Perbandingan DBSCAN Teroptimasi Grid Search dan HDBSCAN untuk Pengelompokan Kemiskinan Berdasarkan Indikator Sosial Ekonomi di Jawa Barat

1st Muhammad Fadhlandhifan Siregar
Program Studi Sains Data
Universitas Telkom
Purwokerto, Indonesia
fadhlansiregar@student.telkomuniversity.ac.id

2ns Ridwan Pandiya
Program Studi Teknik Informatika
Universitas Telkom
Purwokerto, Indonesia
ridwanp@telkomuniversity.ac.id

3rf Aina Latifa Riyana Putri
Program Studi Sains Data
Universitas Telkom
Purwokerto, Indonesia
ainaqp@telkomuniversity.ac.id

***Abstrak*—Penelitian ini membandingkan Density-Based Spatial Clustering of Applications with Noise (DBSCAN) yang dioptimasi menggunakan Grid Search dan Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) untuk mengelompokkan 27 kabupaten/kota di Jawa Barat. Delapan indikator sosial ekonomi tahun 2023 digunakan, yaitu tingkat pengangguran terbuka, indeks pembangunan manusia, PDRB per kapita, Gini Ratio, akses sanitasi layak, akses air minum layak, rasio ketergantungan, dan rumah layak huni. Data diperiksa, ditangani nilai ekstremnya, dan distandardisasi sebelum pemodelan. DBSCAN awal dengan epsilon 34,972 dan MinPts 16 hanya menghasilkan satu klaster. Grid Search memperoleh parameter terbaik epsilon 16,5 dan MinPts 9, membentuk dua klaster yang mencakup 20 wilayah dan tujuh noise. Nilai Silhouette, Davies-Bouldin, dan Dunn masing-masing 0,6279, 0,4882, dan 0,3288. HDBSCAN membentuk dua klaster yang mencakup 25 wilayah dan dua noise dengan nilai 0,5299, 0,6054, dan 0,3703. DBSCAN lebih unggul dalam kekompakan dan pemisahan klaster, sedangkan HDBSCAN memberikan cakupan pengelompokan yang lebih luas.**

***Kata kunci*—kemiskinan, DBSCAN, HDBSCAN, grid search, klasterisasi, Jawa Barat**

I. PENDAHULUAN

Kemiskinan merupakan persoalan multidimensional yang tidak hanya tercermin dari rendahnya pendapatan, tetapi juga dari keterbatasan akses terhadap pekerjaan, pendidikan, layanan dasar, dan hunian yang layak. Perbedaan struktur ekonomi, kualitas sumber daya manusia, ketimpangan, serta ketersediaan infrastruktur dasar menyebabkan karakteristik kemiskinan antardaerah tidak seragam. Oleh sebab itu, analisis pada tingkat kabupaten/kota perlu mempertimbangkan sejumlah indikator secara bersamaan agar pola sosial ekonomi wilayah dapat dipahami secara lebih komprehensif [1], [2].

Penelitian ini menggunakan delapan indikator sosial ekonomi, yaitu Tingkat Pengangguran Terbuka (TPT), Indeks Pembangunan Manusia (IPM), Produk Domestik Regional Bruto (PDRB) per kapita, Gini Ratio, akses sanitasi layak, akses air minum layak, rasio ketergantungan, dan persentase rumah layak huni. TPT menggambarkan penyerapan tenaga kerja, IPM menunjukkan capaian kesehatan, pendidikan, dan standar hidup, sedangkan PDRB per kapita merepresentasikan kapasitas ekonomi rata-rata. Gini Ratio melengkapi indikator ekonomi dengan memperlihatkan tingkat ketimpangan. Empat indikator lainnya merepresentasikan pemenuhan layanan dasar dan beban demografis wilayah.

