

Pengelompokan Kabupaten/Kota Berdasarkan Indikator Perumahan di Pulau Jawa Menggunakan *Hierarchical Density Based Spatial Clustering of Application with Noise* (HDBSCAN)

Rosyida Aurelia Fitri Riayah *, Ilham Faisal Mahdy

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

rosyida.aurelia@gmail.com, ilham.faishal@unisba.ac.id

Abstract. Cluster analysis is a method used to group data based on similar characteristics. One density-based cluster analysis is Hierarchical Density Based Spatial Clustering of Applications with Noise. HDBSCAN is a density-based clustering method capable of forming clusters with different densities and identifying data that does not belong to any cluster (noise). This method automatically forms clusters based on variations in local density. The distance used to measure the proximity between data points is the Euclidean distance. The clustering results are evaluated using Density-Based Clustering Validation (DBCV). DBCV is a validation index specifically designed for density-based methods, focusing on cohesion and separation between clusters. The DBCV value ranges from -1 to 1, with higher values indicating better clusters. In this study, the HDBSCAN method will be applied to data related to the 2023 Housing Indicators with observation points in the regencies/cities on the island of Java. Housing conditions are one of the supporting aspects of community welfare, but disparities in adequate housing still exist on the island of Java. Therefore, it is necessary to map areas based on housing indicators to identify patterns of similarity in characteristics between areas as a basis for development planning. From the cluster analysis results obtained using a minimum cluster size of 7 and a minimum sample size of 7, two clusters and 12 noise points were identified. Cluster 1 consists of 39 regions, and Cluster 2 consists of 60 regions. The DBCV value obtained was 0.5506, indicating a sufficiently stable cluster structure, so the clustering results can be used to interpret the characteristics of each regional group based on the indicators used

Keywords: DBCV, HDBSCAN, Housing Indicators.

Abstrak. Analisis klaster adalah metode yang digunakan untuk mengelompokkan data berdasarkan kemiripan karakteristik. Salah satu analisis klaster berbasis kepadatan adalah *Hierarchical Density Based Spatial Clustering of Application with Noise*. HDBSCAN merupakan klasterisasi berbasis kepadatan yang mampu membentuk klaster dengan kepadatan berbeda dan mengidentifikasi data yang tidak termasuk ke dalam klaster manapun (*noise*). Metode ini membentuk klaster secara otomatis berdasarkan variasi kerapatan lokal. Jarak yang digunakan untuk mengukur kedekatan antar data adalah jarak *euclidean*. Evaluasi hasil klaster dilakukan menggunakan *Density Based Clustering Validation* (DBCV) yang merupakan indeks validasi khusus untuk metode berbasis kepadatan yang memperhatikan kohesi dan separasi antar klaster. Nilai DBCV berada antara -1 hingga 1, dengan nilai yang lebih tinggi menunjukkan klaster semakin baik. Pada penelitian ini, metode HDBSCAN akan diterapkan pada data terkait Indikator Perumahan Tahun 2023 dengan titik pengamatan yakni Kabupaten/Kota yang ada di Pulau Jawa. Kondisi perumahan merupakan salah satu aspek penunjang kesejahteraan masyarakat, namun ketimpangan hunian layak masih terjadi di Pulau Jawa. Karena itu, diperlukan pemetaan wilayah berdasarkan indikator perumahan untuk mengetahui pola kemiripan karakteristik antar wilayah sebagai bahan perencanaan pembangunan. Dari hasil analisis klaster yang dilakukan dengan menggunakan *minimum cluster size = 7* dan *minimum sample = 7*, diperoleh 2 klaster dan 12 *noise*. Klaster 1 terdiri dari 39 daerah dan klaster 2 terdiri dari 60 daerah. Nilai DBCV yang diperoleh sebesar 0,5506 yang menunjukkan struktur klaster cukup stabil, maka hasil klasterisasi dapat digunakan untuk menginterpretasikan karakteristik tiap kelompok wilayah berdasarkan indikator yang digunakan.

Kata Kunci: DBCV, HDBSCAN, Indikator Perumahan.

A. Pendahuluan

Data mining adalah proses analisis data untuk mengidentifikasi pola-pola tertentu dengan tujuan memperoleh informasi yang bermanfaat serta dapat dikembangkan lebih lanjut (Suliman, 2021). Salah satu teknik utama dalam data mining adalah *clustering*. *Clustering* merupakan metode *unsupervised learning* yang digunakan untuk pengelompokan data berdasarkan kemiripan pola tertentu. Pendekatan *clustering* yang sering dipakai adalah *clustering* berbasis kepadatan yang digunakan untuk mengelompokkan data berdasarkan kepadatan suatu area seperti metode HDBSCAN (*Hierarchical Density Based Spatial Clustering of Applications With Noise*). HDBSCAN merupakan pengembangan dari metode *Density Based Spatial Clustering of Application with Noise* (DBSCAN) yang memiliki keterbatasan dalam mengidentifikasi klaster yang memiliki kepadatan berbeda-beda. Algoritma HDBSCAN secara hierarkis membangun struktur klaster dan otomatis memilih klaster yang paling stabil berdasarkan ukuran ketahanan (*stability cluster*). Selain itu, HDBSCAN unggul dalam mendeteksi *noise* yang tidak masuk dalam klaster manapun, sehingga cocok digunakan dalam data analisis dengan kepadatan yang tidak merata seperti pengelompokan Kabupaten/Kota di Pulau Jawa berdasarkan indikator perumahan.

