Jurnal EurekaMatika

Journal homepage: https://ejournal-science.upi.edu/jem/index
doi: https://doi.org/10.17509/jem.v14i2.643

Pengelompokan Provinsi di Indonesia dengan Metode Hierarchical Density Based Spatial Clustering with Noise (HDBSCAN)

(Studi Kasus: Data Perekonomian Provinsi di Indonesia Tahun 2024)

*Ripa Rifayatul Alawiyah, Nar Herrhyanto, dan Husty Serviana Husain**

Program Studi Matematika, FPMIPA, Universitas Pendidikan Indonesia
*Correspondence author: serviana@upi.edu

A B S T R A K

Ketimpangan ekonomi antarprovinsi di Indonesia masih menjadi tantangan dengan data rasio gini tercatat sebesar 0,381 pada September 2024. Penelitian ini mengelompokkan provinsi di Indonesia menggunakan metode HDBSCAN berdasarkan indikator ekonomi tahun 2024 dengan metrik jarak Euclidean dan Manhattan serta dievaluasi menggunakan Silhouette Score (SC) dan Davies-Bouldin Index (DBI). Hasil menunjukkan bahwa pada metrik Euclidean diperoleh parameter terbaik min_samples = 2 dan min_cluster_size = 3 dengan SC = 0,4443, DBI = 0,6761, dan cakupan 89,47%, yang menghasilkan dua klaster dan empat noise. Pada metrik Manhattan diperoleh parameter terbaik min_samples = 1 dan min_cluster_size = 2 dengan SC = 0,5730, DBI = 0,5591, dan cakupan 100%, yang menghasilkan dua klaster tanpa noise. Kedua pendekatan menunjukkan pola pengelompokan yang relatif serupa dalam mengidentifikasi provinsi dengan karakteristik ekonomi yang lebih rendah, sedangkan metrik Euclidean memberikan pembagian yang lebih rinci karena memisahkan provinsi dengan produksi padi dominan. Analisis sensitivitas menunjukkan bahwa pemisahan tersebut hilang ketika indikator produksi padi dikeluarkan, sedangkan kelompok provinsi dengan karakteristik ekonomi rendah tetap bertahan.
© 2026 Kantor Jurnal dan Publikasi UPI

INFORMASI ARTIKEL

Sejarah Artikel:

Diterima 18 Mei 2026
Direvisi 22 Oktober 2026
Disetujui 29 Oktober 2026
Tersedia online 5 November 2026
Dipublikasikan 5 November 2026

Kata Kunci:
Analisis Klaster,
Klasterisasi,
Ketimpangan Ekonomi,
Pembangunan Daerah

A B S T R A C T

Economic inequality between provinces in Indonesia remains a challenge, with the Gini ratio recorded at 0.381 in September 2024. This study clusters the provinces of Indonesia using the HDBSCAN method based on 2024 economic indicators, applying Euclidean and Manhattan distances and evaluating the results with the Silhouette Score (SC) and the Davies-Bouldin Index (DBI). Under the Euclidean metric, the best parameters were min_samples = 2 and min_cluster_size = 3 (SC = 0.4443; DBI = 0.6761; coverage 89.47%), producing two clusters and four noise points. Under the Manhattan metric, the best parameters were min_samples = 1 and min_cluster_size = 2 (SC = 0.5730; DBI = 0.5591; coverage 100%), producing two clusters without noise. Both approaches showed a relatively similar clustering pattern in identifying provinces with lower economic characteristics, while the Euclidean partition provided a finer split by isolating the provinces with dominant rice production. A sensitivity analysis showed that this split disappears when the rice production indicator is removed, whereas the group of provinces with low economic characteristics remains.
© 2026 Kantor Jurnal dan Publikasi UPI

Keywords:
HDBSCAN,
Cluster Analysis,
Clustering,
Economic Inequality,
HDBSCAN,
Regional Development

1. PENDAHULUAN

Perekonomian daerah menentukan kesejahteraan masyarakat. Kesejahteraan yang memadai tidak hanya mencerminkan kekuatan ekonomi, tetapi juga kualitas hidup penduduk (Liwaul, 2023; Simangunsong et al., 2023). Pembangunan ekonomi yang merata menjaga stabilitas sosial dan meningkatkan produktivitas, sedangkan ketimpangan menghambat kemajuan dan memperburuk kondisi sosial (Ningrum et al., 2024). Pemahaman atas kondisi ekonomi antarwilayah karena itu penting untuk merumuskan kebijakan yang efektif.

Di Indonesia, meskipun pertumbuhan ekonomi nasional menunjukkan tren positif, ketimpangan antarprovinsi masih menjadi tantangan (Putri et al., 2023). Rasio gini tercatat 0,381 pada September 2024, meningkat dibandingkan Maret 2024 sebesar 0,379. Beberapa provinsi seperti DKI Jakarta, DI Yogyakarta, dan Papua Selatan bahkan memiliki rasio gini lebih tinggi sehingga ketimpangannya lebih besar. Ketimpangan tersebut berdampak pada kesenjangan akses terhadap sumber daya, pendidikan, kesehatan, dan layanan dasar lainnya (Ambia et.al dalam Purbandono, 2024).

Selain itu, ketimpangan ekonomi juga berkaitan dengan tingkat kemiskinan (Edo & Yasin, 2024). Berdasarkan laporan BPS per September 2024, angka kemiskinan Indonesia tercatat sebesar 8,57%. Namun, dengan ukuran internasional yang ditetapkan Bank Dunia untuk negara menengah ke atas, proporsi penduduk yang tergolong miskin dapat mencapai lebih dari 60%, sehingga sebagian besar masyarakat masih rentan secara ekonomi. Oleh karena itu, diperlukan kebijakan pembangunan yang mendorong pemerataan kesejahteraan dan mengurangi ketimpangan di berbagai daerah (Siregar et al., 2024).

Kebijakan pemerintah berpengaruh besar terhadap pertumbuhan ekonomi dan kesejahteraan masyarakat (Djadjuli, 2018); kebijakan yang tepat mendorong pertumbuhan merata dan memperbaiki ketimpangan, sedangkan kebijakan yang tidak tepat memperburuk keduanya (Hababil et al., 2024). Pemerintah karena itu perlu merumuskan kebijakan berbasis data akurat mengenai kondisi daerah.

Sejalan dengan kebutuhan tersebut, pengelompokan provinsi berdasarkan indikator ekonomi dapat menjadi alternatif untuk memahami kondisi dan karakteristik ekonomi daerah. Dengan mengelompokkan provinsi, pemerintah dapat melihat perbedaan kondisi antarwilayah dan menyesuaikan kebijakan dengan tantangan spesifik setiap daerah (Putri & Suryani dalam Valentina et al., 2025). Pendekatan ini memungkinkan perumusan kebijakan yang lebih terfokus, tepat sasaran, dan efisien, serta membantu memprioritaskan wilayah yang paling membutuhkan perhatian dalam pengurangan ketimpangan dan peningkatan kualitas hidup masyarakat (Putri et al., 2025).

Tingkat perekonomian suatu daerah dapat diukur melalui aspek ekonomi dan aspek sosial ekonomi (Babenko et al., 2021). Aspek ekonomi mencakup PDRB per kapita dan produksi pertanian padi, sedangkan aspek sosial ekonomi meliputi rata-rata upah dan persentase pekerja formal. Dua variabel tambahan yang juga penting adalah persentase kemiskinan dan Indeks Pembangunan Manusia (IPM) (Ersad et al., 2022). Pengelompokan provinsi berdasarkan indikator-indikator tersebut memungkinkan perancangan kebijakan yang lebih spesifik dan tepat sasaran untuk masing-masing wilayah.

Dalam konteks ini, analisis klaster merupakan teknik pengelompokan objek berdasarkan tingkat kemiripan karakteristik, sehingga objek dalam satu klaster memiliki kesamaan satu sama lain dan berbeda karakteristik dari objek pada klaster lain. Dari berbagai pendekatan yang didefinisikan dalam literatur, salah satunya adalah *density-based approach* (pendekatan berbasis kepadatan) (Jahan et al., 2016).

