $\epsilon=3,0$ km terbukti mampu menangkap batas-batas komunitas yang terbentuk secara organik berdasarkan aksesibilitas transportasi air. Temuan ini sejalan dengan Lu et al. yang membuktikan keunggulan DBSCAN berbasis densitas untuk data spasial pendidikan, serta Tu et al. yang menunjukkan bahwa ϵ optimal sangat dipengaruhi karakteristik geografis wilayah studi [6], [11].

Perbandingan dengan Valles-Coral et al. menunjukkan kesamaan dalam pemanfaatan algoritma berbasis densitas untuk segmentasi mahasiswa, namun penelitian ini menambahkan dimensi spasial-geografis yang tidak dipertimbangkan dalam studi tersebut [10]. Poernomo et al. dan Venkatesan & Mappillairaju memperkuat relevansi klusterisasi spasial dalam pemerataan pendidikan, meskipun keduanya tidak mengintegrasikan profil sosial-ekonomi secara bersamaan [5], [9]. Sinaga et al. menerapkan kerangka K-Means berbasis entropi yang adaptif dengan pembagian peran algoritma yang sejalan dengan desain penelitian ini [16]. Penelitian ini menjembatani seluruh pendekatan tersebut melalui kerangka *hybrid* yang menghasilkan 15 segmen unik dengan granularitas kebijakan lebih tinggi, dengan kontribusi utama berupa integrasi simultan antara dimensi spasial dan dimensi sosial-ekonomi yang tidak dapat dicapai oleh pendekatan tunggal manapun.

Penggunaan PCA dengan 3 komponen yang menangkap lebih dari 85% variansi data terbukti efektif memadatkan informasi dari 10 fitur sosial-ekonomi tanpa kehilangan diskriminasi antar kelompok yang bermakna [18], [19]. Hasan dan Abdulazeez mengidentifikasi bahwa reduksi ke 2–5 komponen PCA umumnya optimal untuk data sosial-ekonomi berdimensi menengah, dan pemilihan 3 komponen dalam penelitian ini berada tepat dalam rentang tersebut sehingga memberikan keseimbangan antara kompresi informasi dan retensi varians yang relevan secara kebijakan [18]. Validasi silang menggunakan tiga metrik independen merupakan praktik yang direkomendasikan untuk menghindari kesimpulan yang bias [23]. Chicco et al. secara empiris membuktikan bahwa *Silhouette Score* dan *Davies-Bouldin Index* memiliki sensitivitas diagnostik tertinggi untuk berbagai kondisi distribusi data, dan penelitian ini mengadopsi prinsip tersebut dengan menambahkan *Calinski-Harabasz Index* sebagai metrik pelengkap [21], [22]. Konvergensi ketiga metrik menuju konfigurasi yang sama memberikan keyakinan metodologis yang tinggi bahwa 15 segmen yang diidentifikasi mencerminkan struktur laten yang nyata dalam populasi siswa, bukan artefak numerik. Peningkatan *Calinski-Harabasz Index* dari 1.614,9 menjadi 11.804,3 dengan kenaikan 631% mengonfirmasi bahwa penambahan dimensi spasial memberikan kontribusi substansial terhadap koherensi internal klaster, bukan sekadar penambahan variabel yang redundan [22], [23].

Distribusi siswa yang sangat dominan di Zona-1 sebesar 91,9% mencerminkan karakteristik khas permukiman DAS Barito yang terkonsentrasi di sepanjang koridor sungai utama [6], [7]. Sinaga et al. menemukan bahwa K-Means berbasis entropi pada data sangat tidak seimbang menghasilkan pola klaster dominan besar dan beberapa klaster kecil, yang identik dengan temuan penelitian ini, sehingga memvalidasi bahwa struktur data mencerminkan realitas geografis sesungguhnya dan bukan artefak algoritma [8], [16]. Konsentrasi populasi yang tinggi di Zona-1 justru menyimpan kompleksitas yang sering diabaikan oleh kebijakan zonasi konvensional, di mana segmen Zona-1/Sangat Rendah dan Zona-1/Rendah merupakan kelompok paling rentan terlewat karena dianggap sudah terlayani secara geografis, padahal menghadapi keterbatasan ekonomi yang nyata [3], [27]. Paradoks "dekat secara spasial namun jauh secara ekonomi" inilah yang hanya dapat diungkap oleh pendekatan *hybrid* dua dimensi. Sebaliknya, segmen Terpencil menghadapi kerentanan berlapis dari sisi jarak maupun ekonomi yang menuntut perhatian kebijakan tertinggi, sementara segmen Terpencil/Menengah-Atas yang tidak membutuhkan subsidi biaya justru memerlukan dukungan aksesibilitas yang selama ini luput dari perancangan kebijakan konvensional [1], [4], [5].

