J. Ris. & Ap. Mat. Vol. 08 No. 02 (2024) pp. 122-135
Jurnal Riset dan Aplikasi Matematika (JRAM)
e-ISSN: 2581-0154
URL: journal.unesa.ac.id/index.php/jram

PERBANDINGAN ALGORITMA HDBSCAN DAN *AGGLOMERATIVE HIERARCHICAL CLUSTERING* DALAM KLASTERISASI PADA DATA YANG MENGANDUNG PENCILAN

GHARDAPATY GHALY GHIFFARY[1*], KEVIN ALIFVIANSYAH[2], ANWAR FITRIANTO[3], ERFIANI[4], L.M. RISMAN DWI JUMANSYAH[5].

[1,2,3,4,5]Departemen Statistika dan Sains Data, IPB University

*penulis korespondensi: ghardapatyghiffary@apps.ipb.ac.id

ABSTRAK

Penelitian ini 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 yang mengandung pencilan. Data tersebut meliputi 17 variabel dengan 234 observasi. Metode *ward*, rata-rata dan lengkap hanya mampu menghasilkan gerombol tanpa penciri khas untuk amatan pencilan berbeda dengan metode HDBSCAN. Hasil evaluasi model HDBSCAN memiliki nilai terbaik pada semua ukuran evaluasi dan mampu mendeteksi pencilan secara otomatis. Sebaliknya, metode *ward*, rata-rata dan lengkap belum mampu menangani pencilan secara khusus. Berdasarkan hal tersebut, disimpulkan bahwa untuk data berpencilan, HDBSCAN lebih optimal dibanding metode hierarki agglomerative lainnya karena mampu mengelompokkan data secara konsisten serta mendeteksi dan mengelola pencilan secara efektif.
Kata Kunci: HDBSCAN, Gerombol, PDRB, Pencilan

ABSTRACT

This study compares the performance of the HDBSCAN algorithm with three agglomerative hierarchical methods Ward, average, and complete linkage in clustering Gross Regional Domestic Product at Constant Prices (PDRB ADHK) data from East Java Province in 2023, which contains outliers. The dataset comprises 17 variables and 234 observations. The Ward, average, and complete linkage methods were only able to form clusters without distinguishing outlier observations, in contrast to the HDBSCAN method. Evaluation results showed that HDBSCAN performed best across all evaluation metrics and was able to automatically detect outliers. Conversely, the Ward, average, and complete linkage methods were unable to specifically handle outliers. Based on these findings, it is concluded that for data containing outliers, HDBSCAN is more optimal than other agglomerative hierarchical methods, as it consistently clusters data while effectively detecting and managing outliers.
Keywords: *Clustering, GDRB, HDBSCAN, Outliers*

1 Pendahuluan

Analisis gerombol merupakan teknik multivariat pada *data mining* yang digunakan dalam mengelompokkan objek-objek pada data berdasarkan kesamaan karakteristik yang dimilikinya [1]. Objek akan dikelompokkan ke dalam satu atau lebih gerombol sehingga objek-objek yang

2020 *Mathematics Subject Classification*: 91C20
Diterima: 28-09-24; direvisi: 22-10-24; diterima: 27-10-24

berada dalam kesatuan gerombol akan memiliki kesamaan yang tinggi antara satu dengan lainnya. Salah satu metode analisis gerombol yang biasa digunakan salah satunya adalah *Agglomerative Hierarchical Clustering* (AHC).

Agglomerative Hierarchical Clustering (AHC) teknik gerombol yang bertujuan untuk mengelompokkan data secara bertahap berdasarkan kedekatan atau kesamaan antar data. Proses AHC dimulai dengan menganggap setiap objek sebagai satu kelompok tersendiri, kemudian secara bertahap menggabungkan pasangan kelompok terdekat hingga semua objek berada dalam satu kelompok besar atau memenuhi kriteria penghentian tertentu [2]. Salah satu tantangan pada AHC adalah diperlukan penanganan pada data yang terdapat pencilan.

Pencilan muncul ketika terdapat nilai ekstrim yang berada jauh dari sebagian besar nilai dalam sampel atau populasi data [2]. Pencilan juga menjadi indikasi suatu masalah atau anomali dalam data karena keberadaannya dapat mengganggu proses analisis data. Deteksi pada data pencilan merupakan persoalan penting dalam metode AHC seperti pautan lengkap metode yang menggunakan prinsip jarak terkecil yang diawali dengan mencari jarak terjauh antar dua gerombol dan keduanya membentuk gerombol baru [2] dan pautan rataan metode yang diperoleh dengan menghitung rata-rata jarak seluruh objek suatu gerombol terhadap seluruh objek pada gerombol lainnya [3]. Pautan rataan memperlakukan jarak antara dua kelompok sebagai jarak rata-rata antara semua pasangan item di mana satu anggota pasangan menjadi milik masing-masing kelompok. Kedua metode tersebut menjadi pendekatan yang biasa digunakan dalam penanganan data pencilan. Penelitian yang dilakukan Yusoff dkk. [4], menunjukkan pautan lengkap dan pautan rataan efektif dalam mendeteksi banyak pencilan dalam model *circular regression*. Pada tahun 2017 penelitian yang dilakukan oleh Zulkarnain dkk. [5], menjelaskan mengenai metode pautan rataan dalam mengelompokkan sekolah menengah pertama (SMP) berdasarkan rata-rata nilai ujian nasional, dan mendeteksi pencilan menggunakan *mahalanobis distance* untuk meningkatkan pengelompokan. Namun metode pautan lengkap dan pautan rataan memiliki kelemahan dalam menangani data pencilan. Pautan lengkap dapat menggabungkan gerombol yang tidak sesuai karena pencilan menarik gerombol lain [3], sedangkan pautan rataan dipengaruhi oleh pencilan yang mengaburkan struktur gerombol [4].

HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) merupakan metode analisis gerombol yang diperkenalkan sebagai alternatif dalam menangani data pencilan. Sebagai modifikasi dari algoritma OPTICS, HDBSCAN menggunakan parameter input tunggal, yaitu Minpts (*Minimum Points*) [6]. Algoritma ini tidak hanya mengidentifikasi gerombol dengan kepadatan tinggi, tetapi juga secara otomatis mendeteksi pencilan sebagai data yang tidak termasuk dalam gerombol. HDBSCAN menggunakan metrik *mutual reachability distance* untuk mengidentifikasi titik data berdasarkan kepadatan dan membangun struktur gerombol hirarki. Titik yang tidak memenuhi ambang batas kepadatan akan ditandai sebagai pencilan, dan jika gerombol terputus dengan jumlah titik di bawah batas minimum, titik tersebut dianggap *noise*. Selain itu, HDBSCAN menghitung stabilitas gerombol dan skor pencilan (GLOSH) untuk menilai seberapa jauh titik dari gerombol. Penelitian oleh Handayani dkk. [6], menemukan bahwa HDBSCAN lebih unggul dalam mengelompokkan ulasan pengguna aplikasi *e-learning*, dengan nilai koefisien *silhouette* rata-rata 0,2941, lebih tinggi dibandingkan DBSCAN yang menunjukkan bahwa klaster yang dihasilkan oleh HDBSCAN memiliki jarak intra-klaster yang lebih kecil dan jarak antar-klaster yang lebih besar dibandingkan DBSCAN, sehingga HDBSCAN memiliki keunggulan dalam mengidentifikasi struktur klaster yang lebih jelas dan memisahkan data secara lebih akurat. HDBSCAN juga memberikan keefektifan pada penanganan pencilan pada analisis gerombol pada data *mixed-mail* [7].