Metode DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) merupakan salah satu metode dengan pendekatan berbasis kepadatan data. Algoritma ini mengkategorikan data berkepadatan tinggi ke dalam klaster dan menandai titik berkepadatan rendah sebagai *noise*. DBSCAN menggunakan dua parameter utama, yaitu *Epsilon* sebagai radius atau ambang jarak maksimum antara dua titik yang dianggap satu klaster dan *MinPTS*

sebagai jumlah minimal titik dalam radius tersebut. Dengan demikian, DBSCAN membentuk klaster dengan anggota sebanyak *MinPTS* dalam radius *Epsilon*.

DBSCAN tidak memerlukan informasi mengenai banyaknya klaster yang akan dibentuk, menyediakan informasi mengenai data *outlier* (*noise*), dan hanya memerlukan dua parameter utama yang tidak sensitif terhadap urutan titik dalam *database*. Namun, DBSCAN terbatas dalam menangani data dengan kepadatan bervariasi dan sensitif terhadap penentuan parameter ambang batas kepadatan klaster (Campello et al., 2020). Untuk mengatasi keterbatasan tersebut, HDBSCAN dikembangkan sebagai metode yang lebih fleksibel: HDBSCAN mengidentifikasi klaster dengan kepadatan data bervariasi melalui pendekatan hierarkis dan lebih *robust* menangani *noise*, sehingga lebih sesuai untuk data ekonomi yang bervariasi dan kompleks serta mungkin mengandung *outlier* (Campello et al., 2020; Handayani & Rosyida, 2023).

Beberapa penelitian menunjukkan bahwa HDBSCAN merupakan pengembangan DBSCAN yang lebih fleksibel dan efektif. Handayani & Rosyida (2023) menunjukkan bahwa HDBSCAN mengelompokkan data lebih baik daripada DBSCAN pada *clustering review* aplikasi Zenius di layanan Google Play Store. Wahyuni et al. (2021) menunjukkan bahwa HDBSCAN mengelompokkan wilayah desa/kelurahan tertinggal di Kabupaten Kutai Kartanegara berdasarkan Data Hasil Pendataan Potensi Desa (PODES) Tahun 2018 dengan baik.

Pengelompokan provinsi di Indonesia telah dilakukan dengan metode lain, antara lain perbandingan *K-Means* dan *Agglomerative Clustering* pada Provinsi Jawa Barat (Valentina et al., 2025), *Agglomerative Hierarchical Clustering* dan *Spectral Clustering* pada seluruh wilayah Indonesia (Putri et al., 2025), serta perbandingan DBSCAN dan *K-Means* pada pemetaan segmentasi layanan kesehatan (Shodikin & Taniasari, 2025). Sepanjang penelusuran literatur penulis, belum ditemukan studi yang menerapkan HDBSCAN untuk mengelompokkan provinsi di Indonesia berdasarkan indikator ekonomi. Pernyataan tersebut merupakan batas cakupan penelusuran, bukan klaim bahwa kajian semacam itu tidak pernah ada.

Dalam proses pengelompokan, perhitungan jarak antar titik data memengaruhi pembentukan klaster. Metrik yang umum digunakan adalah jarak *Euclidean* dan *Manhattan*, keduanya dengan karakteristik berbeda: jarak *Euclidean* sederhana dan banyak digunakan, sedangkan jarak *Manhattan* dapat lebih baik dalam menangani data bervariasi tinggi seperti data ekonomi (Muttaqin et al., 2023). Penelitian ini menerapkan kedua metrik tersebut dalam metode HDBSCAN untuk melihat pengaruhnya terhadap hasil pengelompokan provinsi di Indonesia.

Tujuan penelitian ini dinyatakan sebagai berikut. Pertama, mengelompokkan provinsi di Indonesia berdasarkan data perekonomian Provinsi di Indonesia Tahun 2024, yang meliputi Produksi Pertanian Padi, Produk Domestik Regional Bruto (PDRB) per Kapita, Rata-Rata Upah per Bulan, Persentase Penduduk yang Bekerja di Sektor Formal, Persentase Kemiskinan, dan Indeks Pembangunan Manusia (IPM), dengan algoritma HDBSCAN. Kedua, membandingkan hasil pengelompokan ketika HDBSCAN diterapkan dengan metrik jarak *Euclidean* dan metrik jarak *Manhattan*. Ketiga, menginterpretasikan karakteristik setiap klaster yang terbentuk, termasuk provinsi yang ditandai sebagai *noise*, sebagai dasar penetapan prioritas kebijakan pembangunan daerah.

2. METODE

2.1 Data dan Variabel Penelitian

Penelitian ini menggunakan data sekunder dari publikasi Badan Pusat Statistik (BPS) tahun data 2024: (i) publikasi Produk Domestik Regional Bruto Provinsi-Provinsi di Indonesia untuk PDRB per kapita; (ii) publikasi Statistik Indonesia dan publikasi resmi produksi tanaman pangan untuk produksi padi; (iii) Survei Angkatan Kerja Nasional (Sakernas) untuk rata-rata upah per bulan dan persentase penduduk yang bekerja di sektor formal; (iv) publikasi statistik kemiskinan BPS untuk persentase penduduk miskin; serta (v) publikasi Indeks Pembangunan

Manusia untuk nilai IPM. Seluruh data diakses melalui portal resmi BPS (https://www.bps.go.id) untuk 38 provinsi.

Pada penelitian ini PDRB per kapita didefinisikan PDRB atas dasar harga berlaku dibagi jumlah penduduk provinsi (rupiah per orang per tahun). Produksi padi adalah produksi gabah kering giling provinsi per tahun (ton). Rata-rata upah per bulan adalah rata-rata upah atau gaji bersih pekerja berstatus buruh/karyawan/pegawai dalam satu bulan (rupiah), dihitung hanya untuk penduduk yang bekerja sebagai penerima upah sehingga tidak mencakup pekerja informal, pekerja keluarga, dan pekerja tidak dibayar. Persentase penduduk yang bekerja di sektor formal adalah proporsi penduduk bekerja berstatus buruh/karyawan/pegawai terhadap seluruh penduduk bekerja (persen). Persentase kemiskinan adalah persentase penduduk dengan rata-rata pengeluaran per kapita per bulan di bawah garis kemiskinan provinsi (persen). Indeks Pembangunan Manusia (IPM) adalah indeks komposit dimensi kesehatan, pendidikan, dan standar hidup layak dengan skala 0 sampai 100.

Rasio gini tidak dimasukkan sebagai variabel pengelompokan. Rasio gini mengukur ketimpangan distribusi pengeluaran di dalam suatu populasi provinsi, sedangkan penelitian ini membandingkan tingkat pembangunan ekonomi antarprovinsi; memasukkannya akan mencampur dua tingkat analisis yang berbeda sehingga makna klaster yang terbentuk menjadi kabur. Rasio gini hanya digunakan pada bagian pendahuluan sebagai gambaran awal ketimpangan antarwilayah, bukan sebagai variabel pembentuk klaster.

2.2 Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan analisis klaster berbasis algoritma HDBSCAN. Metode ini bersifat eksploratif dan deskriptif untuk mengidentifikasi kesamaan dan perbedaan karakteristik ekonomi antarprovinsi serta menangani *noise*.