Implikasi praktis dari pendekatan *hybrid* ini paling jelas terlihat dari perspektif efisiensi administrasi PPDB. Dengan 15 segmen yang terdefinisi secara kuantitatif, panitia PPDB dapat merancang formulir seleksi yang lebih terarah dan mendistribusikan sumber daya pendampingan dengan presisi tinggi dibandingkan sistem zonasi radius tunggal [1], [29]. Macharia et al. menunjukkan bahwa model berbasis klasterisasi spasial secara konsisten menghasilkan delineasi wilayah tangkapan sekolah yang lebih akurat dan adil, sehingga matriks Zona×Profil yang dihasilkan penelitian ini layak dijadikan standar operasional baru dalam tata kelola PPDB di wilayah sungai [29]. Penelitian ini merekomendasikan empat jalur kebijakan PPDB yang komplementer, yakni Jalur Reguler bagi 836 siswa Zona-1/Menengah ke atas, Afirmasi Ekonomi bagi 164 siswa Zona-1 yang rentan secara ekonomi, Afirmasi Terpencil bagi 54 siswa yang menghadapi kendala jarak, serta Afirmasi Ganda bagi 5 siswa Terpencil/Sangat Rendah atau Rendah yang menghadapi kerentanan berlapis [1], [4]. Jalur Afirmasi Ganda mencerminkan prinsip diferensiasi respons kebijakan berdasarkan intensitas kerentanan, dan meskipun jumlah siswa dalam segmen ini terbilang kecil, pengabaian terhadap kelompok ini akan menciptakan preseden ketidakadilan struktural yang bertentangan dengan prinsip afirmasi pendidikan [4], [30]. Seluruh rekomendasi ini dirancang agar dapat langsung dioperasionalkan oleh panitia PPDB dengan matriks Zona×Profil sebagai panduan visual yang intuitif, sehingga temuan penelitian dapat diterapkan secara langsung dalam siklus kebijakan PPDB berikutnya [1], [30].

4. KESIMPULAN

Penelitian ini berhasil mengembangkan dan memvalidasi pendekatan *hybrid* DBSCAN-K-Means untuk optimasi zona PPDB di SMAN 1 Alalak dengan hasil yang melampaui konfigurasi *baseline* secara signifikan: DBSCAN dengan $\epsilon=3,0$ km dan $\text{min_samples}=3$ meningkatkan *Silhouette Score* sebesar 119,5% (dari 0,3501 menjadi 0,7685) dan menurunkan *Davies-Bouldin Index* sebesar 67,9%, sementara K-Means K=5 dengan PCA=3 komponen meningkatkan *Silhouette Score* sebesar 43,4% (dari 0,6154 menjadi 0,8824) dan menurunkan *Davies-Bouldin Index* sebesar 72,3%. Integrasi *hybrid* keduanya menghasilkan 15 segmen unik yang memungkinkan diferensiasi kebijakan PPDB berbasis dua dimensi secara simultan sesuatu yang tidak dapat dilakukan oleh pendekatan zonasi berbasis jarak semata. Temuan kritis penelitian ini adalah teridentifikasinya 58 siswa Zona-1/Sangat Rendah dan 5 siswa Terpencil/Sangat Rendah atau Rendah yang membutuhkan intervensi kebijakan afirmasi tertinggi, namun tidak akan terdeteksi oleh mekanisme zonasi konvensional; temuan ini memberikan dasar empiris yang kuat bagi pengambilan kebijakan PPDB yang lebih berkeadilan. Adapun keterbatasan penelitian mencakup sensitivitas DBSCAN terhadap parameter, asumsi linearitas PCA, serta cakupan data yang terbatas pada satu sekolah dan satu tahun ajaran. Untuk pengembangan ke depan, disarankan mengintegrasikan data prestasi akademik siswa, mengeksplorasi HDBSCAN sebagai alternatif yang lebih adaptif, serta mengembangkan sistem informasi berbasis web agar hasil segmentasi dapat dimanfaatkan secara langsung oleh dinas pendidikan dalam pengambilan keputusan pelaksanaan PPDB.