Pencilan sering kali muncul dalam analisis PDRB (Produk Domestik Regional Bruto) provinsi Jawa Timur [8], terutama dalam kategori ADHK (Atas Dasar Harga Konstan) 2023.

Pencilan ini dapat terjadi akibat fluktuasi ekonomi yang tidak biasa, perubahan kebijakan, atau bahkan kesalahan pengukuran [9]. Dalam menangani data pencilan yakni dengan mengidentifikasi dan mengeliminasi pengaruh data yang tidak representatif atau mengandung *noise* sehingga analisis yang dilakukan dapat lebih akurat dan mencerminkan kondisi ekonomi yang sebenarnya. Dalam upaya ini, digunakan metode *Agglomerative Hierarchical Clustering* dengan pendekatan pautan lengkap, pautan rataan, dan pautan ward. HDBSCAN metode hierarki berbasis densitas digunakan untuk membandingkan metode *Agglomerative Hierarchical Clustering* yang diharapkan dapat mengidentifikasi metode yang paling efektif dalam mendeteksi pencilan ekstrim di data PDRB Jawa Timur ADHK 2023, sehingga analisis yang dihasilkan lebih akurat dan representatif. Tujuan penelitian ini adalah untuk mengevaluasi dan menentukan metode gerombol yang paling optimal dalam mengidentifikasi data pencilan pada PDRB ADHK Jawa Timur 2023, guna memberikan hasil pengelompokan terbaik dalam mengatasi permasalahan pencilan pada data ekonomi daerah.

2 Tinjauan Pustaka

2.1 Analisis Gerombol

Analisis gerombol merupakan salah satu pendekatan pada analisis multivariat yang bertujuan untuk mengelompokkan data berdasarkan karakteristik data yang sama kedalam sebuah gerombol. Tujuan dari analisis gerombol adalah meminimalkan jarak di dalam gerombol dan memaksimalkan jarak antar gerombol [1]. Secara umum terdapat dua metode dalam pembentukan gerombolnya, yakni metode *hierarchical* dan *non-hierarchical*.

2.2 Metode *Agglomerative Hierarchical Clustering* (AHC)

AHC (*Agglomerative Hierarchical Clustering*) merupakan teknik untuk melakukan eksplorasi data analisis dengan cara mengelompokkan data berdasarkan karakteristik yang sesuai kemudian menjadi kumpulan kelompok yang disebut gerombol. AHC membentuk hirarki sehingga membentuk struktur pohon. Proses yang digunakan dalam pengelompokannya yaitu secara bertingkat atau bertahap. Pada penelitian ini akan menggunakan tiga metode *agglomerative.*

2.2.1 Pautan *Ward*

Metode *ward* merupakan teknik yang digunakan untuk membentuk gerombol yang didasari pada hilangnya informasi akibat pengelompokkan objek menjadi gerombol. Pada prosesnya dihitung dengan menggunakan jumlah total dari deviasi kuadrat pada rata-rata gerombol untuk setiap amatan dengan menggunakan *error sum of squares* (SSE) sebagai fungsi objektif berdasarkan persamaan (1) :

$$\text{SSE} = \sum_{j=1}^{p} \left(\sum_{i=1}^{n} X_{ij}^2 - \frac{1}{n} \left(\sum_{i=1}^{n} X_{ij}^2 \right)^2 \right) \tag{1}$$

X_{ij} merupakan nilai untuk objek ke-i pada gerombol ke-j, sedangkan p adalah banyaknya peubah yang diukur dan n adalah banyaknya objek dalam gerombol yang terbentuk.

2.2.2 Pautan Rataan

Metode *Average linkage* atau pautan rataan menghitung jarak antara dua gerombol sebagai jarak rata-rata yang jarak tersebut dihitung pada masing-masing gerombol berdasarkan persamaan (2) :

$$d_{(AB)X} == \frac{\sum_i \sum_k d_{ik}}{N_{(AB)} N_X} \tag{2}$$

d_{ik} merupakan jarak antara objek ke-i pada gerombol (AB) dan objek ke-k dalam gerombol X. Sedangkan $N_{(AB)}$ dan N_X berturut-turut merupakan banyaknya objek dalam gerombol (AB) dan (X).

2.2.1 Pautan Lengkap

Pada metode ini, *pautan lengkap*atau pautan lengkap didasarkan dalam menemukan jarak *euclidean* maksimum. Jarak antara satu gerombol dan gerombol yang lain diukur berdasarkan objek yang memiliki jarak terjauh. Langkah awalnya mencari nilai minimum dalam $D = \{d_{ik}\}$ dan menggabungkan objek yang bersesuaian. Misalnya A dan B untuk menjadikannya ke dalam gerombol (AB). Menentukan jarak antara (AB) dan gerombol X yang lain dihitung berdasarkan konsep jarak *euclidian* seperti pada persamaan 3.

$$d_{(AB)X} = maks\,\{d_{AX}, d_{BX}\} \tag{3}$$

dengan d_{AX} merupakan jarak terpendek antara gerombol-gerombol A dan X, sementara d_{BX} merupakan jarak terpendek antara gerombol-gerombol B dan X

2.3 Standarisasi Data

Proses standarisasi dilakukan jika pada setiap peubah pada penelitian masih terdapat perbedaan satuan dalam pengukuran yang signifikan. Standarisasi menjadi proses yang penting jika titik data diukur menggunakan skala yang berbeda kemungkinan besar tidak akan memberikan kontribusi yang baik dalam analisis sehingga hasil pada analisis gerombol tidak akurat dan valid. Standarisasi dilakukan dengan mentransformasikan data asli ke dalam satuan yang seragam berdasarkan persamaan 4.

$$X_{scale} = \frac{x - \bar{x}}{simpangan\ baku(x)} \tag{4}$$

x_{scale} merupakan nilai peubah yang sudah melalui proses standarisasi, x adalah peubah sebelum dilakukan standarisasi. Sedangkan $\bar{x}$ merupakan nilai rata-rata pada x dan $simpangan\ baku(x)$ merupakan nilai standar deviasi pada peubah x

2.4 *Principal Component Analysis* (PCA)

PCA adalah teknik reduksi data multivariat yang mentransformasikan satu matriks data menjadi satu kombinasi garis yang lebih sedikit, dengan mempertahankan sebagian besar data asli [8]. Tujuan utama PCA adalah untuk menjelaskan hubungan antara jumlah titik data mentah dan jumlah komponen utama yang dikenal sebagai faktor. Ada banyak faktor (komponen) yang dapat diekstraksi dari kumpulan data awal. Selain itu, PCA adalah teknik statistik yang bertujuan untuk meningkatkan akurasi dari sebuah variabel dengan cara mereduksi dimensinya. Metode ini bekerja dengan menghitung matriks kovarian dari data dan kemudian mencari vektor *eigen* dan nilai *eigen*.