2.3 Tahapan Penelitian

1. Pengumpulan Data Studi Kasus
   Langkah pertama pada penelitian ini dimulai dengan mengumpulkan data setiap variabel yang digunakan pada penelitian untuk mengetahui gambaran data yang akan diteliti
2. Pra-Pemrosesan Data
   a. Pemeriksaan Missing Value
      Tahapan ini mencakup pemeriksaan kelengkapan data dan penanganan nilai yang hilang (*missing value*) melalui imputasi. Berdasarkan pemeriksaan seluruh variabel, tidak ditemukan *missing value*, sehingga seluruh data dinyatakan lengkap dan dapat langsung digunakan pada tahap berikutnya.
   b. Standarisasi Data
      Selanjutnya, data akan distandarisasi menggunakan teknik *Z-score* agar seluruh variabel berada pada skala yang sebanding. Nilai *z-score* dapat dihitung dengan rumus:

$$Z = \frac{X-\mu}{\sigma} \quad (1)$$

      di mana:
      X: Nilai data yang akan dihitung *z-score* nya
      μ: Rata-rata (*mean*) dari data
      σ: Standar deviasi data
3. Pemilihan Metrik Jarak
   Penelitian ini akan menggunakan dua jenis metrik jarak yang umum digunakan, yaitu jarak *Euclidean* dan jarak *Manhattan*. Jarak *Euclidean* dirumuskan sebagai berikut:

$$d(i,j) = \sqrt{\left(\left(x_{i1} - x_{j1}\right)^2 + \cdots + \left(x_{ip} - x_{jp}\right)^2\right)} \quad (2)$$

   Sedangkan jarak *Manhattan* dirumuskan sebagai berikut:

$$d(i,j) = |x_{i1} - x_{j1}| + \cdots + |x_{ip} - x_{jp}| \quad (3)$$

di mana $i = (x_{i1}, \ldots, x_{ip})$ dan $j = (x_{j1}, \ldots, x_{jp})$ merupakan dua objek data yang memiliki p atribut numerik.

4. Penerapan Algoritma HDBSCAN
   a. Implementasi Algoritma
      Selanjutnya, 38 provinsi di Indonesia akan dikelompokkan menggunakan algoritma HDBSCAN. Berikut ini tahapan algoritma HDBSCAN:
   1) Perhitungan *core distance*
      Dengan menetapkan nilai *min_samples* sebesar k, *core distance* titik x_p didefinisikan sebagai jarak dari x_p ke tetangga terdekat ke-k, yaitu $d_{core}(x_p) = d\left(N_k(x_p)\right)$, dengan $N_k(x_p)$ adalah himpunan k tetangga terdekat dari x_p dan d adalah metrik jarak yang dipilih, yaitu Persamaan (2) untuk jarak *Euclidean* atau Persamaan (3) untuk jarak *Manhattan*. Bentuk umum perhitungan *core distance* disajikan pada Persamaan (4).

$$d_{\text{core}}(x_p) = d(N_k(x_p)) \quad (4)$$

   2) Perhitungan *mutual reachability distance* (MRD)
      Jarak antar titik selanjutnya ditransformasikan menjadi *mutual reachability distance* (MRD) dengan mengambil nilai maksimum antara *core distance* kedua titik dan jarak langsung antar titik, sebagaimana disajikan pada Persamaan (5). Transformasi ini membuat dua titik di daerah renggang tetap berjauhan meskipun jarak langsungnya kecil, sedangkan dua titik di daerah padat tetap berdekatan, sehingga titik di tepi daerah padat tidak langsung bergabung dengan klaster yang berdekatan.

$$d_{\text{mreach}}(b) = \{d(b)\} \quad (5)$$

   3) Pembentukan *minimum spanning tree* (MST)
      Berdasarkan nilai MRD, dibentuk graf berbobot lengkap $G = (E)$. Himpunan simpul V berisi 38 provinsi, himpunan sisi E menghubungkan setiap pasangan provinsi, dan bobot setiap sisi (b) adalah $d_{mreach}(b)$ seperti pada Persamaan (5). Karena setiap pasangan provinsi dihubungkan oleh satu sisi, graf memuat $38 \times 37/2 = 703$ sisi. *Minimum spanning tree* (MST) adalah pohon rentang yang menghubungkan seluruh 38 simpul dengan total bobot sisi minimum, sehingga hanya tersusun atas 37 sisi dan tidak memuat siklus. Setiap sisi pada MST merepresentasikan dua provinsi yang paling layak digabungkan pada tingkat kepadatan tertentu, sedangkan bobotnya menyatakan tingkat jarak *mutual reachability* yang harus ditempuh agar keduanya berada dalam satu komponen. Pencarian MST dilakukan dengan algoritma Prim. Algoritma ini dimulai dari satu simpul awal, kemudian pada setiap langkah menambahkan sisi berbobot terkecil yang menghubungkan simpul yang sudah berada di dalam pohon dengan simpul yang belum berada di dalam pohon, sampai seluruh simpul terhubung. Cara kerja tersebut menjamin pohon yang terbentuk selalu terhubung dan tidak memuat siklus. Untuk graf lengkap dengan n simpul, algoritma Prim memiliki kompleksitas waktu $O(n^2)$, sehingga pada 38 provinsi prosesnya berlangsung singkat.
   4) Pembentukan hierarki klaster
      MST yang terbentuk diurutkan sisi-sisinya dari bobot terbesar ke terkecil dan dipotong satu per satu; setiap pemotongan menambah satu komponen terhubung sehingga terbentuk struktur klaster hierarkis yang digambarkan sebagai dendrogram. Struktur tersebut diringkas menjadi *condensed tree* dengan memperhatikan *min_cluster_size*: komponen yang jumlah anggotanya kurang dari

min_cluster_size diperlakukan sebagai bagian klaster induknya dan tidak dianggap klaster yang berdiri sendiri.

5) Seleksi klaster dan penentuan *noise*
Setiap klaster kandidat pada *condensed tree* dievaluasi stabilitasnya dengan konsep *excess of mass*. Stabilitas klaster C_i dihitung dengan Persamaan (6), dengan $\lambda = 1/jarak$, $\lambda_{birth}(C_i)$ adalah nilai λ pada saat C_i pertama kali terbentuk, $\lambda_{max}(C_i)$ adalah nilai λ pada saat titik x keluar dari klaster C_i, dan $|C_i|$ adalah jumlah anggota klaster. Nilai $\lambda_{max}(C_i)$ sama dengan kebalikan jarak *mutual reachability* pada saat x terpisah dari klaster. Klaster yang memiliki nilai stabilitas tinggi dipertahankan, sedangkan klaster yang stabilitasnya rendah atau ukurannya kurang dari *min_cluster_size* dilebur ke klaster induknya. Titik yang pada akhir proses tidak termasuk klaster mana pun diberi label *noise*, yaitu -1. Pustaka hdbscan yang digunakan dalam penelitian ini menerapkan metode pemilihan klaster *excess of mass* secara bawaan (*cluster_selection_method = 'eom'*), sehingga klaster yang dipertahankan adalah klaster dengan stabilitas terbesar sesuai Persamaan (6).

$$S(C_i) = \sum_{x \in C_i} [\lambda_{\max}(x, C_i) - \lambda_{\text{birth}}(C_i)] \quad (6)$$

Tahapan tersebut diimplementasikan menggunakan bahasa pemrograman Python dengan bantuan pustaka *hdbscan* (McInnes et al., 2017). Perhitungan dijalankan dengan scikit-learn 1.6.1; karena scikit-learn menghitung titik itu sendiri dalam *min_samples* sedangkan pustaka *hdbscan* tidak, nilai *min_samples* dinaikkan satu agar setara dengan definisi pada Persamaan (4).

5. Penentuan Parameter
Penentuan parameter *min_cluster_size* dan *min_samples* dilakukan dengan menguji beberapa kombinasi nilai dalam rentang tertentu untuk mencari klaster terbaik, yaitu rentang 1 hingga 7 untuk *min_samples* dan 2 hingga 7 untuk *min_cluster_size*.
6. Penggunaan Metrik Jarak
Proses ini dilakukan dua kali dengan menggunakan metrik *Euclidean* dan *Manhattan* dalam membentuk klaster.
7. Validasi dan Evaluasi Hasil Klaster
Setelah hasil klaster diperoleh, evaluasi dilakukan dengan indeks validasi internal, yaitu *Silhouette Score* (Rousseeuw, 1987) dan *Davies-Bouldin Index* (Davies & Bouldin, 1979). Nilai *Silhouette Score* yang lebih tinggi menunjukkan kualitas klaster yang lebih baik, sedangkan nilai DBI yang lebih rendah mengindikasikan pemisahan klaster yang semakin optimal. Kedua indeks dihitung hanya pada provinsi yang memperoleh label klaster, sedangkan provinsi berlabel *noise* tidak diikutsertakan, sehingga nilainya kondisional terhadap himpunan provinsi yang dipertahankan; setiap nilai karena itu dilaporkan bersama jumlah provinsi yang dievaluasi dan cakupannya, yaitu persentase provinsi berlabel klaster terhadap seluruh 38 provinsi. *Silhouette Score* dihitung dengan jarak yang sama dengan metrik pengelompokan, sedangkan *Davies-Bouldin Index* memakai definisi baku, yaitu jarak *Euclidean* ke pusat klaster. Rumus kedua indeks disajikan sebagai berikut.