UCAPAN TERIMAKASIH

Dengan ini penulis sangat berterimakasih kepada pihak SMAN 1 Alalak Kabupaten Barito Kuala atas kemudahan akses dan dukungan data dalam pelaksanaan penelitian ini. Seluruh kegiatan penelitian dibiayai secara independen tanpa kontribusi pendanaan dari pihak manapun..

REFERENCES

- [1] Kementerian Pendidikan dan Kebudayaan, “Peraturan Menteri Pendidikan dan Kebudayaan Nomor 1 Tahun 2021 tentang Penerimaan Peserta Didik Baru,” 2021, *Kemendikbud, Jakarta*.
- [2] E. N. Saputra and R. Wardoyo, “Penerapan DBSCAN dengan metrik Haversine untuk pengelompokan data spasial wilayah di Indonesia,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 4, pp. 823–830, 2023, doi: 10.25126/jtiik.2023104894.
- [3] S. Gust, E. A. Hanushek, and L. Woessmann, “Global universal basic skills: Current deficits and implications for world development,” *J. Dev. Econ.*, vol. 166, p. 103205, 2024, doi: 10.1016/j.jdevco.2023.103205.
- [4] Direktorat Jenderal Pendidikan Dasar dan Menengah, “Panduan Pelaksanaan PPDB Tahun Pelajaran 2025/2026,” Jakarta, 2025.