2.5 Pencilan (*Outliers*)

Pencilan merupakan data yang terletak jauh atau terpencil dari data lainnya dalam suatu populasi. Keberadaan pencilan dapat menyebabkan distribusi data menjadi tidak normal, menghasilkan bias dalam penaksiran parameter, dan mempengaruhi hasil analisis data. Meskipun pencilan sering dianggap sebagai noise dan memiliki perilaku berbeda dari mayoritas data lainnya, akan tetapi pencilan seringkali mengandung informasi yang sangat berharga. Tidak semua data dengan pencilan dapat ditransformasi dengan cara yang sama, karena setiap kasus data memiliki karakteristik yang berbeda.

2.6 HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*)

HDBSCAN merupakan konsep dan algoritma modifikasi dari algoritma *ordering points to identify the clustering structure* (OPTICS) dengan parameter tunggal yang disebut M_{pts}[10]. HDBSCAN menghasilkan struktur gerombol hirarki yang dapat digunakan untuk tugas pasca pemrosesan, seperti mendeteksi dan menyederhanakan identifikasi pencilan. metode ini dapat mengisolasi pencilan secara lebih efisien dan menyoroti pola-pola signifikan dalam data.
Tahapan pada HDBSCAN adalah sebagai berikut:

1. Menginisialisasi parameter $M_{pts} = 2,3,4,5$ dan 6
2. Tentukan titik x_p yang akan menjadi *core point* secara acak
3. Menghitung semua jarak titik terhadap *core point* pada persamaan (5)

$$d_{p,q} = \| x_p - x_q \| = \sqrt{\sum_{k=1}^{n} \left| x_{pk} x_{qk} \right|^2} \tag{5}$$

4. Mentransformasi ruang fitur untuk semua titik data dengan metrik *density-reachable* terhadap x_p yang berada pada $d_{core}(x_p)$. Jika banyak titik yang memiliki hubungan yang kuat terhadap x_p lebih dari atau sama dengan banyaknya M_{pts}, maka titik x_p menjadi *core point* sehingga gerombol terbentuk dan dilanjutkan ke titik lainnya. Jika titik yang berhubungan kuat terhadap x_p, maka gerombol yang terbentuk merupakan gerombol yang sama dengan gerombol bagi *core point* sebelumnya.
5. Menghitung nilai *Silhouette coefficient* dari setiap kombinasi parameter M_{pts} berdasarkan persamaan (6)

$$SC = \frac{\sum_{i=0}^{C} n_i SC_2(p)}{\sum_{i=0}^{C} n_i} \tag{6}$$

6. Mengulangi langkah 2 hingga 4 untuk setiap titik-titik yang belum membentuk gerombol.
7. Mengekstraksi gerombol akhir. Setiap titik data kemudian diberi label sesuai dengan gerombol yang terbentuk. Beberapa titik mungkin tidak termasuk dalam klaster mana pun, diklasifikasikan sebagai “noise”, dan biasanya diberi label -1.

2.7 *Silhouette Score*

Metode yang digunakan untuk mengevaluasi dalam melihat kualitas dan kekuatan gerombol adalah *silhouette score* (SC). Metode validasi gerombol ini menggabungkan metode *cohesion* dan *separation*. Tahapan perhitungannya adalah sebagai berikut.

1. Menghitung jarak rata-rata dari suatu data ke-i dengan semua data yang berada dalam gerombol yang sama dengan persamaan (7)

$$a_i = \frac{1}{n_k - 1} \sum_{r=1}^{n_k - 1} d(x_i, x_r), r \neq i \tag{7}$$

   Dengan $k = 1,2,3 \ldots, K$

2. Menghitung jarak rata-rata suatu data ke-i dengan semua data yang berada pada gerombol yang berbeda dengan menggunakan persamaan (8)

$$d_i(k) = \frac{1}{n_k} \sum_{r=1}^{n_k} d(x_i, x_r) \tag{8}$$

Kemudian diambil nilai terkecilnya dengan persamaan (9)

$$b_i = \{d_i(k)\}, r \neq i \tag{9}$$

3. Menghitung nilai SC untuk setiap data ke-i

$$SC_1(i) = \frac{b_i - a_i}{max\{a_i, b_i\}}, i = 1,2, \dots, n \tag{10}$$

Nilai *silhouette score* dari sebuah gerombol $SC_2(k)$ diperoleh dengan menghitung rata-rata nilai $SC_1(i)$ pada semua data yang tergabung dalam gerombol tersebut dengan menggunakan persamaan (11)

$$SC_2(k) = \frac{1}{n_k}\sum_{r=1}^{n_k} SC_1(i) \tag{11}$$

Kemudian menentukan nilai *silhouette score* keseluruhan diperoleh dari menghitung rata-rata nilai $SC_2(k)$ dari semua gerombol dengan persamaan (12).

$$SC = \frac{\sum_{k=1}^{K}(n_k \times SC_2(k))}{\sum_{k=1}^{K} n_k} \tag{12}$$

Hasil perhitungan nilai *silhouette score* dapat bervariasi antara -1 sampai dengan 1, jika nilai $SC = 1$, memberikan informasi bahwa titik data x_i sudah berada dalam gerombol yang tepat. Jika nilai $SC = 0$, maka titik data x_i berada di antara dua gerombol. Jika $SC = -1$ memberikan informasi struktur gerombol yang dihasilkan oleh objek x_i lebih tepat dimasukkan dalam gerombol yang lain.