Pengelompokan wilayah berdasarkan kemiripan indikator dapat dilakukan menggunakan metode klasterisasi berbasis kepadatan. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) mampu membentuk klaster tanpa menentukan jumlah kelompok sejak awal dan dapat mengidentifikasi observasi yang tidak memenuhi struktur kepadatan sebagai noise [3]. Namun, hasil DBSCAN sangat sensitif terhadap parameter epsilon dan minimum points (MinPts). Pemilihan parameter yang tidak sesuai dapat menyatukan seluruh observasi ke dalam satu klaster atau menghasilkan noise secara berlebihan. Sejumlah penelitian menunjukkan bahwa optimasi parameter dapat meningkatkan kualitas pembentukan klaster DBSCAN [4]-[6].

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) dikembangkan untuk menangani variasi kepadatan melalui pembentukan hierarki klaster dan pemilihan kelompok yang stabil [7]. Metode ini tidak memerlukan epsilon secara eksplisit dan cenderung lebih adaptif terhadap kepadatan lokal. Meskipun demikian, keluasan cakupan pengelompokan tidak selalu sejalan dengan kekompakan serta pemisahan klaster. Perbandingan kedua metode perlu mempertimbangkan lebih dari satu metrik validasi internal dan proporsi wilayah yang berhasil dikelompokkan.

Penelitian terdahulu telah menerapkan DBSCAN pada data kesehatan, kriminalitas, status desa, transportasi, serta pola sosial ekonomi [3], [8]-[10]. Penelitian lain mengembangkan optimasi DBSCAN menggunakan pendekatan evolusioner, pencarian sistematis, atau perbaikan perhitungan ketetanggaan [4]-[6]. Namun, perbandingan DBSCAN yang dioptimasi melalui Grid Search dengan HDBSCAN pada data kemiskinan multidimensional tingkat kabupaten/kota di Jawa Barat masih terbatas. Penelitian ini bertujuan menemukan kombinasi epsilon dan MinPts yang optimal, membandingkan kualitas hasil DBSCAN dan HDBSCAN, serta menginterpretasikan karakteristik kelompok wilayah yang terbentuk.

II. KAJIAN TEORI

A. Klasterisasi Berbasis Kepadatan

DBSCAN mengelompokkan observasi berdasarkan kepadatan lokal menggunakan dua parameter. Epsilon menentukan radius ketetanggaan, sedangkan MinPts menentukan jumlah minimum observasi yang diperlukan agar suatu titik menjadi core point. Titik yang berada dalam jangkauan core point dapat menjadi border point, sementara titik yang tidak memenuhi hubungan kepadatan dikategorikan sebagai noise. Keunggulan utama DBSCAN adalah kemampuannya menangani bentuk klaster tidak beraturan dan mendeteksi observasi ekstrem [3], [10].

HDBSCAN memperluas prinsip DBSCAN melalui core distance dan mutual reachability distance. Jarak tersebut digunakan untuk membangun minimum spanning tree, membentuk hierarki kepadatan, dan memilih klaster berdasarkan kestabilannya [7]. Pendekatan hierarkis memungkinkan struktur dengan kepadatan berbeda tetap teridentifikasi tanpa menetapkan satu nilai epsilon untuk seluruh data.

B. Grid Search dan Standardisasi

Grid Search melakukan evaluasi sistematis terhadap seluruh kombinasi parameter pada ruang pencarian yang telah ditentukan. Dalam penelitian ini, setiap kombinasi epsilon dan MinPts digunakan untuk membentuk klaster DBSCAN, kemudian hasil yang memenuhi sekurang-kurangnya dua klaster dievaluasi untuk menentukan konfigurasi terbaik. Pendekatan ini bersifat transparan dan dapat direplikasi, meskipun kebutuhan komputasinya meningkat seiring bertambahnya kombinasi parameter [5], [11].

DBSCAN dan HDBSCAN menggunakan jarak antartitik sehingga perbedaan skala variabel dapat menimbulkan dominasi pada fitur dengan rentang besar, terutama PDRB per kapita. Standardisasi mengubah setiap variabel agar memiliki rata-rata nol dan simpangan baku satu. Dengan demikian, seluruh indikator memberikan kontribusi yang lebih seimbang dalam perhitungan jarak.