Kesejahteraan masyarakat merupakan salah satu tujuan utama dalam pembangunan nasional berkelanjutan (*Sustainable Development Goals/SDGs*) khususnya pada TPB 3 (kesehatan), TPB 6 (akses air dan sanitasi), dan TPB 11 (perumahan layak) (Alisjahbana & Murniningtyas, 2021). Perumahan layak menjadi aspek penting dalam kehidupan serta merupakan hak dasar warga negara. Menurut BPS dan Kementerian PUPR, indikator perumahan menjadi cerminan dari tingkat kesejahteraan masyarakat. Berdasarkan data BPS, sekitar 36,85% rumah tangga di Indonesia masih tinggal di rumah tidak layak huni. Artinya, sekitar 36-37 dari setiap 100 rumah tangga hidup dalam kondisi hunian tidak layak, hal ini juga diperparah dengan peningkatan urbanisasi. BPS memprediksi pada tahun 2035 penduduk Indonesia yang tinggal di wilayah perkotaan akan mencapai 66.6% dan diprediksi akan terus meningkat sehingga dapat meningkatkan jumlah kawasan kumuh (Fadlurohman & Nur, 2023).

Pulau Jawa merupakan pusat aktivitas ekonomi dan urbanisasi nasional sekaligus wilayah dengan jumlah penduduk terbanyak. Sebagai pusat pertumbuhan ekonomi, Pulau Jawa menghadapi tantangan besar dalam penyediaan hunian layak bagi masyarakat. Meskipun Pulau Jawa merupakan pusat pertumbuhan ekonomi, namun tidak semua wilayah mencapai nilai yang optimal. Provinsi DIY menjadi wilayah dengan nilai tertinggi yang memiliki akses hunian layak yakni sekitar 86%, sedangkan Provinsi DKI Jakarta memiliki nilai terendah yaitu sekitar 39% bahkan di tahun 2022 mengalami penurunan dari tahun sebelumnya. Ketimpangan ini menunjukkan bahwa pembangunan perumahan belum merata. Hal ini dapat berdampak pada kualitas hidup masyarakat baik dari aspek kesehatan hingga kesejahteraan masyarakat. Oleh karena itu, diperlukan analisis spasial yang dapat membantu memahami karakteristik wilayahnya.

Penelitian dengan metode HDBSCAN sebelumnya pernah dilakukan oleh (Amalina & Fauzan, 2025) yang mengidentifikasi desa-desa potensial di Kabupaten Buleleng. Dari penelitian tersebut, diperoleh 2 klaster dengan nilai *silhouette coefficient* sebesar 0,37. Penelitian lain yang dilakukan oleh (Wahyuni et al., 2021) yang mengelompokkan wilayah Desa/Kelurahan tertinggal di Kabupaten Kutai Kertanegara dengan metode HDBSCAN, juga memperoleh 2 klaster dengan nilai *silhouette coefficient* sebesar 0,184. Pada penelitian (Fitrianto, 2024) yang membandingkan performa algoritma HDBSCAN dan tiga metode hierarki *agglomerative* yaitu *ward*, *rataan* dan *lengkap* dalam mengelompokkan data Produk Domestik Regional Bruto Atas Dasar Harga Konstan (PDRB ADHK) Provinsi Jawa Timur tahun 2023. HDBSCAN terbukti lebih efektif dibandingkan *Agglomerative Hierarchical Clustering* (AHC) dalam menangani data yang mengandung pencilan. Berdasarkan hal tersebut, penelitian ini juga menggunakan metode HDBSCAN (*Hierarcical Density Basesd Spatial Clustering of Applications with Noise*) untuk mengelompokkan Kabupaten/Kota di Pulau Jawa berdasarkan indikator perumahan pada tahun 2023. Metode ini dapat mengelompokkan tanpa harus menentukan jumlah klaster di awal serta mampu mengidentifikasi daerah dengan kepadatan bervariasi dan mendeteksi *noise*.

Berdasarkan latar belakang yang telah dijabarkan, tujuan dari penelitian ini yaitu:

1. Mengidentifikasi pengelompokan Kabupaten/Kota di Pulau Jawa berdasarkan Indikator Perumahan Tahun 2023 menggunakan metode *Hierarchical Density-Based Spatial Clustering of Application with Noise* (HDBSCAN).