$$s(i) = \frac{b(i) - a(i)}{max[a(i), b(i)]} \qquad DB = \frac{1}{c} \sum_{i=1}^{c} max_{i \neq j} \left\{ \frac{d(x_i) + d(x_j)}{d(c_i, c_j)} \right\} \quad (7)$$

di mana:
$a(i)$: rata-rata jarak i terhadap semua objek di klaster tertentu
$b(i)$: rata-rata jarak i terhadap semua objek di klaster lain
c: jumlah klaster

i: label klaster
$d(x_i)$: jarak antara objek pada klaster ke-i terhadap pusat klaster
$d(x_j)$: jarak antara objek pada klaster ke-j terhadap pusat klaster
$d(c_i, c_j)$: jarak antara objek pada klaster i dan j

8. Kriteria Pemilihan Konfigurasi Terbaik
   Karena nilai *Silhouette Score* dan DBI dapat bergerak ke arah berlawanan, kriteria pemilihan konfigurasi terbaik ditetapkan secara eksplisit dan hierarkis. Kriteria pertama adalah cakupan, yaitu proporsi provinsi yang memperoleh label klaster, paling sedikit 75% atau setara dengan paling banyak sembilan provinsi berlabel *noise*; ambang ini dipilih agar perbandingan nilai indeks tidak didominasi oleh sebagian kecil provinsi dan struktur klaster tetap mewakili populasi 38 provinsi. Kriteria kedua, di antara konfigurasi yang memenuhi kriteria pertama dipilih *Silhouette Score* tertinggi; kriteria ketiga, apabila SC sama, dipilih DBI terendah; kriteria keempat, apabila masih sama, dipilih *min_cluster_size* lebih kecil karena menghasilkan pembagian lebih rinci.
   Silhouette Score dijadikan kriteria utama karena dihitung dari perbandingan jarak suatu titik terhadap klaster sendiri dan jarak terhadap klaster lain, sehingga tidak mengasumsikan bentuk klaster tertentu. Sebaliknya, DBI membandingkan sebaran internal klaster dengan jarak antar pusat klaster dan cenderung memberi nilai baik pada klaster kompak berbentuk cembung; pada klaster berbasis kepadatan yang bentuknya tidak beraturan, DBI karena itu hanya berfungsi sebagai penegas.
9. Interpretasi
   Setelah hasil klaster terbentuk, langkah berikutnya adalah menginterpretasi masing-masing klaster berdasarkan karakteristiknya.

3. HASIL DAN PEMBAHASAN

Untuk melakukan pengelompokan dengan metode HDBSCAN, dilakukan uji coba dengan parameter dalam rentang yang telah ditentukan pada dua metrik jarak, yaitu *Euclidean* dan *Manhattan*. Setiap konfigurasi dievaluasi dengan *Silhouette Score* (SC) dan *Davies-Bouldin Index* (DBI), serta dilaporkan jumlah klaster, jumlah *noise*, dan cakupan, yaitu persentase provinsi yang memperoleh label klaster. Konfigurasi terbaik dipilih dengan kriteria hierarkis pada tahap Kriteria Pemilihan Konfigurasi Terbaik, yaitu cakupan minimal 75%, kemudian SC tertinggi, lalu DBI terendah; hasilnya dibahas pada Subbab 3.3.

3.1 Hasil Klasterisasi HDBSCAN dengan Jarak *Euclidean*

Tabel 1. Uji Coba Menggunakan Jarak *Euclidean* dengan *min_samples* = 1

min_cluster_size	Klaster	*Silhouette Score*	*DB Index*	*Noise*
2	9	0,3274	0,7877	7
3	3	0,3678	0,8363	9
4	2	0,3540	0,9605	12
5	2	0,3409	0,9057	20
6	2	0,3409	0,9057	20
7	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38

Berdasarkan Tabel 1 (*min_samples* = 1; *min_cluster_size* 2 hingga 7), hasil terbaik diperoleh pada *min_cluster_size* = 3 dengan SC 0,3678 dan DBI 0,8363, yang membentuk 3 klaster dan 9 *noise*.

Tabel 2. Uji Coba Menggunakan Jarak *Euclidean* dengan *min_samples* = 2

min_cluster_size	Klaster	*Silhouette Score*	*DB Index*	*Noise*
2	3	0,3643	0,6980	12
3	2	0,4443	0,6761	4
4	2	0,3255	0,8037	22
5	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
6	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
7	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38

Berdasarkan Tabel 2 (*min_samples* = 2), hasil terbaik diperoleh pada *min_cluster_size* = 3 dengan SC 0,4443 dan DBI 0,6761, yang membentuk 2 klaster dan 4 *noise*.

Tabel 3. Uji Coba Menggunakan Jarak *Euclidean* dengan *min_samples* = 3

min_cluster_size	Klaster	*Silhouette Score*	*DB Index*	*Noise*
2	2	0,4159	0,7408	28
3	2	0,4159	0,7408	28
4	2	0,4159	0,7408	28
5	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
6	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
7	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38

Berdasarkan Tabel 3 (*min_samples* = 3), hasil terbaik diperoleh pada *min_cluster_size* = 2, 3, dan 4 dengan SC dan DBI yang sama, yaitu 0,4159 dan 0,7408, membentuk 2 klaster dan 28 *noise*.

Tabel 4. Uji Coba Menggunakan Jarak *Euclidean* dengan *min_samples* = 4

min_cluster_size	Klaster	*Silhouette Score*	*DB Index*	*Noise*
2	2	0,3705	0,6922	32
3	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
4	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
5	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
6	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
7	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38

Berdasarkan Tabel 4 (*min_samples* = 4), hasil terbaik diperoleh pada *min_cluster_size* = 2 dengan SC 0,3705 dan DBI 0,6922, yang membentuk 2 klaster dan 32 *noise*.

Pengujian dengan *min_samples* = 5, 6, dan 7 pada metrik *Euclidean*, untuk seluruh *min_cluster_size* dari 2 sampai 7, tidak menghasilkan klaster yang dapat dievaluasi: seluruh 38 provinsi berlabel *noise* sehingga SC dan DBI tidak terdefinisi. Karena seluruh baris bernilai sama, hasilnya dirangkum di sini dan pada Subbab 3.2, bukan disajikan dalam tabel.

Kegagalan HDBSCAN membentuk klaster pada nilai *min_samples* yang besar dapat dijelaskan melalui Persamaan (4) dan Persamaan (5). Nilai *min_samples* menentukan tetangga terdekat ke-k yang dipakai dalam perhitungan *core distance*. Semakin besar k, semakin jauh jarak ke tetangga ke-k sehingga *core distance* seluruh titik meningkat. Peningkatan tersebut menaikkan bobot MRD pada setiap sisi dan membuat seluruh titik tampak berada di daerah berkerapatan rendah. Ketika pohon hierarki diringkas dengan *min_cluster_size*, tidak ada komponen yang memenuhi ukuran minimum sehingga seluruh titik diberi label *noise*. Dengan 38 provinsi, *min_samples* = 5 berarti setiap titik harus memiliki paling sedikit lima tetangga yang berdekatan, suatu syarat yang terlalu ketat untuk sebaran 38 titik pada enam dimensi. Selain itu, *min_samples* yang mendekati atau melampaui *min_cluster_size* membuat klaster sulit memenuhi syarat ukuran minimum.

3.2 Hasil Klasterisasi HDBSCAN dengan Jarak *Manhattan*

Tabel 5. Uji Coba Menggunakan Jarak *Manhattan* dengan *min_samples* = 1

min_cluster_size	Klaster	*Silhouette Score*	*DB Index*	*Noise*
2	2	0,5730	0,5591	0
3	2	0,2637	1,0774	6
4	2	0,2637	1,0774	6
5	2	0,4052	0,7822	21
6	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
7	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38

Berdasarkan Tabel 5 (*min_samples* = 1), hasil terbaik diperoleh pada *min_cluster_size* = 2 dengan SC 0,5730 dan DBI 0,5591, yang membentuk 2 klaster tanpa *noise*.