C. Evaluasi Klaster

Kualitas klaster dievaluasi menggunakan Silhouette Score, Davies-Bouldin Index (DBI), dan Dunn Index. Silhouette mengukur kedekatan observasi dengan klasternya sekaligus jaraknya dari klaster lain; nilai yang mendekati satu menunjukkan struktur yang lebih baik. DBI mengukur rasio dispersi internal terhadap pemisahan antarklaster, sehingga nilai yang lebih rendah lebih baik. Dunn Index membandingkan jarak minimum antarklaster dengan diameter maksimum klaster; nilai yang lebih tinggi menunjukkan pemisahan yang lebih baik. Karena DBSCAN dan HDBSCAN menghasilkan noise, seluruh metrik dihitung hanya pada observasi non-noise agar evaluasi mencerminkan struktur klaster yang terbentuk [9], [12].

III. METODE

A. Data Penelitian

Data penelitian merupakan data sekunder tahun 2023 yang mencakup 27 kabupaten/kota di Provinsi Jawa Barat. Data dihimpun dari publikasi Badan Pusat Statistik dan disatukan berdasarkan nama wilayah serta periode pengamatan yang sama. Delapan indikator numerik yang digunakan disajikan pada Tabel 1.

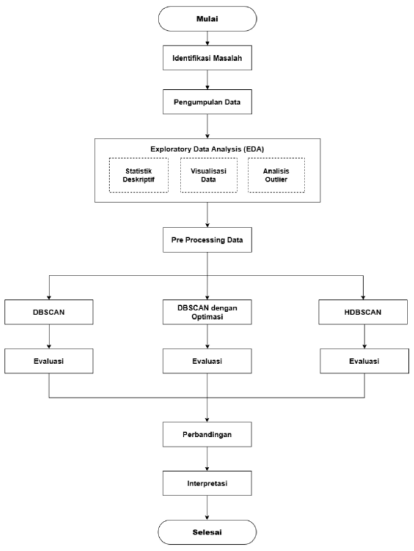
TABEL 1
(Indikator Penelitian)

Indikator	Representasi
TPT	penyerapan angkatan kerja
IPM	kualitas kesehatan, pendidikan, dan standar hidup
PDRB per kapita	kapasitas ekonomi rata-rata wilayah
Gini Ratio	ketimpangan distribusi pengeluaran
Sanitasi layak	akses rumah tangga terhadap sanitasi
Air minum layak	akses rumah tangga terhadap air minum
Rasio ketergantungan	beban penduduk usia nonproduktif
Rumah layak huni	kualitas dan kelayakan tempat tinggal

B. Praproses dan Rancangan Eksperimen

Tahap praproses diawali dengan pemeriksaan nilai kosong dan duplikasi. Distribusi data dianalisis melalui statistik deskriptif, histogram, dan boxplot. Nilai ekstrem ditangani secara terbatas melalui winsorization agar pengaruh observasi ekstrem terhadap rata-rata dan simpangan baku tidak berlebihan, tanpa menghapus wilayah dari dataset. Selanjutnya, seluruh variabel distandardisasi menggunakan StandardScaler.

Eksperimen terdiri atas tiga skenario. Pertama, DBSCAN diterapkan menggunakan parameter awal yang diperoleh dari grafik k-distance dengan nilai MinPts sebesar dua kali jumlah variabel. Kedua, parameter DBSCAN dioptimasi melalui Grid Search. Ruang pencarian terdiri atas epsilon 16,5-34,5 dengan interval 0,5 dan MinPts 8-15, sehingga terdapat 296 kombinasi. Konfigurasi terbaik dipilih berdasarkan struktur klaster valid dan nilai Silhouette tertinggi dengan mempertimbangkan DBI, Dunn Index, serta jumlah noise. Ketiga, HDBSCAN diterapkan dengan min_cluster_size sebesar 3 dan min_samples sebesar 2. Setiap hasil dievaluasi menggunakan metrik yang sama dan dibandingkan berdasarkan kualitas internal, cakupan pengelompokan, serta waktu eksekusi.