2. Menganalisis kinerja metode *Hierarchical Density-Based Spatial Clustering of Application with Noise* (HDBSCAN) dalam mengelompokkan Kabupaten/Kota di Pulau Jawa berdasarkan Indikator Perumahan Tahun 2023.

B. Metode

Penelitian ini, menggunakan metode pendekatan kuantitatif dengan data sekunder yaitu data indikator perumahan tahun 2023 pada Kabupaten/Kota di Provinsi Jawa Barat, Jawa Tengah, Jawa Timur, DI Yogyakarta, dan DKI Jakarta dengan total 111 data yang diperoleh dari *website* Badan Pusat Statistik (BPS). Metode analisis yang digunakan adalah *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) untuk mengelompokkan wilayah berdasarkan 9 variabel indikator perumahan. Berikut variabel yang digunakan.

Tabel 1. Deskripsi Variabel Penelitian

Variabel	Nama Variabel	Satuan
X ₁	Persentase Rumah Tangga menurut Status Kepemilikan Rumah Milik Sendiri	Persen (%)
X ₂	Persentase Rumah Tangga menurut Status Kepemilikan Rumah Kontrak/sewa	Persen (%)
X ₃	Persentase Rumah Tangga menurut Jenis Atap Terluas yang Layak	Persen (%)
X ₄	Persentase Rumah Tangga menurut Jenis Dinding Terluas yang Layak	Persen (%)
X ₅	Persentase Rumah Tangga menurut Jenis Lantai Terluas yang Layak	Persen (%)
X ₆	Persentase Rumah Tangga menurut Sumber Air Minum Layak	Persen (%)
X ₇	Persentase Rumah Tangga menurut Sanitasi Layak	Persen (%)
X ₈	Persentase Rumah Tangga menurut Sumber Penerangan dari Listrik	Persen (%)
X ₉	Persentase Rumah Tangga menurut Luas Hunian per kapita $\leq 7,2\text{m}^2$	Persen (%)

Hierarchical Density Based Spatial Clustering of Application with Noise (HDBSCAN)

HDBSCAN merupakan gabungan antara metode hierarki dan berbasis kepadatan. Kepadatan ini mengacu pada jumlah titik dalam suatu wilayah, dimana wilayah dengan konsentrasi titik data yang tinggi dianggap sebagai klaster, sedangkan wilayah dengan konsentrasi titik data yang rendah akan diklasifikasikan sebagai noise (Han et al., 2011). HDBSCAN menggunakan dua parameter yaitu *Minimum sampel* dan *minimum cluster size*. *Minimum sampel* mengatur jumlah tetangga minimum untuk titik tertentu agar dianggap sebagai inti klaster (*core point*) yang digunakan dalam perhitungan jarak untuk menentukan jarak keterjangkauan antara titik, sedangkan *minimum cluster size* menentukan batas bawah jumlah titik data agar membentuk klaster yang valid (Stewart & Al-Khassaweneh, 2022). HDBSCAN menggunakan *Minimum Spanning Tree* (MST) untuk menentukan hubungan antar titik dalam data dengan ukuran *reachability distance* yang membantu dalam mengidentifikasi struktur hierarki. Dengan menghapus sisi dalam MST secara bertahap, HDBSCAN membentuk klaster berdasarkan kepadatan berbeda, dimana klaster dengan kepadatan tinggi bertahan lebih lama dibandingkan klaster dengan kepadatan rendah. Hasil akhirnya adalah klaster yang lebih stabil dan representatif terhadap variasi kepadatan dalam data (Pepe, n.d.). Proses HDBSCAN terdiri dari beberapa langkah utama (Stewart & Al-Khassaweneh, 2022):

1. Menghitung *Core Distance*

Core distance merupakan jarak terdekat untuk mencapai titik tetangga ke- k yang digunakan untuk menentukan kepadatan suatu titik dikelilingi oleh titik data lainnya.

$$d_{core} = k\text{-th_nearest_neighbor_distance}(i) \quad \dots(1)$$

dengan, d_{core} : jarak inti titik ke- i , k : jumlah tetanga yang ditentukan, $kth_nearest_neighbor_distance(i)$: jarak antar titik ke- i dengan tetangga ke- k terdekat berdasarkan jarak *euclidean*

2. Menghitung *Mutual Reachability Distance* (MRD)

Mutual Reachability Distance merupakan jarak yang diperhitungkan untuk membentuk kluster berdasarkan kepadatan titik. MRD dinotasikan:

$$d_{mreach}(x_p, x_q) = \max\{d_{core}(x_p), d_{core}(x_q), d(x_p, x_q)\} \quad \dots(2)$$

dengan, $d_{core}(x_p)$: *core distance* dari titik x_p , $d_{core}(x_q)$: *core distance* dari titik x_q , $d(x_p, x_q)$: jarak *euclidean* antara titik x_p dan x_q

3. *Minimum Spanning Tree* (MST)

Minimum Spanning Tree (MST) merupakan struktur graf bobot yang menghubungkan semua titik dalam data dengan panjang total minimum dan memastikan seluruh titik terhubung, sehingga dapat merepresentasikan struktur kepadatan data secara keseluruhan sebagai dasar bagi HDBSCAN dalam proses pembentukan kluster.