3 Hasil dan Pembahasan

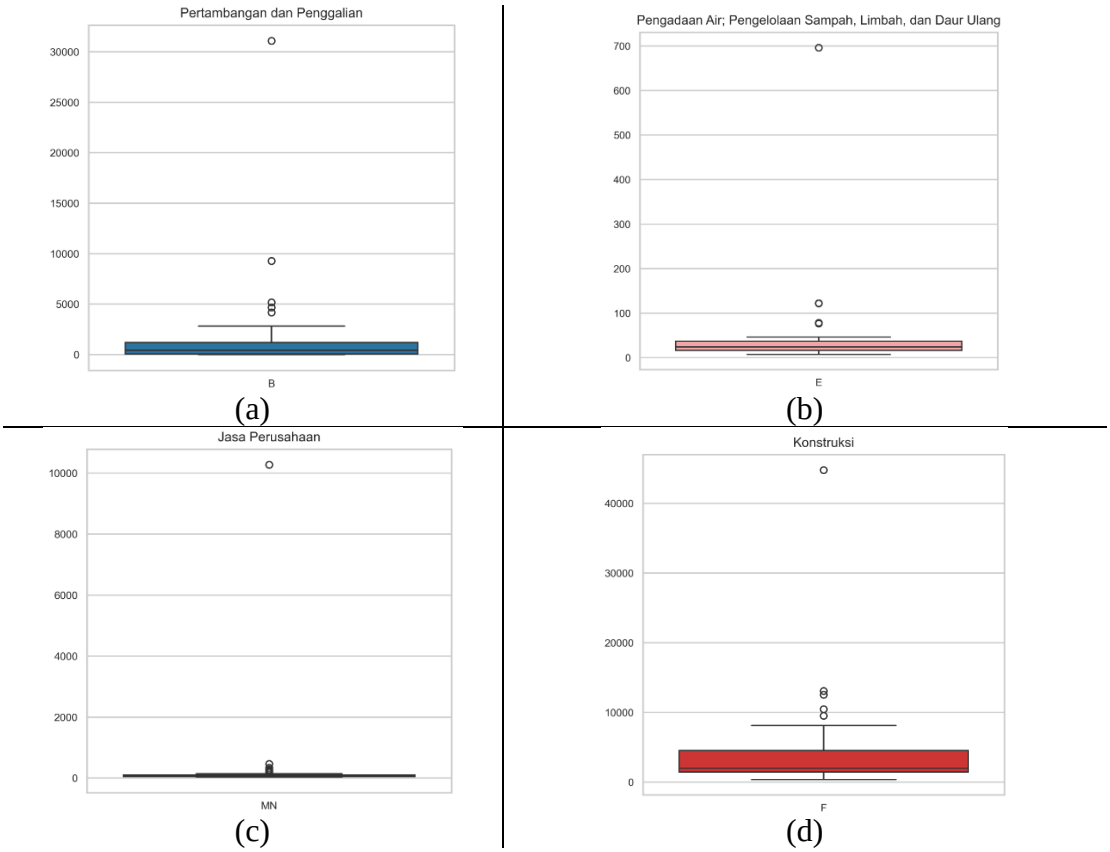
3.1 Sumber Data

Penelitian ini menggunakan data BPS (Badan Pusat Statistik) Provinsi Jawa Timur yang diperoleh dari website BPS. Dataset tersebut berisi Produk Domestik Regional Bruto Atas Dasar Harga Konstan (PDRB ADHK) pada tahun 2023.

3.2 Eksplorasi Data

3.2.1 Visualisasi Peubah

Eksplorasi data merupakan suatu metode yang digunakan untuk melihat gambaran umum pada data. Terdapat 17 peubah yang digunakan dalam Penelitian ini dengan sebaran peubah merupakan numerik. Melalui analisis statistika deskriptif dan visualisasi pada data, terdapat amatan yang dikatakan merupakan pencilan ekstrim seperti contohnya peubah Lapangan Usaha Sektor C (Industri Pengolahan). Tetapi ada juga peubah yang memiliki sebaran data cukup simetris sehingga nilai informasi dan sebaran pada data dinilai bukan suatu permasalahan yang harus diatasi. Berikut merupakan visualisasi data untuk peubah yang memiliki nilai amatan pencilan ekstrem.

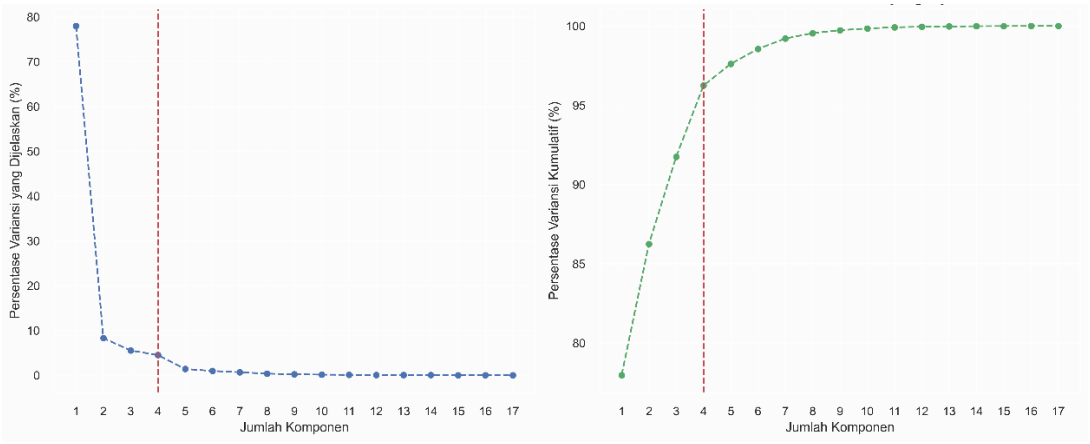


Gambar 1. Peubah Sektor B (a); Sektor E (b); Sektor MN (c) dan; Sektor F (d)

Berdasarkan **Gambar.1** ditemukan bahwa pada peubah yang memiliki pencilan ekstrem, sebaran data yang dimiliki cenderung tidak normal dan terdapat amatan yang terlampau jauh dari rentang normal data. Secara keseluruhan, amatan pencilan ekstrim ditemukan pada 15 dari 17 peubah berdasarkan pola visualisasi *boxplot*. Pencilan ekstrim dapat mempengaruhi nilai kualitas dan performa suatu pengelompokan sehingga membutuhkan algoritma yang bersifat *robust* serta penanganan lebih lanjut.

3.2.2 Reduksi Peubah

Principal component analysis (PCA) digunakan untuk mereduksi dimensi data yang tinggi menjadi dimensi yang rendah dengan mempertahankan informasi peubah asal pada ragamnya. PCA juga dapat mengatasi masalah multikolinearitas antar peubah. Pada dataset ini terdapat multikolinearitas pada beberapa peubah sehingga perlu dilakukan reduksi data menggunakan PCA. Jumlah komponen dipilih berdasarkan rasio ragam dan persentase kumulatif yang dijelaskan guna memastikan bahwa sebagian besar ragam dapat dijelaskan dengan komponen hasil reduksi.



(a) (b)

Gambar 2. *Scree plot* (a); Persentase Kumulatif Ragam (b)

Berdasarkan **Gambar 2**. terdapat dua grafik yang digunakan untuk menentukan jumlah komponen optimal dalam reduksi dimensi data. *Scree plot* pada **Gambar 2(a).** menunjukkan bahwa penurunan persentase varians yang dijelaskan signifikan hingga komponen ke-4, sedangkan setelahnya penurunan menjadi lebih lambat, menandakan titik siku (*elbow point*) pada komponen ke-2. **Gambar 2(b)**. menampilkan persentase ragam kumulatif yang menunjukkan bahwa 4 komponen pertama telah cukup menjelaskan informasi dari total ragam. Analisis lebih lanjut ditampilkan dalam bentuk table sebagai penjelas visualisasi dan nilai keputusannya. Berikut merupakan tabel hasil PCA dalam menentukan komponen terbentuk.