GAMBAR 1
(Alur penelitian)

IV. HASIL DAN PEMBAHASAN

A. Karakteristik Data

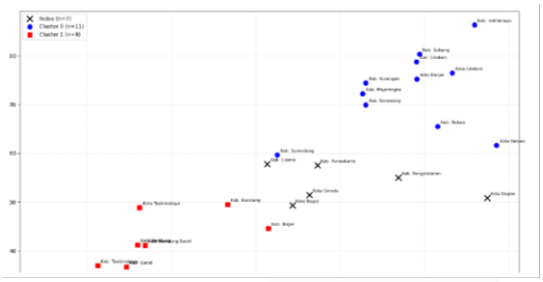
Dataset memiliki variasi yang nyata antarkabupaten/kota. PDRB per kapita menunjukkan penyebaran paling lebar dan distribusi menceng ke kanan karena hanya beberapa wilayah memiliki kapasitas ekonomi yang sangat tinggi. Akses sanitasi layak dan rumah layak

huni juga memiliki rentang besar, sedangkan rasio ketergantungan relatif lebih seragam. Kondisi tersebut menunjukkan bahwa pengelompokan berdasarkan satu indikator tidak cukup untuk menggambarkan perbedaan karakteristik wilayah. Standardisasi diperlukan agar PDRB per kapita tidak mendominasi perhitungan jarak.

B. DBSCAN Tanpa Optimasi dan Hasil Grid Search

Grafik k-distance dengan MinPts awal 16 menghasilkan epsilon sebesar 34,972. Kombinasi tersebut menggabungkan seluruh 27 wilayah ke dalam satu klaster. Karena tidak terdapat sekurang-kurangnya dua klaster, Silhouette Score, DBI, dan Dunn Index tidak dapat dihitung. Hasil ini memperlihatkan sensitivitas DBSCAN terhadap pemilihan epsilon; radius yang terlalu besar menghilangkan struktur perbedaan kepadatan pada data.

Grid Search menghasilkan parameter terbaik epsilon 16,5 dan MinPts 9. DBSCAN teroptimasi membentuk dua klaster dengan 20 wilayah non-noise dan tujuh wilayah berlabel noise. Klaster 0 terdiri atas 11 wilayah, sedangkan Klaster 1 terdiri atas sembilan wilayah. Evaluasi pada data non-noise menghasilkan Silhouette Score 0,6279, DBI 0,4882, dan Dunn Index 0,3288. Perubahan dari satu klaster menjadi dua klaster yang dapat dievaluasi menunjukkan bahwa pencarian parameter sistematis memperbaiki kemampuan DBSCAN dalam mengenali struktur data.



GAMBAR 2

(Visualisasi klaster DBSCAN teroptimasi)

C. Hasil HDBSCAN

HDBSCAN membentuk dua klaster yang mencakup 25 wilayah dan hanya menghasilkan dua noise, yaitu Kabupaten Pangandaran dan Kota Depok. Klaster 0 berisi sepuluh wilayah, sedangkan Klaster 1 berisi 15 wilayah. Silhouette Score yang diperoleh sebesar 0,5299, DBI sebesar 0,6054, dan Dunn Index sebesar 0,3703. Dibandingkan DBSCAN, HDBSCAN memasukkan lima wilayah tambahan ke dalam struktur klaster. Hal ini menunjukkan bahwa pendekatan hierarki lebih adaptif terhadap perbedaan kepadatan lokal.



Gambar 3. Visualisasi klaster HDBSCAN

D. Perbandingan Kualitas dan Cakupan

TABEL 2

(Perbandingan Hasil Klasterisasi)

Ukuran	DBSCAN	HDBSCAN
Jumlah klaster	2	2
Data terklaster	20	25
Noise	7	2
Cakupan	74,07%	92,59%
Silhouette	0,6279	0,5299
DBI	0,4882	0,6054
Dunn	0,3288	0,3703
Median waktu	1,0603 detik	4,1073 detik

Tabel 2 menunjukkan bahwa DBSCAN lebih baik berdasarkan dua metrik utama. Silhouette yang lebih tinggi menandakan anggota klaster DBSCAN lebih konsisten terhadap kelompoknya, sedangkan DBI yang lebih rendah menunjukkan dispersi internal dan pemisahan antarklaster yang lebih baik. Sebaliknya, Dunn Index HDBSCAN sedikit lebih tinggi. Perbedaan tersebut menunjukkan bahwa penilaian kualitas tidak dapat didasarkan pada satu metrik saja karena setiap indeks menekankan aspek geometris yang berbeda.