Tabel 6. Uji Coba Menggunakan Jarak *Manhattan* dengan *min_samples* = 2

min_cluster_size	Klaster	*Silhouette Score*	*DB Index*	*Noise*
2	2	0,2306	0,9092	10
3	2	0,4379	0,7486	23
4	2	0,4379	0,7486	23
5	2	0,4379	0,7486	23
6	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
7	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38

Berdasarkan Tabel 6 (*min_samples* = 2), hasil terbaik diperoleh pada *min_cluster_size* = 3, 4, dan 5 dengan SC dan DBI yang sama, yaitu 0,4379 dan 0,7486, membentuk 2 klaster dan 23 *noise*.

Tabel 7. Uji Coba Menggunakan Jarak *Manhattan* dengan *min_samples* = 3

min_cluster_size	Klaster	*Silhouette Score*	*DB Index*	*Noise*
2	2	0,4851	0,6024	29
3	2	0,4851	0,6024	29
4	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
5	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
6	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
7	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38

Berdasarkan Tabel 7 (*min_samples* = 3), hasil terbaik diperoleh pada *min_cluster_size* = 2 dan 3 dengan SC dan DBI yang sama, yaitu 0,4851 dan 0,6024, membentuk 2 klaster dan 29 *noise*.

Tabel 8. Uji Coba Menggunakan Jarak *Manhattan* dengan *min_samples* = 4

min_cluster_size	Klaster	*Silhouette Score*	*DB Index*	*Noise*
2	3	0,4065	0,8236	29
3	2	0,4284	0,8520	29
4	2	0,4284	0,8520	29
5	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
6	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38
7	-	Tidak bisa dievaluasi	Tidak bisa dievaluasi	38

Berdasarkan Tabel 8 (*min_samples* = 4), hasil terbaik diperoleh pada *min_cluster_size* = 3 dan 4 dengan SC 0,4284 dan DBI 0,8520, membentuk 2 klaster dan 29 *noise*. Pengujian *min_samples* = 5, 6, dan 7 pada metrik *Manhattan* juga hampir seluruhnya tidak menghasilkan klaster yang dapat dievaluasi karena seluruh 38 provinsi berlabel *noise*, kecuali *min_samples* = 5 dengan *min_cluster_size* = 2 yang membentuk dua klaster dengan 31 provinsi berlabel *noise*. Polanya sama dengan metrik *Euclidean* sehingga penjelasannya disatukan pada Subbab 3.1.

3.3 Evaluasi dan Pemilihan Konfigurasi Terbaik

Dari 84 konfigurasi yang diuji, hanya 26 konfigurasi yang menghasilkan nilai SC dan DBI yang dapat dievaluasi. Pada tabel-tabel sebelumnya terlihat bahwa SC dan DBI dihitung pada himpunan provinsi yang berbeda karena jumlah provinsi berlabel *noise* tidak sama antar konfigurasi. Sebagai contoh, pada metrik *Euclidean* konfigurasi *min_samples* = 1 dan *min_cluster_size* = 2 mengevaluasi 31 provinsi dengan 7 provinsi berlabel *noise*, sedangkan konfigurasi *min_samples* = 3 dan *min_cluster_size* = 2 hanya mengevaluasi 10 provinsi dengan 28 provinsi berlabel *noise*; kedua nilai SC berasal dari populasi yang berbeda sehingga tidak dapat dibandingkan secara langsung.

Pemilihan yang semata-mata mengoptimalkan SC tinggi dan DBI rendah juga cenderung memilih konfigurasi yang membuang provinsi dengan karakter berbeda. Provinsi seperti DKI Jakarta dan Kepulauan Riau memang memperburuk nilai indeks apabila diikutsertakan, tetapi pembuangannya menghilangkan informasi yang menjadi salah satu tujuan penelitian ini. Konfigurasi karena itu tidak dipilih berdasarkan nilai indeks saja, melainkan dengan kriteria hierarkis pada tahap Kriteria Pemilihan Konfigurasi Terbaik.

Dengan menerapkan kriteria tersebut, konfigurasi yang memenuhi ambang cakupan minimal 75% disajikan pada Tabel 9.

Tabel 9. Konfigurasi yang Memenuhi Ambang Cakupan Minimal 75% pada Kedua Metrik Jarak

Metrik	*min_samples*	*min_cluster_size*	Klaster	*Noise*	Cakupan (%)	SC	DBI
Euclidean	2	3	2	4	89,47	0,4443	0,6761
Euclidean	1	2	9	7	81,58	0,3274	0,7877
Euclidean	1	3	3	9	76,32	0,3678	0,8363
Manhattan	1	2	2	0	100,00	0,5730	0,5591
Manhattan	1	3	2	6	84,21	0,2637	1,0774
Manhattan	1	4	2	6	84,21	0,2637	1,0774

Pada metrik *Euclidean* terdapat tiga konfigurasi yang memenuhi ambang cakupan, dengan SC tertinggi pada *min_samples* = 2 dan *min_cluster_size* = 3 (SC = 0,4443; DBI = 0,6761; cakupan 89,47%). Pada metrik *Manhattan* juga terdapat tiga konfigurasi yang memenuhi ambang, dengan SC tertinggi pada *min_samples* = 1 dan *min_cluster_size* = 2 (SC = 0,5730; DBI = 0,5591; cakupan 100%). Kedua konfigurasi tersebut sama dengan yang dilaporkan pada Subbab 3.1 dan Subbab 3.2, sehingga penerapan kriteria yang eksplisit tidak mengubah kesimpulan. Hasil pemilihan juga tidak berubah apabila ambang cakupan digeser dalam rentang 70% sampai 85%.

Cakupan yang tinggi tidak dengan sendirinya berarti kualitas klaster lebih baik. Pada metrik *Manhattan*, konfigurasi terpilih menempatkan seluruh 38 provinsi ke dalam dua klaster tanpa *noise*, termasuk DKI Jakarta dan Kepulauan Riau yang karakternya ekstrem; penempatan tersebut perlu ditafsirkan hati-hati karena HDBSCAN pada dasarnya mengasumsikan adanya titik yang tidak termasuk klaster mana pun. Sebaliknya, hasil metrik *Euclidean* mempertahankan empat provinsi sebagai *noise*, dan keempat provinsi tersebut justru menjadi bagian penting dari temuan penelitian ini.

3.4 Karakteristik dan Interpretasi Klaster Terbaik

Klaster optimal yang terbentuk menggunakan metrik jarak *Euclidean* disajikan pada Tabel 10.

Tabel 10. Hasil Klasterisasi Optimal Menggunakan Jarak *Euclidean*

Klaster 1	Klaster 2	*Noise*
Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kepulauan Bangka Belitung, DI Yogyakarta, Banten, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Papua Barat Daya, Papua, Papua Selatan	Jawa Barat, Jawa Tengah, Jawa Timur	Kepulauan Riau, DKI Jakarta, Papua Tengah, Papua Pegunungan

Berdasarkan Table 10, klaster 1 terdiri dari 31 provinsi dengan rata-rata PDRB per kapita Rp76.897.712, produksi pertanian padi 849.851,23 ton, upah per bulan Rp3.138.486,452, persentase pekerja formal 40,86%, persentase kemiskinan 10,09%, dan IPM 72,71. Klaster 2 yang terdiri dari 3 provinsi memiliki rata-rata PDRB per kapita Rp59.940.867, produksi padi 8.929.537,42 ton, upah per bulan Rp2.964.442,67, persentase pekerja formal 41,49%, persentase kemiskinan 8,74%, dan IPM 74,133.