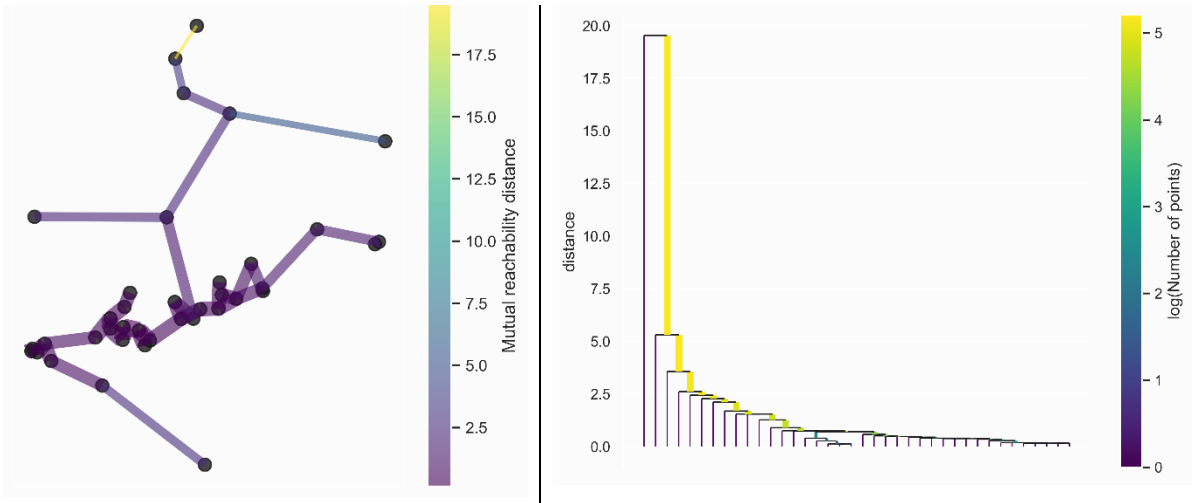
Tabel 1. Ragam pada PCA

Komponen	Inisiasi Eigen		
	Ragam	Persentase Ragam	Kumulatif (%)
1	13.61	77.97	77.97
2	1.44	8.27	86.19
3	0.96	5.51	91.7
4	0.78	4.48	96.15
5	0.23	1.37	97.55

Berdasarkan **Tabel 1.** setelah pembentukan 4 komponen, peningkatan ragam yang dapat dijelaskan peubah baru hanya meningkat 1.37% dan tidak cukup berpengaruh dalam memberikan informasi tambahan. Sehingga dengan kumulatif persentase ragam yang dapat dipertahankan sebesar 96.15% dapat membentuk 4 komponen utama yang baru.

3.3 HDBSCAN

HDBSCAN merupakan suatu algoritma yang dapat mengelompokkan data berdasarkan kepadatan dan bersifat memiliki suatu tingkatan (hierarki). Metode ini juga bersifat *robust* yang artinya tahan terhadap pencilan. Menggunakan jarak mahalanobis guna menempatkan amatan yang merupakan pencilan ekstrim dapat ditangani. Nilai parameter optimum yang digunakan pada data ini adalah *minimum cluster size* = 5, *minimum sample* = 2, *number cluster* = 3, *silhouette coefficient* sebesar 0.1995. Menggunakan nilai parameter tersebut menunjukkan bagaimana kelompok - kelompok terhubung secara optimal dengan mempertimbangkan kepadatan dan jarak antar titik. Visualisasi antar kelompok yang terhubung divisualisasikan menggunakan pohon rentang minimum dan *dendrogram*.

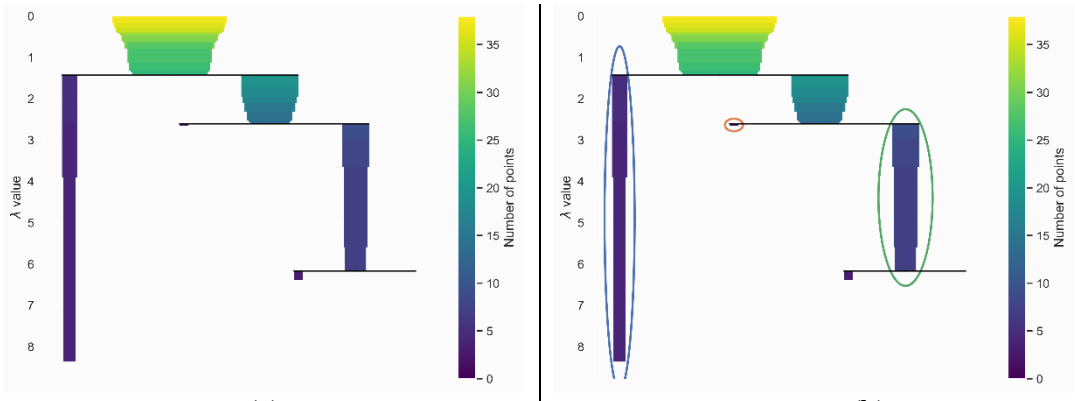


(a) (b)

(a) (b)

Gambar 3. Pohon Rentang Minimum (a) dan Dendrogram HDBSCAN

Pohon rentang minimum dan *dendrogram* memberikan representasi mengenai hubungan hierarkis antar kelompok data. Temuan ini memiliki implikasi penting, terutama dalam menangani data dengan distribusi yang tidak teratur serta kehadiran pencilan ekstrim, seperti yang terlihat pada data PDRB Lapangan Usaha Provinsi Jawa Timur. Struktur hierarki pengelompokan dapat divisualisasikan dalam bentuk *dendrogram*, yang menghasilkan tiga kelompok utama. Kelompok - kelompok ini diidentifikasi melalui *condensed tree*, yang merepresentasikan kepadatan hierarki dalam pohon rentang minimum dan dibentuk berdasarkan kelompok dengan kepadatan minimum.



(a) (b)

Gambar 4. *Condensed Tree* (a) dan Kepadatan Kelompok Terbentuk (b)

Terdapat 2 sumbu y dimana y_1 menunjukkan nilai λ dari 0 sampai 8 merupakan Tingkat hirarki kelompok dimana semakin mendekati 0 maka akan semakin dalam kelompok tersebut pada Tingkat hirarki kelompok sedangkan nilai y_2 menunjukkan titik amatannya pada **Gambar 4(a).** Menggunakan HDSCAN didapatkan tiga kelompok utama yang terbentuk. Pada **Gambar 4(b)**. kelompok-kelompok ini ditandai dengan lingkaran berwarna berbeda: kelompok 0 (ditandai dengan lingkaran berwarna biru), kelompok 1 (ditandai dengan lingkaran berwarna oranye), dan kelompok 2 (ditandai dengan lingkaran berwarna hijau). Selain itu, terdapat juga kelompok -1, yang mengindikasikan *noise* atau data yang tidak cukup padat untuk dimasukkan ke dalam kelompok mana pun.