Cakupan pengelompokan memperlihatkan trade-off yang lebih jelas. DBSCAN hanya mengelompokkan 74,07% wilayah karena MinPts 9 menerapkan persyaratan kepadatan yang relatif ketat. HDBSCAN mengelompokkan 92,59% wilayah dan mengurangi noise dari tujuh menjadi dua. Dengan demikian, DBSCAN menghasilkan kelompok non-noise yang lebih kompak dan terpisah, sedangkan HDBSCAN lebih representatif untuk pemetaan sebagian besar wilayah. Perbandingan metrik perlu dibaca bersama jumlah data yang disertakan karena metrik DBSCAN dihitung pada 20 observasi, sedangkan metrik HDBSCAN dihitung pada 25 observasi.

Pada lingkungan pengujian dan ukuran data yang kecil, median waktu Grid Search DBSCAN sebesar 1,0603 detik, sedangkan HDBSCAN sebesar 4,1073 detik. Hasil ini bersifat spesifik terhadap implementasi, perangkat keras, dan dataset penelitian. HDBSCAN menjalankan tahapan tambahan berupa pembentukan mutual reachability distance, minimum spanning tree, hierarki klaster, dan evaluasi stabilitas. Oleh karena itu, hasil waktu eksekusi tidak dapat digeneralisasi secara langsung pada data yang lebih besar.

E. Interpretasi Karakteristik Wilayah

Kedua metode menghasilkan pola interpretasi yang relatif konsisten. Klaster 0 memiliki PDRB per kapita, akses sanitasi layak, akses air minum layak, dan persentase rumah layak huni yang relatif lebih tinggi. Gini Ratio serta rasio ketergantungannya cenderung lebih rendah. Karakteristik tersebut menunjukkan kapasitas ekonomi, pemerataan, fasilitas dasar, dan kualitas hunian yang relatif lebih baik dibandingkan Klaster 1.

Klaster 1 memiliki akses sanitasi dan rumah layak huni yang lebih rendah, disertai rasio ketergantungan dan ketimpangan yang relatif lebih tinggi. Perbedaan TPT dan IPM antara kedua klaster tidak sebesar perbedaan pada fasilitas dasar dan kondisi hunian. Temuan ini menunjukkan bahwa dimensi kemiskinan wilayah tidak hanya dibentuk oleh indikator ekonomi dan ketenagakerjaan, tetapi juga oleh kualitas layanan dasar.

Kabupaten Pangandaran dan Kota Depok tetap menjadi noise pada HDBSCAN karena memiliki kombinasi indikator yang tidak membentuk kepadatan stabil dengan dua kelompok utama. Kabupaten Pangandaran memiliki TPT sangat rendah tetapi akses air minum dan rasio ketergantungan yang berbeda dari mayoritas wilayah. Kota Depok memiliki IPM serta akses layanan dasar yang tinggi, tetapi pola rumah layak huni dan ketimpangannya berbeda. Status noise tidak menunjukkan bahwa wilayah tersebut

lebih baik atau lebih buruk, melainkan menandakan karakteristik yang perlu dianalisis secara individual dan tidak dipaksakan ke dalam kelompok umum.

Bagi analisis kebijakan, HDBSCAN dapat digunakan sebagai acuan ketika prioritasnya adalah cakupan pemetaan yang luas. DBSCAN teroptimasi lebih sesuai ketika fokus analisis diarahkan pada pembentukan kelompok inti yang kompak. Hasil klaster tidak dapat digunakan sebagai peringkat kemiskinan secara langsung karena label hanya menyatakan kemiripan multidimensional. Interpretasi kebijakan tetap memerlukan pemeriksaan nilai indikator setiap wilayah dan konteks lokal.