4. Membangun Hierarki Kluster

Setelah MST terbentuk berdasarkan MRD, hierarki kluster akan dibangun dengan menghapus sisi (*edge*) besar secara bertahap. MST dipotong sedikit demi sedikit berdasarkan bobot sisi dari terpanjang ke yang terpendek. Setiap kali suatu sisi dipotong, maka terjadi pemisahan (*split*) kluster. Jika hasil split membentuk kelompok dengan jumlah titik kurang dari minimum cluster size, maka dianggap sebagai *noise*.

5. Menentukan Kluster yang Stabil

Kluster dengan *stability score* tertinggi dianggap sebagai kluster paling konsisten dalam merepresentasikan struktur pengelompokan data. *Stability cluster* $S(C_i)$ dihitung dengan:

$$S(C_i) = \sum_{p \in \text{Cluster}} (\lambda_p - \lambda_{brith}) \quad \dots(3)$$

dengan, $S(C_i)$: *stability cluster*, λ_{brith} : nilai λ saat kluster C terbentuk, λ_p : nilai λ saat titik p keluar dari kluster (jadi *noise*).

Density-Based Clustering Validation (DBCV)

DBCV merupakan metode evaluasi hasil clustering yang digunakan khusus untuk algoritma clustering berbasis kepadatan seperti DBSCAN atau HDBSCAN. DBCV menilai kualitas kluster dengan mempertimbangkan hubungan kepadatan antar objek, bukan jarak antar pusat. Indeks ini menggunakan konsep *core distance* dan *mutual reachability distance* untuk membangun *minimum spanning tree* di setiap kluster, sehingga dapat menangkap baik bentuk atau kepadatan kluster secara bersamaan. DBCV dapat dihitung dengan membandingkan dua aspek utama, yaitu *density sparseness* dan *density separation* menggunakan rumus (Moulavi et al., 2014):

$$VC(C_i) = \frac{\min_{1 \leq j \leq l, j \neq i} (DSPC(C_i, C_j)) - DSC(C_i)}{\max_{1 \leq j \leq l, j \neq i} (\min_{1 \leq j \leq l, j \neq i} (DSPC(C_i, C_j)) - DSC(C_i))} \quad \dots(4)$$

dengan, $VC(C_i)$: *Validity Coefficient of Cluster* C_i , DSC: nilai maksimum dari semua *edge* di *Minimum Spanning Tree* (MST) internal kluster C_i , DSPC: nilai minimum *mutual reachability distance* antara anggota kluster C_i dan kluster C_j .

Kemudian, DBCV keseluruhan dihitung sebagai:

$$DBCV = \sum_{i=1}^l \frac{|C_i|}{|O|} \times VC(C_i) \quad \dots(5)$$

dengan, $|C_i|$: jumlah titik data dalam kluster C_i , $|O|$: jumlah total titik data, l : jumlah kluster.

Nilai DBCV berkisar dari -1 hingga 1, dimana nilai mendekati 1 menunjukkan struktur kluster yang lebih baik, nilai DBCV mendekati -1 menunjukkan struktur yang buruk atau berlawanan.

C. Hasil Penelitian dan Pembahasan

Bagian Data Pre-Processing

Setelah pemeriksaan data, diperoleh bahwa data pada penelitian ini memiliki missing value yaitu pada variabel X_9 di tiga wilayah (Boyolali, Sragen, dan Blora). Selanjutnya akan dilakukan penanganan terhadap data yang hilang dengan imputasi *median*:

$$Median = \frac{X_{16} + X_{17}}{2} = \frac{1,47 + 1,55}{2} = 1,51 \quad \dots(6)$$

Diperoleh hasil 1,51 yang digunakan pada variabel X_9 di ketiga wilayah yang nilainya hilang tersebut.

Pemeriksaan Multikolinearitas

Pemeriksaan ini dilakukan dengan memeriksa nilai *Variance Inflation Factor* (VIF) untuk setiap variabel bebas.

Tabel 2. Hasil Pemeriksaan Multikolinearitas

Variabel	VIF
X_1	4,8387
X_2	4,4913
X_3	1,2781
X_4	1,9318
X_5	2,2971
X_6	1,4198
X_7	1,4121
X_8	1,1514
X_9	1,4923

Berdasarkan nilai pada tabel di atas, terlihat bahwa masing-masing variabel memiliki nilai VIF kurang dari 10. Sehingga disimpulkan bahwa pada masing-masing variabel tidak memiliki multikolinearitas yang tinggi.