Profilisasi dari tiap kelompok untuk jumlah amatan (anggota) dari tiap kelompok dapat dilihat dari skala warna pada batang di bawah *condensed tree*. Berdasarkan legenda warna,

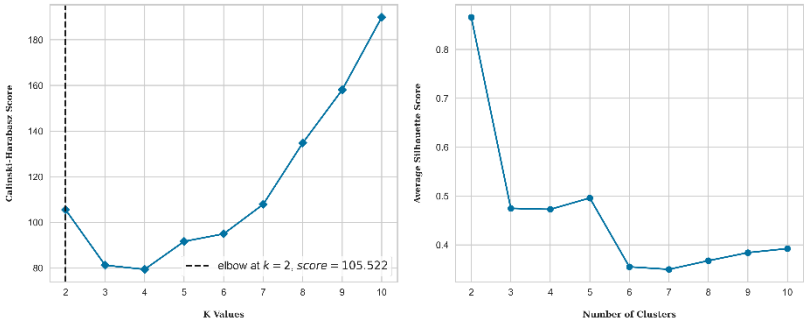
kelompok 0 memiliki sekitar 5 anggota, kelompok 1 memiliki 3 anggota, dan kelompok 2 memiliki 9 anggota. Sementara itu, kelompok -1 terdiri dari 21 anggota yang dianggap sebagai pencilan atau *noise*. Hal ini disebabkan oleh kepadatan data yang terlalu rendah atau berbeda jauh dari kelompok lain.

3.4 Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering merupakan suatu metode pengelompokkan hirarki dimana titik tiap amatan dianggap sebagai sebuah kelompok lalu akan digabungkan secara bertahap berdasarkan kemiripan dan kesamaan amatan tersebut. Penghitungan jarak antar kelompok akan menggunakan metode Pautan *ward*, rataan dan lengkap dimana jarak yang digunakan adalah *euclidean*.

3.4.1 Pautan *Ward*

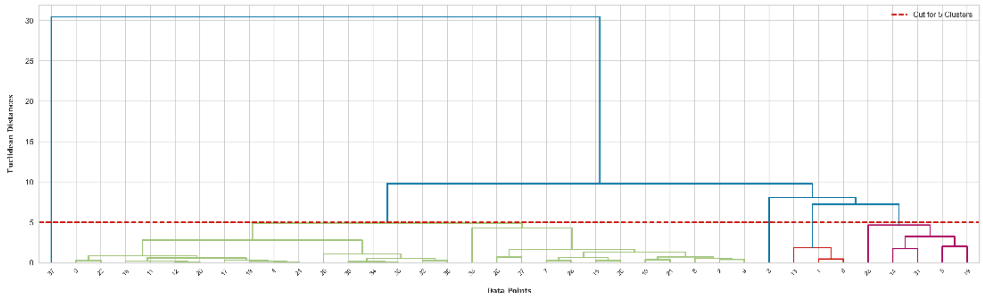
Pautan *ward* menggunakan konsep meminimalkan jumlah ragam yang ada pada kelompok, dimana tiap kelompok akan menghasilkan nilai ragam terkecil. Berikut adalah visualisasi *dendrogram* menggunakan pautan *ward*.



(a) (b)

Gambar 5. *k*-optimum Callinski-Harabasz (a); *k*-optimum Silhouette Score (b) – Pautan *Ward*

Nilai *k*-optimum untuk pautan *ward* yang ditunjukkan **Gambar 5.** menggunakan nilai *Callinski-Harabasz* dan *silhouette* adalah $k = 2$. Menggunakan nilai $k = 2$, mampu untuk membagi pengelompokkan Kabupaten/kota di Provinsi Jawa Timur berdasarkan PDRD Sektor Lapangan Usaha, tetapi terdapat nilai amatan yang berbeda dan bertimpangan pada kota Surabaya sehingga nilai k untuk pengelompokkan dibagi menjadi $k = 5$ berdasarkan nilai terbaik kedua pada *Callinski-Harrabasz* dan *silhouette*. Adapun hasil pengelompokkan menggunakan pautan *ward* ditampilkan sebagai berikut:



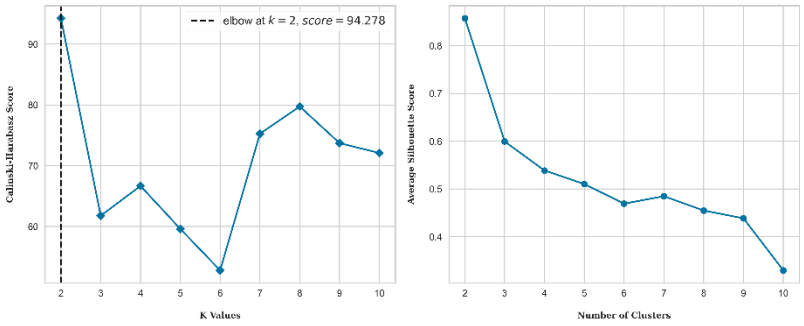
Gambar 6. *Dendrogram* Pautan *Ward*

Dendrogram pautan *ward* pada **Gambar 6.** menggambarkan pengelompokan data yang menghasilkan 5 *k* (kelompok). Penggerombolan pada kelompok ini meminimalkan ragam antar kelompok, sehingga menghasilkan pembagian yang seimbang. Sebagian besar amatan tergabung dalam 1 kelompok besar, sementara beberapa kelompok lebih kecil terbentuk pada jarak penggabungan yang lebih tinggi, menunjukkan adanya ragam atau pencilan dalam

amatan. Profilisasi jumlah anggota tiap kelompok adalah sebagai berikut: kelompok 1 memiliki 28 anggota, kelompok 2 memiliki 3 anggota, kelompok 3 memiliki 5 anggota, kelompok 4 memiliki 1 anggota, dan kelompok 5 memiliki 1 anggota.

3.4.2 Pautan Rataan

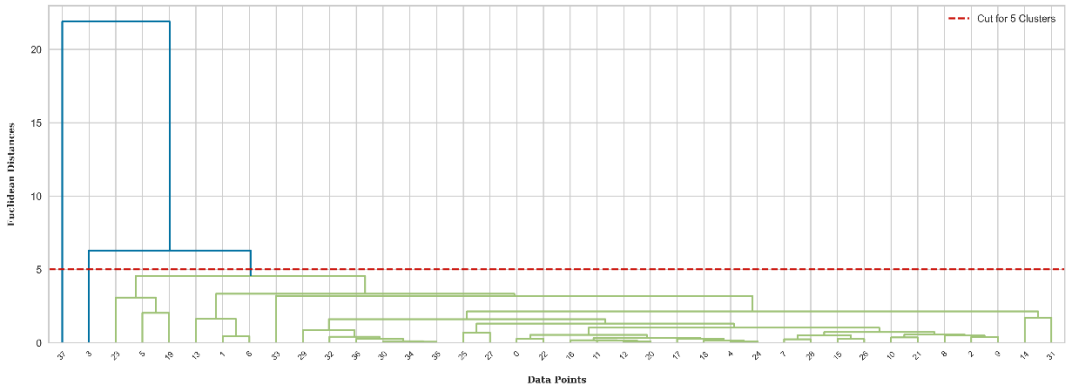
Pautan rataan menggunakan konsep menghitung jarak antar 2 kelompok sebagai rata-rata jarak antar semua kombinasi amatan pada kelompok. Berikut adalah visualisasi *dendrogram* menggunakan pautan rataan.