V. KESIMPULAN

Optimasi Grid Search berhasil memperbaiki kinerja DBSCAN dari satu klaster yang tidak dapat dievaluasi menjadi dua klaster dengan parameter epsilon 16,5 dan MinPts 9. DBSCAN teroptimasi mengelompokkan 20 wilayah dengan Silhouette Score 0,6279, Davies-Bouldin Index 0,4882, dan Dunn Index 0,3288. HDBSCAN mengelompokkan 25 wilayah dengan Silhouette Score 0,5299, Davies-Bouldin Index 0,6054, dan Dunn Index 0,3703. DBSCAN menghasilkan klaster non-noise yang lebih kompak dan terpisah berdasarkan Silhouette dan DBI, sedangkan HDBSCAN memiliki Dunn Index serta cakupan pengelompokan yang lebih tinggi dan jumlah noise yang lebih sedikit. Kedua metode mengidentifikasi satu kelompok dengan kapasitas ekonomi, fasilitas dasar, dan kondisi hunian relatif lebih baik serta satu kelompok dengan keterbatasan yang lebih besar. HDBSCAN lebih representatif untuk pemetaan mayoritas kabupaten/kota, sementara DBSCAN bermanfaat untuk memperoleh kelompok inti yang lebih ketat. Penelitian selanjutnya perlu menggunakan data multiyear, menguji kestabilan klaster, serta membandingkan metode berbasis kepadatan dengan pendekatan klasterisasi lain agar pola perubahan sosial ekonomi dapat dianalisis secara temporal.

REFERENSI

[1] I. M. Zuhendra and R. Hidayat, “Penerapan Data Mining untuk Klusterisasi Tingkat Kemiskinan Berdasarkan Data Terpadu Kesejahteraan Sosial (DTKS),” SKANIKA, vol. 7, no. 1, pp. 32-41, 2024.

[2] L. Priseptian and W. P. Primandhana, “Analisis faktor-faktor yang mempengaruhi kemiskinan,” Forum Ekonomi, vol. 24, no. 1, pp. 45-53, 2022.

[3] N. A. Yolandari et al., “Analisis perbandingan K-Means dan DBSCAN dalam pengelompokan data travel review ratings menggunakan evaluasi Silhouette Index dan Davies-Bouldin Index,” Jurnal Informatika dan Teknik Elektro Terapan, vol. 13, no. 3, 2025.

[4] X. Zhang and S. Zhou, “WOA-DBSCAN: Application of Whale Optimization Algorithm in DBSCAN Parameter Adaption,” IEEE Access, vol. 11, pp. 91861-91878, 2023.

[5] G. J. Monko and M. Kimura, “Optimized DBSCAN parameter selection: Stratified sampling for epsilon and grid search for minimum samples,” in Proc. AIRCC, pp. 43-61, 2023.

[6] X. Han, C. Armenakis, and M. Jadidi, “DBSCAN optimization for improving marine trajectory clustering and anomaly detection,” Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., pp. 455-461, 2020.

[7] G. Stewart and M. Al-Khassaweneh, “An implementation of the HDBSCAN clustering algorithm,” Applied Sciences, vol. 12, no. 5, 2022.

[8] I. N. Simbolon and P. D. Friskila, “Analisis dan evaluasi algoritma DBSCAN pada tuberkulosis,” Jurnal Informatika dan Teknik Elektro Terapan, vol. 12, no. 3S1, 2024.

[9] A. Saputra and R. Yusuf, “Perbandingan algoritma DBSCAN dan K-Means dalam segmentasi pelanggan pengguna transportasi publik Transjakarta menggunakan metode RFM,” MALCOM, vol. 4, no. 4, pp. 1346-1361, 2024.

[10] Z. N. Fauziyah, “Implementation of ST-DBSCAN algorithm to cluster regency/city in West Java Province based on natural disaster occurrence in 2020-2022,” Journal of Applied Statistics and Data Science, vol. 2, no. 1, pp. 36-44, 2025.

[11] Y. Zhao, W. Zhang, and X. Liu, “Grid search with a weighted error function: Hyper-parameter optimization for financial time series forecasting,” Applied Soft Computing, vol. 154, 2024.

[12] M. Syahrul et al., “Application of the DBSCAN algorithm in MSME clustering using the Silhouette Coefficient method,” 2024.