Dua hal perlu ditegaskan dalam menafsirkan klaster 2 pada metrik *Euclidean*. Pertama, klaster tersebut beranggotakan tepat tiga provinsi dengan produksi padi terbesar secara nasional, yaitu Jawa Barat, Jawa Tengah, dan Jawa Timur; pemisahan tersebut terutama mencerminkan skala produksi padi yang jauh di atas provinsi lain, bukan kinerja ekonomi yang lebih baik secara umum, karena rata-rata PDRB per kapita dan upah klaster ini justru lebih rendah daripada klaster 1. Klaster 2 karena itu lebih tepat ditafsirkan sebagai kelompok provinsi dengan produksi padi dominan. Kedua, produksi padi merupakan besaran total yang bergantung pada luas wilayah dan luas lahan pertanian sehingga bersifat *size-dependent*, berbeda dengan kelima variabel lainnya yang dinyatakan per satuan atau proporsi; standardisasi *z-score* menyetarakan skala antar variabel tetapi tidak menghilangkan

ketergantungan tersebut. Pengujian sensitivitas dengan mengeluarkan indikator tersebut disajikan pada Subbab 3.6.

Keempat provinsi yang diklasifikasikan sebagai *noise* menunjukkan karakteristik ekonomi yang ekstrem. Kepulauan Riau dan DKI Jakarta memiliki karakteristik ekonomi ekstrem tinggi (sangat maju): PDRB per kapita, rata-rata upah per bulan, persentase penduduk yang bekerja di sektor formal, dan IPM jauh lebih tinggi, serta persentase kemiskinan jauh lebih rendah dibandingkan kedua klaster. Adapun Papua Tengah dan Papua Pegunungan memiliki karakteristik ekonomi ekstrem rendah (tertinggal): produksi pertanian padi, persentase pekerja formal, dan IPM jauh lebih rendah dibanding kedua klaster meski upah per bulan dan PDRB per kapita bervariasi. Kombinasi variabel tersebut membentuk pola unik yang tidak sesuai dengan klaster mana pun, sehingga pemerintah dapat memprioritaskan kebijakan pembangunan ekonomi pada kedua provinsi tersebut untuk mengurangi kesenjangan pembangunan.Klaster optimal menggunakan metrik jarak *Manhattan* disajikan pada Tabel 11.

Tabel 11. Hasil Klasterisasi Optimal Menggunakan Jarak *Manhattan*

Klaster 1	Klaster 2	*Noise*
Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, DKI Jakarta, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, Banten, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Papua Barat Daya, Papua, Papua Selatan	Papua Tengah, Papua Pegunungan	Tidak Ada

Berdasarkan Tabel 11, klaster 1 terdiri dari 36 provinsi dengan rata-rata PDRB per kapita Rp85.262.736,7, produksi pertanian padi 1.476.017 ton, upah per bulan Rp3.247.344, persentase pekerja formal 42,32%, persentase kemiskinan 9,67%, dan IPM 73,27. Klaster 2 yang terdiri dari 2 provinsi memiliki rata-rata PDRB per kapita Rp68.439.455, produksi padi 3.057,38 ton, upah per bulan Rp4.580.517, persentase pekerja formal 8,61%, persentase kemiskinan 28,63%, dan IPM 56,59.

Perbandingan karakteristik antar klaster menunjukkan bahwa klaster 1 memiliki nilai rata-rata PDRB per kapita, produksi pertanian padi, persentase penduduk yang bekerja di sektor formal, serta IPM yang lebih tinggi dibandingkan klaster 2. Sedangkan klaster 2 memiliki rata-rata upah per bulan dan persentase kemiskinan yang lebih tinggi dibandingkan klaster 1. Hasil ini menunjukkan bahwa klaster 1 sebagai klaster besar yang terbentuk memiliki indikator ekonomi yang lebih baik dari klaster 2. Temuan ini dapat menjadi dasar bagi pemerintah untuk memprioritaskan provinsi pada klaster 2 sebagai sasaran utama dalam perencanaan dan evaluasi kebijakan pembangunan.

Arah indikator upah pada klaster 2 metrik *Manhattan* perlu dijelaskan. Rata-rata upah per bulan dihitung dari upah atau gaji bersih pekerja berstatus buruh/karyawan/pegawai, sehingga nilainya hanya mencakup penduduk yang bekerja sebagai penerima upah dan tidak memasukkan pekerja informal, pekerja keluarga, maupun pekerja tidak dibayar. Pada Papua Tengah dan Papua Pegunungan, proporsi penduduk bekerja yang berstatus formal hanya 8,61%, sehingga rata-rata upah yang relatif tinggi (Rp4.580.517) mencerminkan komposisi sektor formal yang sempit tetapi bergaji relatif tinggi, bukan tingginya kesejahteraan penduduk secara umum. Kedua provinsi tersebut tetap memiliki persentase kemiskinan tertinggi (28,63%) dan IPM terendah (56,59) sehingga penafsirannya sebagai wilayah dengan karakteristik ekonomi rendah tetap konsisten.

3.5 Perbandingan Metrik *Euclidean* dan *Manhattan*

Kedua metrik menghasilkan jumlah klaster yang sama pada konfigurasi terbaik masing-masing, yaitu dua klaster, tetapi komposisi anggotanya berbeda. Tabulasi silang keanggotaan klaster pada kedua metrik disajikan pada Tabel 12.

Tabel 12. Tabulasi Silang Keanggotaan Klaster Hasil Metrik *Euclidean* dan Metrik *Manhattan*

Klaster *Euclidean*	Klaster 1 (*Manhattan*)	Klaster 2 (*Manhattan*)	Jumlah
Klaster 1	31	0	31
Klaster 2	3	0	3
Noise	2	2	4
Jumlah	36	2	38

Tabel 12 menunjukkan bahwa partisi *Euclidean* merupakan penghalusan dari partisi *Manhattan*. Seluruh 31 provinsi klaster 1 dan 3 provinsi klaster 2 pada metrik *Euclidean* bergabung menjadi satu klaster pada metrik *Manhattan*, sedangkan dua dari empat provinsi *noise Euclidean*, yaitu Papua Tengah dan Papua Pegunungan, membentuk klaster tersendiri. Tingkat kesesuaian kedua partisi diukur dengan *Adjusted Rand Index* (ARI) sebesar 0,35 apabila label *noise* diperlakukan sebagai kelompok tersendiri, tergolong rendah karena kedua partisi memiliki jumlah klaster yang berbeda.

Silhouette Score dihitung dari perbandingan jarak intraklaster dan jarak ke klaster terdekat, sedangkan DBI dari perbandingan sebaran internal klaster dan jarak antar pusat klaster; keduanya tanpa satuan sehingga berada pada skala yang sama. Namun, *Silhouette Score* dihitung dengan jarak yang sama dengan metrik pengelompokan (*Euclidean* atau *Manhattan*), sedangkan *Davies-Bouldin Index* memakai definisi baku, yaitu jarak *Euclidean* ke pusat klaster, sehingga nilainya menggambarkan kualitas pemisahan relatif di dalam geometri metrik yang dievaluasi dan bukan hasil pengukuran pada satu skala jarak yang sama. Nilai SC dan DBI karena itu diperlakukan sebagai indeks deskriptif, bukan uji statistik formal yang membandingkan dua populasi.

Dengan kriteria tersebut, konfigurasi terbaik pada metrik *Manhattan* memiliki SC lebih tinggi (0,5730) dan DBI lebih rendah (0,5591) dibandingkan konfigurasi terbaik metrik *Euclidean* (0,4443 dan 0,6761), sehingga secara numerik klasterisasi dengan metrik *Manhattan* memberikan pemisahan klaster yang lebih baik. Namun, keunggulan tersebut diperoleh dengan memasukkan seluruh 38 provinsi ke dalam klaster, termasuk dua provinsi berkarakteristik ekstrem, sedangkan metrik *Euclidean* memisahkan provinsi dengan produksi padi dominan dan menandai empat provinsi ekstrem sebagai *noise*. Metrik *Manhattan* karena itu lebih unggul apabila ukuran keberhasilan adalah kepadatan dan keterpisahan klaster, sedangkan metrik *Euclidean* lebih informatif untuk mengidentifikasi provinsi yang menyimpang dari pola umum.