(a) (b)

Gambar 7. *k*-optimum Callinski-Harabasz (a); *k*-optimum Silhouette Score (b) – Pautan Rataan

Berdasarkan **Gambar 7**, nilai *k*-optimum yang ditunjukkan oleh pautan rataan menggunakan nilai *Callinski-Harrabasz* dan *silhouette* adalah $k = 2$. Menggunakan nilai $k = 2$, mampu untuk membagi pengelompokkan Kabupaten/kota di Provinsi Jawa Timur berdasarkan PDRD Sektor Lapangan Usaha, tetapi terdapat nilai amatan yang berbeda dan bertimpangan pada Kota Surabaya sehingga nilai *k* untuk pengelompokkan dibagi menjadi $k = 4$ berdasarkan nilai terbaik kedua pada *Callinski-Harrabasz* dan *silhouette*. Adapun hasil pengelompokkan menggunakan pautan *ward* ditampilkan sebagai berikut:



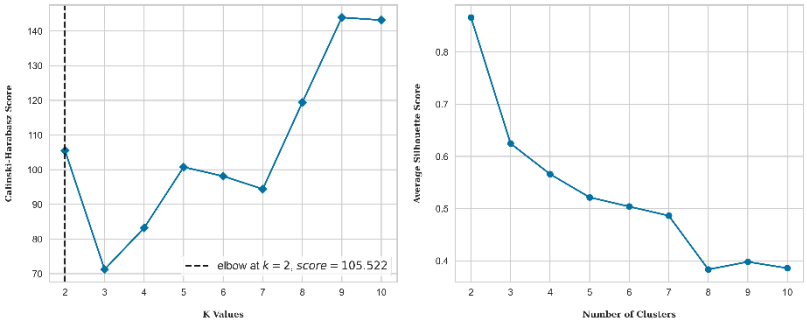
Gambar 8. *Dendrogram* Pautan Rataan

Dendrogram pautan rataan pada **Gambar 8.** menggambarkan pengelompokan data yang menghasilkan 4 *k* (kelompok). Pengelompokan ini dilakukan dengan cara menggabungkan kelompok berdasarkan rata-rata jarak antar amatan. Sebagian besar amatan tergabung dalam 1 kelompok besar, sementara beberapa kelompok lebih kecil terbentuk pada jarak penggabungan yang lebih tinggi, menunjukkan adanya ragam atau pencilan dalam amatan. Profilisasi jumlah anggota tiap kelompok adalah sebagai berikut: kelompok 1 memiliki 3 anggota, kelompok 2 memiliki 33 anggota, kelompok 3 memiliki 1 anggota, dan kelompok 4 memiliki 1 anggota. Hasil penggerombolan ini menunjukkan adanya struktur hierarki yang beragam dalam data,

dengan beberapa amatan yang hanya membentuk 1 kelompok karena nilai amatannya merupakan pencilan.

3.4.3 Pautan Lengkap

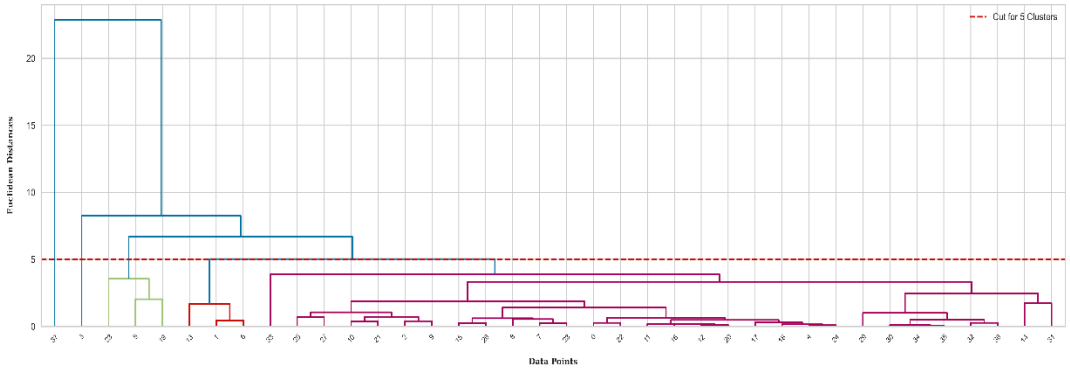
Pautan *lengkap* menggunakan konsep menghitung jarak antar 2 kelompok berdsaarkan jarak antara 2 titik amatan terjauh. Hal ini menunjukkan kelompok yang terbentuk memiliki jarak terdekat terkecil. Berikut adalah visualisasi dendrogram menggunakan pautan *lengkap*.



(a) (b)

Gambar 9. *k*-optimum Callinski-Harabasz (a); *k*-optimum Silhouette Score (b) – Pautan Lengkap

Berdasarkan **Gambar 9.** nilai *k*-optimum yang ditunjukkan oleh pautan lengkap menggunakan nilai *Callinski-Harrabasz* dan *silhouette* adalah k = 2. Menggunakan nilai *k* = 2, mampu untuk membagi pengelompokkan Kabupaten/kota di Provinsi Jawa Timur berdasarkan PDRD Sektor Lapangan Usaha, tetapi terdapat nilai amatan yang berbeda dan bertimpangan pada Kota Surabaya sehingga nilai k untuk pengelompokkan dibagi menjadi *k* = 5 berdasarkan nilai terbaik kedua pada *Callinski-Harrabasz* dan *silhouette*. Adapun hasil pengelompokkan menggunakan pautan *ward* ditampilkan sebagai berikut:



Gambar 10. *Dendrogram* Pautan Lengkap

Dendrogram pautan lengkap pada **Gambar 10.** menggambarkan pengelompokan data yang menghasilkan 5 *k* (kelompok). Pengelompokkan data dilakukan dengan menggabungkan kelompok berdasarkan jarak maksimum antar amatan, sehingga penggabungan dilakukan dengan mempertimbangkan amatan terjauh antara dua kelompok. Sebagian besar amatan tergabung dalam 1 kelompok besar, sementara beberapa kelompok lebih kecil terbentuk pada jarak penggabungan yang lebih tinggi, menunjukkan adanya ragam atau pencilan dalam amatan. Profilisasi jumlah anggota tiap kelompok adalah sebagai berikut: kelompok 1 memiliki 3 anggota, kelompok 2 memiliki 3 anggota, kelompok 3 memiliki 30 anggota, kelompok 4 memiliki 1 anggota, dan kelompok 5 memiliki 1 anggota. Hasil pengelompokan menunjukkan terdapat keragaman antar kelompok dan amatan yang sulit untuk dikelompokkan bersama

amatan lain. Temuan ini mengindikasikan bahwa amatan tersebut memiliki perbedaan yang signifikan dalam hal jarak maksimum antar amatan dengan kelompok lainnya.