Analisis HDBSCAN

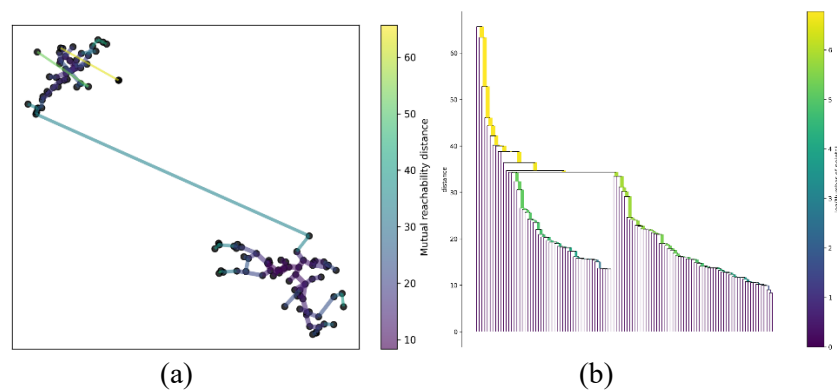
Dalam pengelompokkan menggunakan metode HDBSCAN, terdapat 2 parameter penting yang mempengaruhi proses pengelompokannya, yakni *minimum cluster size* dan *minimum sample*.

Tabel 3. Kombinasi Parameter dengan DBCV Tertinggi

<i>Minimum Cluster Size</i>	<i>Minimum Sampel</i>	DBCV
7	7	0,5506
9	9	0,5474
8	8	0,5449
10	10	0,5424
12	12	0,5393

Berdasarkan hasil eksplorasi kombinasi parameter *minimum cluster size* dan *minimum sampel*, diperoleh DBCV optimal sebesar 0,5506 dengan kombinasi *minimum cluster size* = 7 dan *minimum sampel* = 7.

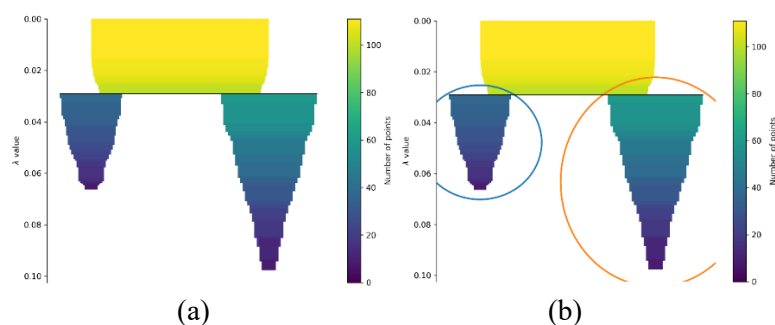
Setelah diperoleh nilai *minimum cluster size* dan *minimum sample* yang optimal, maka selanjutnya menghitung *core distance* dimana *minimum sample* menentukan tetangga ke- k yang digunakan. Pada penelitian ini, karena *min sample* = 7, maka *core distance* dari suatu titik adalah jarak ke tetangga terdekat ketujuh. Selanjutnya, dilakukan perhitungan *mutual reachability distance* dan dibentuk *minimum spanning tree* (MST) berdasarkan jarak-jarak tersebut.



Gambar 1. (a) Minimum Spanning Tree (MST), (b) Cluster Hierarki Graph

Gambar 4(a) menggambarkan *minimum spanning tree* (MST) dari data indikator perumahan tahun 2023. Berdasarkan Gambar 1(a), grafik tersebut memberikan visualisasi MST yang menggunakan *mutual reachability distance* (MRD) sebagai dasar perhitungan antar titik. Warna pada garis menunjukkan besarnya nilai *mutual reachability distance*, dimana warna yang semakin gelap menandakan semakin kecil jaraknya. Sebaran titik pada dua area besar yang terpisah itu mengindikasikan terdapat dua struktur kelompok utama dalam data yang akan diperkuat melalui pembentukan klaster stabil di tahap selanjutnya.

Selanjutnya, dibangun hierarki klaster dengan mengurutkan atau menghapus garis dalam MST berdasarkan nilai bobotnya, dari yang terbesar ke terkecil. Proses ini akan menghasilkan struktur berbentuk pohon atau dikenal dengan *dendrogram* dan divisualisasikan dalam bentuk grafik *single linkage tree* seperti ditunjukkan pada Gambar 1(b). Sumbu vertikal (*distance*) menunjukkan jarak atau kepadatan yang dihitung dengan *mutual reachability distance*. Warna batang menunjukkan logaritma jumlah titik dalam tiap cabang, dimana warna ungu menunjukkan jumlah titik yang sedikit dan warna kuning menunjukkan jumlah titik yang banyak. Sementara batang horizontal menunjukkan penggabungan klaster berdasarkan jarak tertentu. Semakin tinggi suatu cabang di grafik, semakin besar jarak antar anggotanya dan kemungkinan menjadi batas pemisah antar klaster. Klaster-klaster yang stabil akan membentuk cabang panjang dengan vertikal tebal dan tidak cepat hilang, sedangkan cabang yang pendek dan cepat hilang atau putus diidentifikasi sebagai *noise*.