Perbedaan struktur kedua partisi dijelaskan oleh cara kedua metrik menjumlahkan selisih antar variabel. Jarak *Euclidean* mengkuadratkan selisih setiap variabel sehingga satu variabel dengan selisih ekstrem mendominasi nilai jarak, sedangkan jarak *Manhattan* menjumlahkan nilai mutlak selisih sehingga pengaruh variabel ekstrem lebih teredam. Produksi padi merupakan variabel dengan sebaran paling timpang: rata-rata klaster 2 pada metrik *Euclidean* mencapai 8.929.537 ton atau sekitar sepuluh kali rata-rata klaster 1 sebesar 849.851 ton. Pada metrik *Euclidean* selisih tersebut cukup untuk memisahkan ketiga provinsi penghasil padi utama menjadi klaster tersendiri, sedangkan pada metrik *Manhattan* tidak cukup untuk melawan akumulasi selisih kelima variabel lainnya.

3.6 Analisis Stabilitas dan Sensitivitas

Ketidakstabilan jumlah klaster pada Tabel 1 sampai Tabel 8 sebagian besar merupakan akibat perubahan *min_samples* dan *min_cluster_size*, bukan perubahan metrik. Pada *min_samples* = 1, *core distance* suatu titik adalah jarak ke tetangga terdekatnya sehingga nilainya sangat kecil; akibatnya MRD mendekati jarak asli dan HDBSCAN membentuk banyak klaster kecil, yaitu sembilan klaster. Ketika *min_cluster_size* dinaikkan menjadi 3, klaster kecil tersebut tidak lagi memenuhi ukuran minimum sehingga digabungkan dan jumlah klaster

turun menjadi dua atau tiga. Setelah *min_samples* mencapai 5, tidak ada klaster yang terbentuk. Rentang hasil yang lebar tersebut mencerminkan kepekaan HDBSCAN terhadap kedua parameternya pada data berukuran kecil.

Struktur yang bertahan pada seluruh konfigurasi yang dapat dievaluasi adalah terpisahnya dua provinsi Papua, yaitu Papua Tengah dan Papua Pegunungan, dari klaster besar: keduanya berlabel *noise* pada 24 dari 26 konfigurasi dan membentuk klaster tersendiri pada dua konfigurasi lainnya, yaitu *Euclidean* dan *Manhattan* dengan *min_samples* = 1 dan *min_cluster_size* = 2. Kedua provinsi tersebut juga memiliki persentase kemiskinan tertinggi dan IPM terendah. Sebaliknya, pemisahan provinsi penghasil padi utama hanya muncul pada metrik *Euclidean* dan hilang ketika metrik diganti, sehingga tidak dapat dikatakan stabil.

Stabilitas klaster juga diuji melalui *resampling bootstrap* dengan 500 replikasi: 38 provinsi diambil secara acak dengan pengembalian, proses klasterisasi diulang pada konfigurasi terbaik masing-masing metrik, lalu pemulihan setiap klaster diukur dengan *Jaccard index* antara keanggotaan klaster pada model dasar dan klaster yang paling mirip pada setiap replikasi. Karena observasi hanya 38 dan *resampling* dengan pengembalian menimbulkan titik kembar, hasilnya sangat bergantung pada cara menangani duplikat dan cara mencocokkan klaster replikasi dengan klaster dasar; pada pengujian ini nilai *Jaccard* berubah cukup besar di antara kedua cara tersebut sehingga satu angka tunggal tidak dapat dilaporkan sebagai ukuran yang mantap, dan hasil tersebut hanya menunjukkan bahwa stabilitas klaster pada data sebesar ini terbatas.

Analisis sensitivitas berikutnya ditujukan pada indikator produksi padi. Sesuai penjelasan pada Subbab 2.1 bahwa produksi padi merupakan besaran total, analisis diulang dengan mengeluarkan produksi padi sehingga hanya lima indikator yang bersifat per satuan atau proporsi yang digunakan. Hasilnya disajikan pada Tabel 13.

Tabel 13. Perbandingan Struktur Klaster pada Model Enam Indikator dan Varian tanpa Indikator Produksi Padi

Varian indikator	Metrik	Para-meter	Klas-ter	*Noise*	Cakupan (%)	SC	DBI	Anggota klaster kecil
Enam indikator (model utama)	*Euclidean*	ms = 2; mcs = 3	2	4	89,47	0,4443	0,6761	Jawa Barat, Jawa Tengah, Jawa Timur (3 provinsi)
Lima indikator (tanpa produksi padi)	*Euclidean*	ms = 1; mcs = 3	2	8	78,95	0,3843	0,9328	Papua Barat, Papua Barat Daya, Papua, Papua Selatan (4 provinsi)
Lima indikator (tanpa produksi padi)	*Manhattan*	ms = 1; mcs = 3	2	5	86,84	0,3359	0,9780	Papua Barat, Papua Barat Daya, Papua, Papua Selatan (4 provinsi)

Pada varian tanpa produksi padi, konfigurasi terbaik tetap membentuk dua klaster, tetapi komposisinya berubah: klaster kecil tidak lagi berisi tiga provinsi penghasil padi utama, melainkan empat provinsi di wilayah Papua, yaitu Papua Barat, Papua Barat Daya, Papua, dan Papua Selatan, sedangkan Jawa Barat, Jawa Tengah, dan Jawa Timur bergabung kembali ke klaster besar. Perubahan tersebut menunjukkan bahwa pemisahan tiga provinsi penghasil padi utama pada model enam indikator ditentukan oleh variabel produksi padi dan bukan oleh perbedaan kinerja ekonomi secara umum. Sebaliknya, kelompok provinsi dengan karakteristik pembangunan rendah tetap bertahan meskipun indikator produksi padi dikeluarkan, sehingga temuan tersebut lebih kokoh.

Nilai SC dan DBI pada Tabel 13 tidak dapat dibandingkan antarvarian karena ruang indikator yang digunakan berbeda; yang dapat dibandingkan adalah struktur klaster yang terbentuk. Membagi produksi padi dengan total produksi nasional juga tidak akan mengubah

hasil, karena pembagian dengan suatu konstanta saling menghapus pada proses standardisasi *z-score*. Normalisasi yang benar-benar berpengaruh adalah normalisasi terhadap ukuran provinsi, seperti jumlah penduduk atau luas lahan pertanian, yang datanya tidak tersedia dalam penelitian ini.

4. KESIMPULAN

Penelitian ini mengelompokkan 38 provinsi di Indonesia berdasarkan enam indikator ekonomi tahun 2024 menggunakan metode HDBSCAN dengan dua metrik jarak. Konfigurasi terbaik pada metrik Euclidean adalah min_samples = 2 dan min_cluster_size = 3 dengan SC = 0,4443, DBI = 0,6761, dan cakupan 89,47%, yang membentuk dua klaster dan menandai empat provinsi sebagai noise. Konfigurasi terbaik pada metrik Manhattan adalah min_samples = 1 dan min_cluster_size = 2 dengan SC = 0,5730, DBI = 0,5591, dan cakupan 100%, yang membentuk dua klaster tanpa noise.

Kedua metrik menghasilkan struktur yang konsisten pada tingkat kasar, yaitu satu kelompok besar provinsi dan satu kelompok provinsi dengan karakteristik ekonomi rendah, dan partisi Euclidean merupakan penghalusan dari partisi Manhattan. Secara numerik, metrik Manhattan memberikan pemisahan klaster yang lebih baik berdasarkan Silhouette Score dan Davies-Bouldin Index, sedangkan metrik Euclidean memberikan pembagian lebih rinci dan menandai provinsi berkarakteristik ekstrem sebagai noise. Analisis sensitivitas menunjukkan bahwa pemisahan tiga provinsi penghasil padi utama ditentukan oleh indikator produksi padi, sedangkan kelompok provinsi dengan karakteristik pembangunan rendah tetap bertahan.

Papua Tengah dan Papua Pegunungan secara konsisten teridentifikasi sebagai wilayah dengan karakteristik ekonomi paling rendah pada kedua metrik, dengan persentase kemiskinan tertinggi dan IPM terendah, sehingga perlu menjadi prioritas dalam perencanaan kebijakan pembangunan guna mengurangi kesenjangan antarwilayah.