3.5 Evaluasi dan Kualitas Kelompok

Evaluasi kualitas kelompok dalam penelitian ini dilakukan menggunakan metrik *Silhouette Score*, *Davies-Bouldin Index*, *Calinski-Harabasz Index*, *Dunn Index*, dan *Within-Cluster Sum of Squares* (WCSS). *Silhouette Score* digunakan untuk mengukur kedekatan objek dengan kelompok dibandingkan dengan kelompok lainnya. *Davies-Bouldin Index* bertujuan mengevaluasi rasio antara jarak dalam kelompok dan jarak antar kelompok, dengan nilai yang lebih rendah menandakan hasil yang lebih baik. *Calinski-Harabasz Index* untuk mengukur ragam antar-dalam kelompok (nilai lebih tinggi menunjukkan pemisahan kelompok yang baik). *Dunn Index* menilai jarak terdekat antar kelompok dibandingkan diameter kelompok terbesar, dan WCSS menghitung total ragam dalam klaster untuk mengukur kompaksi. Metrik ini digunakan untuk menilai validitas hasil pengelompokkan dan mengidentifikasi potensi pencilan yang memengaruhi kualitas pengelompokan.

Tabel 2. Evaluasi Model

Kelompok (Metode)	Metrik Evaluasi				
	Silhouette Score	*Davies Bouldin*	*Calinski Harabasz*	*Dunn Index*	*Within-Cluster Sum of Squares*
HDBSCAN	0.624	0.492	61.465	0.696	0.789
Pautan *Ward*	0.496	0.546	91.657	0.267	51.34
Pautan Rataan	0.566	0.403	83.21	0.012	74.52
Pautan Lengkap	0.522	0.405	100.761	0.389	47.05

Merujuk **Tabel 2.** mengenai evaluasi model menggunakan metrik evaluasi, HDBSCAN menunjukkan performa yang sangat baik di semua metrik yang diuji. Metode ini memiliki nilai tertinggi pada *Silhouette Score* (0.624), *Davies-Bouldin Index* (0.492), *Calinski-Harabasz Index* (61.465), *Dunn Index* (0.696), dan WCSS (0.789). Nilai evaluasi ini mengindikasikan bahwa kelompok yang dihasilkan oleh HDBSCAN tidak hanya terpisah dengan baik, tetapi juga menunjukkan kualitas kelompok yang konsisten di berbagai aspek evaluasi. Sebaliknya, metode pautan *ward*, pautan rataan, dan pautan lengkap masing-masing menunjukkan kelebihan dan kelemahan yang berbeda di metrik-metrik tersebut, dengan nilai yang tidak mencapai standar yang ditetapkan oleh HDBSCAN pada sebagian besar metrik evaluasi. HDBSCAN efektif dalam mengelompokkan data dengan sebaran terdapat pencilan karena kemampuannya untuk menangani kelompok dengan bentuk yang tidak teratur dan mendeteksi titik-titik pencilan.

4 Kesimpulan

Berdasarkan hasil pengelompokan 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. Metode AHC seperti *ward*, rataan, dan lengkap mampu melakukan pengelompokan tetapi hasil evaluasi menunjukkan bahwa HDBSCAN menghasilkan kelompok yang lebih baik berdasarkan berbagai metrik evaluasi. Keunggulan pengelompokkan HDBSCAN adalah kemampuannya mendeteksi dan memberi label khusus pada pencilan secara otomatis, sementara AHC tidak memberikan penanganan khusus terhadap pencilan.

Penelitian ini juga memberikan hasil analisis terhadap PDRB ADHK Provinsi Jawa Timur, karena pencilan seringkali menunjukkan peristiwa ekonomi yang tidak biasa terjadi, seperti perubahan kebijakan atau fluktuasi besar di beberapa sektor industri di PDRB. Pengelompokkan menggunakan HDBSCAN dilakukan guna menangani pencilan dinilai

mampu untuk menganalisis data PDRB yang kompleks, terutama dalam wilayah dengan distribusi yang tidak normal atau pencilan yang signifikan serta memberikan wawasan yang lebih tepat tentang dinamika ekonomi daerah per sektor ekonomi.

Daftar Pustaka

[1] Chaudhry, M., Shafi, I., Mahnoor, M., Vargas, D. L. R., Thompson, E. B., & Ashraf, I. (2023). A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective. *Symmetry*, *15*(9), 1679.

[2] O. Eric U. dan O. Michael O., "Overview of Agglomerative Hierarchical Clustering Methods," *British Journal of Computer, Networking and Information Technology*, vol. 7, no. 2, (pp. 14–23), Jun 2024, doi: 10.52589/bjcnit-cv9poogw.

[3] Vinisha, F.A., & Sujihelen, L. (2022). Study on Missing Values and Outlier Detection in Concurrence with Data Quality Enhancement for Efficient Data Processing. *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 1600-1607.

[4] YUSOFF, N. S. W. (2023). A New Single Linkage Robust Clustering Outlier Detection Procedures for Multivariate Data. *Sains Malaysiana*, *52*(8), 2431-2451.

[5] Zulkarnain, F. (2017). Implementasi Metode Average Linkage dalam Aplikasi Pengelompokan Sekolah Menengah Pertama (SMP) Berdasarkan Nilai Rata-rata Hasil Ujian Nasional Kabupaten Bondowoso.

[6] 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*.

[7] Stadlthanner, D., Steinkellner, H., Landschü, C., & Kaever, D. (2024). A Hierarchical Density-Based Clustering Method Applied to Mixed-Mail in Austria. *Logistics Research*, *17*(1).

[8] Amni, W. O. S. A., Jaya, A. K., & Ilyas, N. (2024). Perbandingan Analisis Komponen Utama Robust Minimum Covarian Determinant dengan Least Trimmed Square pada Data Produk Domestik Regional Bruto. *ESTIMASI: Journal of Statistics and Its Application*, 266-281.

[9] Wijoyo, N. A. (2016). Peramalan Nilai Tukar Rupiah Terhadap USD dengan Menggunakan Model GARCH. *Kajian Ekonomi dan Keuangan*, *20*(2), 169-189.

[10] Asyaky, M. S., & Mandala, R. (2021, September). Improving the performance of HDBSCAN on short text clustering by using word embedding and UMAP. In *2021 8th international conference on advanced informatics: Concepts, theory and applications (ICAICTA)* (pp. 1-6). IEEE.