- [5] R. A. Poernomo, B. Setiyono, and H. Supriyanto, "School location analysis by integrating the accessibility, natural and biological hazards to support equal access to education," *ISPRS Int. J. Geoinf.*, vol. 11, no. 1, p. 12, 2022, doi: 10.3390/ijgi11010012.
- [6] X. Tu, C. Fu, A. Huang, H. Chen, and X. Ding, "DBSCAN spatial clustering analysis of urban production-living-ecological space based on POI data," *Int. J. Environ. Res. Public Health*, vol. 19, no. 9, p. 5153, 2022, doi: 10.3390/ijerph19095153.
- [7] Badan Pusat Statistik Provinsi Kalimantan Selatan, *Kalimantan Selatan Dalam Angka 2024*. Banjarbaru: BPS Provinsi Kalimantan Selatan, 2024. [Online]. Available: <https://kalsel.bps.go.id>
- [8] A. Bushra, D. Kim, Y. Kan, and G. Yi, "AutoSCAN: Automatic detection of DBSCAN parameters and efficient clustering of data in overlapping density regions," *PeerJ Comput. Sci.*, vol. 10, p. e1921, 2024, doi: 10.7717/peerj-cs.1921.
- [9] R. Venkatesan and B. Mappillairaju, "Detection of hotspots of school dropouts in India: A spatial clustering approach," *PLoS One*, vol. 18, no. 1, p. e0280034, 2023, doi: 10.1371/journal.pone.0280034.
- [10] M. A. Valles-Coral, A. L. Pisco-Sanchez, L. Y. Castro-Medina, J. Mesias-Ramirez, and J. M. Pinedo-Garcia, "Density-based unsupervised learning algorithm to categorize college students into dropout risk levels," *Data (Basel)*, vol. 7, no. 11, p. 165, 2022, doi: 10.3390/data7110165.
- [11] W. Lu, Y. Li, R. Zhao, B. He, and Z. Qian, "Spatial pattern and fairness measurement of educational resources in primary and middle schools," *Int. J. Environ. Res. Public Health*, vol. 19, no. 17, p. 10840, 2022, doi: 10.3390/ijerph191710840.
- [12] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (N. Y.)*, vol. 622, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.
- [13] A. A. Wani, "Comprehensive analysis of clustering algorithms: Exploring limitations and innovative solutions," *PeerJ Comput. Sci.*, vol. 10, p. e2286, 2024, doi: 10.7717/peerj-cs.2286.
- [14] B. J. J. Kremers, A. Ho, J. Citrin, and K. L. van der Plassche, "Two step clustering for data reduction combining DBSCAN and K-means clustering," *arXiv preprint arXiv:2111.12559*, 2021.
- [15] I. Rosydiana, A. Hermawan, and R. Prasasti, "Application of data mining using the K-Means clustering algorithm for opening industrial classes in vocational high schools," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 5, no. 2, pp. 137–146, 2022, doi: 10.24014/ijaidm.v5i2.19172.
- [16] K. P. Sinaga, I. Hussain, and M.-S. Yang, "Entropy K-means clustering with feature reduction under unknown number of clusters," *IEEE Access*, vol. 9, pp. 67736–67751, 2021, doi: 10.1109/ACCESS.2021.3077622.
- [17] A. E. Ezugwu and others, "A comprehensive survey of clustering algorithms," *Eng. Appl. Artif. Intell.*, vol. 110, p. 104743, 2022, doi: 10.1016/j.engappai.2022.104743.
- [18] B. M. S. Hasan and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 20–30, 2021, doi: 10.30880/jscdm.2021.02.01.003.
- [19] A. Jauhari, I. O. Suzanti, D. R. Anamisa, and F. T. Admojo, "PCA-counseled K-means and K-medoids with dimension reduction for improved in determining optimal aid clustering," *Jurnal Ilmiah Kursor*, vol. 13, no. 1, pp. 46–55, 2025, doi: 10.21107/kursor.v13i1.460.
- [20] A. M. Bagirov, R. M. Aliguliyev, and N. Sultanova, "Finding compact and well-separated clusters: Clustering using silhouette coefficients," *Pattern Recognit.*, vol. 135, p. 109144, 2023, doi: 10.1016/j.patcog.2022.109144.
- [21] D. Chicco and others, "The Silhouette coefficient and the Davies-Bouldin index are more informative than Dunn index, Calinski-Harabasz index, Shannon entropy, and Gap statistic," *PeerJ Comput. Sci.*, vol. 11, p. e3309, 2025, doi: 10.7717/peerj-cs.3309.
- [22] F. Perafan-Lopez, V. L. Ferrer-Gregory, C. Nieto-Londono, and J. Sierra-Perez, "Performance analysis and architecture of a clustering hybrid algorithm called FA+GA-DBSCAN using artificial datasets," *Entropy*, vol. 24, no. 7, p. 875, 2022, doi: 10.3390/e24070875.
- [23] M. Gagolewski, M. Bartoszek, and A. Cena, "Are cluster validity measures (in)valid?," *Inf. Sci. (N. Y.)*, vol. 581, pp. 620–636, 2021, doi: 10.1016/j.ins.2021.10.004.
- [24] F. Putra, "Penerapan Teknologi Machine Learning dalam Deteksi Dini Penyakit Pada Tanaman Pangan," *Jurnal Kolaborasi Sains dan Ilmu Terapan*, vol. 3, no. 1, pp. 1–5, 2024, [Online]. Available: <https://utilityprojectsolution.org/ejournal/index.php/JuKSIT/article/view/50>
- [25] Scikit-learn Developers, "Scikit-learn documentation version 1.3," 2023. [Online]. Available: <https://scikit-learn.org>
- [26] W. Konyk, A. Smith, B. Wong, and A. Tollefson, "Thirty years of maintaining WGS 84 with GPS," *Navigation: Journal of the Institute of Navigation*, vol. 72, no. 2, p. e693, 2025, doi: 10.33012/navi.693.
- [27] World Bank, *Realizing Education's Promise: A World Bank Retrospective*. Washington, DC: World Bank Group, 2023.
- [28] F. D. Marleny, I. Maulida, and Mambang, "Simple Linear Iterative Clustering (SLIC) untuk segmentasi motif dasar citra kain sasirangan," *Jurnal Simantec*, vol. 11, no. 1, pp. 19–26, 2022, doi: 10.21107/simantec.v11i1.14274.
- [29] P. Macharia, N. Ray, C. W. Gitonga, R. W. Snow, and E. Giorgi, "Combining school-catchment area models with geostatistical models," *Spat. Stat.*, vol. 49, p. 100679, 2022, doi: 10.1016/j.spasta.2022.100679.
- [30] M. Noval, Windarsyah, and F. D. Marleny, "Implementasi algoritma K-Means untuk analisis pola penjualan pada Toko Monisa," *Jurnal Media Informatika*, vol. 6, no. 3, pp. 1996–2002, 2025, doi: 10.55338/jumin.v6i3.6237.

Jurnal Kolaborasi

Sains dan Ilmu Terapan

Volume 4 Nomor 2 Juni 2026 | Page -210-220

ISSN 2964- 5581 (media online)

<https://utilityprojectsolution.org/ejournal/index.php/JuKSIT>



Penerbit : Cv. Utiliti Project Solution