Penelitian ini memiliki beberapa keterbatasan. Pertama, analisis hanya menggunakan data satu tahun sehingga perubahan karakteristik ekonomi antarwaktu tidak dapat ditangkap. Kedua, observasi terbatas pada 38 provinsi sehingga hasil HDBSCAN sangat bergantung pada nilai min_samples dan min_cluster_size; uji stabilitas bootstrap belum menghasilkan ukuran kuantitatif yang mantap. Ketiga, indikator yang digunakan mencampur variabel per satuan dengan variabel total; pengujian tanpa produksi padi menunjukkan pemisahan provinsi penghasil padi utama hilang ketika indikator tersebut dikeluarkan, sehingga normalisasi terhadap jumlah penduduk atau luas lahan pertanian masih diperlukan. Keempat, Silhouette Score dan Davies-Bouldin Index dihitung hanya pada provinsi berlabel klaster sehingga nilainya bergantung pada cakupan evaluasi. Kelima, rasio gini tidak dimasukkan karena mengukur ketimpangan di dalam suatu provinsi, bukan ketimpangan antarprovinsi.

Berdasarkan keterbatasan tersebut, penelitian lanjutan disarankan untuk (i) menormalkan produksi padi terhadap jumlah penduduk atau luas lahan pertanian, atau menggantinya dengan indikator yang tidak bergantung pada ukuran wilayah, agar klaster yang terbentuk mencerminkan kinerja ekonomi; (ii) menambahkan periode pengamatan agar stabilitas klaster dapat diuji antarwaktu dan melalui resampling; (iii) membandingkan HDBSCAN dengan metode klasterisasi lain, seperti K-Means, Agglomerative Hierarchical Clustering, dan Spectral Clustering, dengan indeks validasi yang sesuai; serta (iv) menguji penambahan indikator distribusi seperti rasio gini sebagai variabel pelengkap dalam analisis terpisah.

5. DAFTAR PUSTAKA

Babenko, V., Panchyshyn, A., Zomchak, L., Nehrey, M., Artym-Drohomyretska, Z., & Lahotskyi, T. (2021). Classical machine learning methods in economics research: Macro and micro level examples. *Wseas Transactions on Business and Economics*, *18*, 209–217. https://doi.org/10.37394/23207.2021.18.22

Campello, R. J. G. B., Kröger, P., Sander, J., & Zimek, A. (2020). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(2). https://doi.org/10.1002/widm.30

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Advances in Knowledge Discovery and Data Mining (PAKDD 2013), Lecture Notes in Computer Science, 7819, 160–172. Springer. https://doi.org/10.1007/978-3-642-37456-2_14

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Djadjuli, R. D. (2018). Peran pemerintah dalam pembangunan ekonomi daerah. *Dinamika: Jurnal Ilmiah Ilmu Administrasi Negara*, 5(2), 8–21. http://dx.doi.org/10.25157/dinamika.v5i2.1409

Edo, A., & Yasin, M. (2024). Dampak kesenjangan akses pendidikan dan faktor ekonomi keluarga terhadap mobilitas sosial. *SINOVA: Jurnal Ilmu Pendidikan & Sosial*, *2*(3), 317–326. https://doi.org/10.71382/sinova.v2i3.175

Ersad, M., Amri, A., & Zulgani. (2022). Dampak IPM, tingkat pengangguran dan tingkat kemiskinan terhadap ketimpangan pendapatan di Sumatera Bagian Selatan. *Jurnal Paradigma Ekonomika*, *17*(2), 2684–7868. https://doi.org/10.22437/jpe.v17i2.15614

Hababil, M. P., Firdaus, M. K., Nazhmi, N., Alghifary, M. R., Hamdani, M. D., & Fadilla, A. (2024). Analisis pengaruh pemerataan ekonomi dalam upaya menghapus ketimpangan sosial-ekonomi antar masyarakat. *Journal of Macroeconomics and Social Development*, *1*(4), 1–9. https://doi.org/10.47134/jmsd.v1i4.276

Handayani, F. D., & Rosyida, I. (2023). Clustering review pengguna aplikasi Zenius pada layanan Google Play Store menggunakan metode DBSCAN dan HDBSCAN. *Emerging Statistics and Data Science Journal*, *1*(2), 178-191. https://doi.org/10.20885/esds.vol1.iss.2.art19

Jahan, R., Javed, N., & Ahmad, S. F. (2016). Cluster analysis techniques in data mining: A review. *IJSRD-International Journal for Scientific Research & Development*, *4*(2), 1088-1091. https://doi.org/10.1145/331499.331504

Liwaul. (2023). Pertumbuhan ekonomi kerakyatan melalui kewirausahaan lokal untuk meningkatkan kesejahteraan masyarakat. *Jurnal Kesejahteraan Dan Pelayanan Sosial*, *4*(2), 104–123. https://doi.org/10.52423/jkps.v4i2.15

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. Journal of Open Source Software, 2(11), 205. https://doi.org/10.21105/joss.00205

Muttaqin, Z., Fernando, D., & Sulastriani, S. (2023). Implementasi unsupervised learning pada nilai jasmani kesamaptaan sekolah polisi negara dengan metode clustering analysis. *Jurnal PROSISKO*, *10*(1), 18–23. https://doi.org/10.30656/prosisko.v10i1.6269

Ningrum, E. P., M, S., Nursyamsi, S. E., & Siregar, N. (2024). Faktor terkait kesenjangan ekonomi dan kesejahteraan. *Jurnal Riset-Akuntansi Dan Keuangan*, *7*(2), 116–126. https://doi.org/10.36815/prive.v7i2.3480

Purbandono, R. (2024). Dampak ketimpangan ekonomi terhadap stabilitas sosial: Analisis peran kebijakan pemerintah dalam mengurangi kesenjangan ekonomi. *MULTIPLE: Journal of Global and Multidisciplinary*, *2*(12), 3869–3883.

Putri, D. F., Andriani, Sari, W. R., & Nabbila, F. L. (2023). Problematika ketimpangan ekonomi dalam sistem ekonomi di Indonesia. *Jurnal Ilmiah Ekonomi Dan Manajemen*, *1*(4), 163–172. https://doi.org/10.61722/jiem.v1i4.330

Putri, S. A., Putri, T. R., Trimono, & Idhom, M. (2025). Segmentasi wilayah Indonesia berdasarkan IHK menggunakan AHC dan spectral clustering. *Jurnal Teknik Informatika Dan Multimedia*, *5*(1), 65–76. https://doi.org/10.51903/informatika.v5i1.1010

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Shodikin, M., & Taniasari, N. (2025). Analisis komparasi algoritma DBSCAN dan K-Means dalam pemetaan segmentasi pasien rawat inap menggunakan model RFMT. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(4), 1231-1241. https://doi.org/10.57152/malcom.v5i4.2218

Simangunsong, N. A., Wardani, D. A., Reksapramudya, A., Arrahman, M. I., Wulandari, S., & Islam, P. E. (2023). Peran pertumbuhan ekonomi dalam menunjang kesejahteraan masyarakat. *Jurnal Kajian Ekonomi & Bisnis Islam*, *4*(5), 1289–1298. https://doi.org/10.47467/elmal.v4i5.2808

Siregar, A. R. S., Lubis, N. H., Isfa, F., & Muzayyana. (2024). Pentingnya pemerataan pembangunan ekonomi dalam meningkatkan kesejahteraan masyarakat Indonesia. CEMERLANG: Jurnal Manajemen Dan Ekonomi Bisnis, 5(1), 51–64. https://doi.org/10.55606/cemerlang.v5i1.3424

Valentina, T., Sabela, S. N., War, S. S. M., & Saputra, W. S. J. (2025). Segmentasi Provinsi Jawa Barat berdasarkan indikator sosial-ekonomi menggunakan metode K-Means dan Agglomerative Clustering. RAGAM: Journal of Statistics and Its Application, 4(1), 50–60. https://doi.org/10.20527/ragam.v4i1.15589

Wahyuni, N. A., Hayati, M. N., & Rizki, N. 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). *Jurnal EKSPONENSIAL*, *12*(1), 47-52. https://doi.org/10.30872/eksponensial.v12i1.758