Gambar 2. Condensed Tree

Berdasarkan Gambar 4.3 visualisasi *condensed tree* tersebut menggambarkan proses pembentukan dan penggabungan klaster hierarki dari data indikator perumahan. Pada Gambar 2(a) di atas, diperoleh 2 klaster yang terbentuk yang ditandai dengan lingkaran dengan warna berbeda pada Gambar 2(b). Klaster 1 ditandai dengan lingkaran berwarna biru dan klaster 2 ditandai dengan lingkaran berwarna orange. Sementara itu, terdapat kelompok *noise* atau data yang tidak cukup padat untuk dimasukkan ke kelompok manapun. Anggota dari setiap klaster dilihat dari skala warna pada batang di bawah *condensed tree*. Berdasarkan gambar di atas, klaster 1 memiliki 39 anggota, klaster 2 memiliki 60 anggota dan 12 *noise*. *Noise* ini tidak masuk dalam klaster manapun karena tidak berada dalam area yang cukup padat atau tidak stabil, sehingga tidak tahan terhadap perubahan tingkat kepadatan dan menyebabkan titik-titik dibuang dan dianggap *noise*.

Setelah klaster terbentuk, langkah selanjutnya adalah menghitung *stability* untuk setiap klaster. Nilai ini digunakan untuk mengukur kekuatan atau konsistensi struktur klaster dalam hierarki kepadatan. Semakin besar nilainya, maka semakin kuat keterhubungan antar titik di dalamnya. Pada penelitian ini, diperoleh 2 klaster dengan nilai sebagai berikut.

Tabel 4. Stability Cluster

<i>Cluster</i>	<i>Stability Cluster</i>
1	1,4576
2	4,1098

Klaster 1 memiliki nilai *stability* sebesar 1,4576, sedangkan klaster 2 memiliki nilai *stability cluster* lebih tinggi dibanding klaster 1 yakni sebesar 4,1098. Artinya klaster pertama menunjukkan kelompok wilayah dengan variasi karakteristik yang lebih tinggi dibanding klaster kedua yang berisi wilayah dengan indikator lebih seragam.

Evaluasi Klaster

Evaluasi klaster pada penelitian ini menggunakan metode *Density-Based Clustering Validation* (DBCV). Metode ini dirancang khusus untuk mengevaluasi hasil klaster berbasis kepadatan seperti HDBSCAN. Evaluasi dilakukan berdasarkan hasil klasterisasi terhadap indikator perumahan berdasarkan Kabupaten/Kota di Provinsi Jawa Barat, Jawa Timur, Jawa Tengah, DKI Jakarta dan DI Yogyakarta. Diperoleh nilai DBCV sebesar 0,5506 yang mengindikasikan struktur klaster terbentuk dengan baik. Nilai ini menunjukkan objek-objek dalam satu klaster memiliki kepadatan yang cukup tinggi dan objek antar klasternya memiliki kepadatan yang berbeda. Secara umum, nilai DBCV ini mampu menghasilkan klaster yang valid untuk menggambarkan pengelompokan kabupaten/kota berdasarkan indikator perumahan tahun 2023.

Interpretasi Hasil



Gambar 3. hasil visualisasi menggunakan HDBSCAN

Berdasarkan hasil visualisasi menggunakan indikator perumahan yang telah dikelompokkan menggunakan HDBSCAN. Setiap titik pada gambar, menggambarkan satu wilayah kabupaten/kota, dan masing-masing klaster ditunjukkan dengan warna yang berbeda. Berdasarkan visualisasi di atas, diperoleh 2 klaster yaitu klaster 1 digambarkan dengan warna biru yang berjumlah 39 anggota dan klaster 2 digambarkan dengan warna orange yang berjumlah 60 anggota. Sementara warna merah menggambarkan wilayah yang tidak masuk dalam klaster manapun atau *noise* yang berjumlah 12 titik. Berikut ringkasan anggota pada masing-masing klaster.

Tabel 5. Distribusi Desa Berdasarkan Metode HDBSCAN

Klaster	Anggota Klaster	Banyaknya Anggota
1	Kota Surabaya, Cilacap, Banyumas, Kebumen, Purworejo, Magelang, Boyolali, Klaten, Sukoharjo, Wonogiri, Karanganyar, Sragen, Rembang, Pati, Kudus, Jepara, Demak, Semarang, Temanggung, Kendal, Batang, Pekalongan, Pemalang, Tegal, Brebes, Kota Magelang, Kota Surakarta, Kota Salatiga, Kota Semarang, Kota Pekalongan, Kota Tegal, Kepulauan Seribu, Kota Jakarta Selatan, Kota Jakarta Timur, Kota Jakarta Pusat, Kota Jakarta Barat, Kota Jakarta Utara, Bantul, Sleman.	39
2	Bogor, Sukabumi, Cianjur, Bandung, Garut, Tasikmalaya, Ciamis, Kuningan, Cirebon, Majalengka, Sumedang, Indramayu, Subang, Purwakarta, Karawang, Bekasi, Bandung Barat, Pangandaran, Kota Bogor, Kota Sukabumi, Kota Cirebon, Kota Bekasi, Kota Depok, Kota Cimahi, Kota Tasikmalaya, Kota Banjar, Pacitan, Ponorogo, Trenggalek, Tulungagung, Blitar, Kediri, Malang, Lumajang, Jember, Banyuwangi, Bondowoso, Probolinggo, Pasuruan, Sidoarjo, Mojokerto, Jombang, Nganjuk, Madiun, Magetan, Ngawi, Tuban, Lamongan, Gresik, Sampang, Pemekasan, Sumenep, Kota Kediri, Kota Blitar, Kota Malang, Kota Probolinggo, Kota Pasuruan, Kota Mojokerto, Kota Madiun, Kota Batu.	60
Noise	Kota Bandung, Situbondo, Bojonegoro, Bangkalan, Purbalingga, Banjarnegara, Wonosobo, Grobogan, Blora, Kulon Progo, Gunung Kidul, Yogyakarta.	12

Selanjutnya dilakukan analisis rata-rata dari masing-masing klaster terhadap rata-rata seluruh objek. Rata-rata masing-masing variabel secara keseluruhan dinotasikan sebagai $\bar{X}$, sedangkan rata-rata pada variabel masing-masing klaster dinotasikan sebagai $\bar{X}_c$. Berikut ringkasan karakteristik dari masing-masing variabel pada klaster 1 dan klaster 2 terhadap rata-rata keseluruhan.

Tabel 6. Rata-rata klaster

Klaster	Variabel	$\bar{X}_c$	$\bar{X}$
Klaster 1	X ₁	83,2685	86,1368
	X ₂	5,3741	4,1960
	X ₃	95,1610	95,8458
	X ₄	91,3592	90,6451
	X ₅	94,2354	93,8236
	X ₆	95,2420	94,9257
	X ₇	88,9449	83,5346
	X ₈	99,9928	99,9749
	X ₉	6,2938	49,3646
Klaster 2	X ₁	87,4413	86,1368
	X ₂	3,5155	4,1960
	X ₃	98,0175	95,8458
	X ₄	93,3907	90,6451
	X ₅	96,7683	93,8236
	X ₆	95,1158	94,9257
	X ₇	81,4185	83,5346
	X ₈	99,9625	99,9749
	X ₉	82,9737	49,3646

Noise	X ₁	88,9358	86,1368
	X ₂	3,7700	4,1960
	X ₃	87,2125	95,8458
	X ₄	74,5967	90,6451
	X ₅	77,7617	93,8236
	X ₆	89,6967	94,9257
	X ₇	76,5317	83,5346
	X ₈	99,9783	99,9749
	X ₉	21,2992	49,3646

Berdasarkan Tabel 6, klaster 1 menunjukkan kondisi perumahan dengan X₂, X₄, X₅, X₆, X₇, X₈ yang lebih tinggi dari rata-rata secara keseluruhan serta proporsi rumah sempit yang rendah, menandakan hunian lebih layak. Klaster 2 unggul pada struktur fisik rumah X₁, X₃- X₆, X₈, namun sanitasi lebih rendah dan presentasi rumah sempit yang tinggi mencerminkan keterbatasan ruang hunian. Sementara itu, kelompok *noise* menunjukkan pola kepemilikan yang tidak konsisten sebagaimana kelompok lain, sehingga tidak dimasukkan ke dalam klaster manapun.

D. Kesimpulan

Berdasarkan hasil pengelompokan data indikator perumahan Kabupaten/Kota di Pulau Jawa Tahun 2023 dengan menggunakan metode HDBSCAN menghasilkan 2 klaster dengan 12 *noise*. Hasil tersebut diperoleh dengan menggunakan *minimum cluster size* sebesar 7 dan *minimum sample* sebesar 7. Klaster 1 terdiri dari 39 wilayah dan klaster 2 terdiri dari 60 wilayah. Klaster 1 dan klaster 2 memiliki karakteristik perumahan yang berbeda secara signifikan. Klaster 1 cenderung memiliki hunian lebih layak dari segi luas hunian, sebaliknya klaster 2 memerlukan perhatian lebih pada aspek luas hunian (X₉) karena rata-rata persentasenya yang tinggi mengindikasikan permasalahan serius dalam aspek hunian sempit. Sedangkan untuk klaster 2, unggul dalam aspek kepemilikan dan kelengkapan infrastruktur perumahan. Sementara itu, wilayah-wilayah dalam kategori *noise* menunjukkan pola indikator perumahan yang tidak konsisten sehingga tidak dapat dikelompokkan ke dalam klaster manapun. Pengelompokan ini menghasilkan klaster yang cukup baik, dilihat dari nilai *Density-Based Clustering Validation* (DBCV) yakni sebesar 0,5506 yang berarti setiap klaster mampu terpisahkan dengan baik menggunakan metode HDBSCAN.

Berdasarkan hasil penelitian, untuk penelitian selanjutnya disarankan untuk mencakup wilayah lebih luas seperti seluruh Provinsi di Indonesia termasuk Provinsi Banten yang pada penelitian ini tidak dapat dianalisis karena keterbatasan data. Dengan cakupan wilayah yang lebih luas, hasil pengelompokan akan lebih representatif secara nasional. Hasil clustering ini juga dapat dijadikan dasar dalam merancang strategi peningkatan akses hunian yang layak terutama di wilayah yang teridentifikasi ke dalam hunian tidak layak untuk dapat meningkatkan kualitas hidup serta kesejahteraan masyarakat di setiap Kabupaten/Kota di Pulau Jawa.

Ucapan Terimakasih

Penelitian ini tidak akan selesai tanpa bantuan serta dukungan dari banyak pihak. Penulis ingin mengucapkan terima kasih kepada semua pihak yang telah ikut membantu dalam penyelesaian penelitian ini. Terutama kepada Bapak Ilham Faishal Mahdy, S.Stat., M. Stat., sebagai dosen pembimbing yang telah memberikan ilmu, saran, dan waktu yang berharga bagi penulis. Tidak lupa juga kepada orang tua serta teman-teman saya yang telah memberikan dukungan moral sampai akhir.

Daftar Pustaka

Ajeng Mega Pratiwi, & Mutaqin, A. K. (2021). Penerapan Algoritma Naïve Bayes Classifier dalam Memprediksi Status Keberlanjutan Polis Nasabah Asuransi PT.X. *Jurnal Riset Statistika*, 1(2), 117–126. <https://doi.org/10.29313/jrs.v1i2.435>

- Alisjahbana, A. S., & Murniningtyas, E. (2021). Sustainable Transport, Sustainable Development. In *Sustainable Transport, Sustainable Development*. <https://doi.org/10.18356/9789210010788>
- Amalina, D. N., & Fauzan, A. (2025). *A Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) Approach for Identifying Potential Villages in Buleleng Regency. December 2024*. <https://doi.org/10.17977/um018v7i22024p187-199>
- Dewi, E. F., & Hajarisman, N. (2023). Penanganan Data Hilang pada Pemodelan Persamaan Terstruktur melalui Metode Full Information Maximum Likelihood (FIML). *DataMath: Journal of Statistics and Mathematics*, 1(1), 11–18. <https://doi.org/10.29313/datamath.v1i1.10>
- Fadlurohman, A., & Nur, I. M. (2023). Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Perumahan dan Kesehatan Lingkungan Menggunakan Metode K-Medoids. *Prosiding Seminar Nasional UNIMUS*, 6, 1168–1180.
- Fitrianto, N. (2024). *PERBANDINGAN ALGORITMA HDBSCAN DAN AGGLOMERATIVE HIERARCHICAL CLUSTERING DALAM KLASERISASI PADA DATA YANG MENGANDUNG*. 08(02), 122–135.
- Han, J., Kambe, M., & Pe, J. (2011). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. <https://doi.org/10.1016/C2009-0-61819-5>
- Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., & Sander, J. (2014). Density-based clustering validation. *SIAM International Conference on Data Mining 2014, SDM 2014*, 2(i), 839–847. <https://doi.org/10.1137/1.9781611973440.96>
- Nastiti, R. S., & Mutaqin, A. K. (2022). Penerapan Model Komposit Weibull-Pareto pada data Klaim Asuransi Harta Benda. *Jurnal Riset Statistika*, 43–49. <https://doi.org/10.29313/jrs.vi.903>
- Pepe, B. (n.d.). *Understanding HDBSCAN and Density-Based Clustering*.
- Stewart, G., & Al-Khassaweneh, M. (2022). An Implementation of the HDBSCAN* Clustering Algorithm. *Applied Sciences (Switzerland)*, 12(5), 1–21. <https://doi.org/10.3390/app12052405>
- Suliman, S. (2021). Implementasi Data Mining Terhadap Prestasi Belajar Mahasiswa Berdasarkan Pergaulan dan Sosial Ekonomi Dengan Algoritma K-Means Clustering. *Simkom*, 6(1), 1–11. <https://doi.org/10.51717/simkom.v6i1.48>
- Wahyuni, N. A., Hayati, M. N., & Rizki, A. (2021). Metode Hierarchical Density-Based Spatial Clustering of Application with Noise (HDBSCAN) Pada Wilayah Desa / Kelurahan Tertinggal di Kabupaten Kutai Kartanegara (Studi Kasus : Data Hasil Pendataan Potensi Desa (PODES) Tahun 2018) Hierarchical Densit. *Eksponensial*, 12(1), 47–52.