Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

PERBANDINGAN ALGORITMA HDBSCAN DAN *AGGLOMERATIVE HIERARCHICAL CLUSTERING* DALAM MENGELOMPOKKAN DATA KETENAGAKERJAAN YANG *OUTLIERS*

Citra Amelia Intan Permadani[1], Aviolla Terza Damaliana[2], Mohammad Idhom[3]
1,2,3) Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur Indonesia

Article Info

Article history:

Received:14 Juli 2025
Revised: 15 Juli 2025
Accepted: 16 Juli 2025

ABSTRACT

Abstrak

Ketenagakerjaan merupakan indikator penting dalam mendukung pembangunan ekonomi nasional. Namun, distribusi tenaga kerja di Indonesia masih menunjukkan ketimpangan antarprovinsi. Beberapa provinsi memiliki kontribusi ekonomi dan tingkat pekerjaan formal yang tinggi, sementara yang lain tertinggal. Penelitian ini bertujuan mengidentifikasi pola distribusi ketenagakerjaan antarprovinsi dengan menerapkan analisis klaster menggunakan delapan variabel dari data BPS. Mengingat adanya pencilan dalam data, deteksi outlier dilakukan menggunakan metode *Local Outlier Factor* (LOF) yang mengidentifikasi enam provinsi sebagai outlier yaitu Jawa Barat, Jawa Tengah, Jawa Timur, DKI Jakarta, Banten, dan Sumatera Utara. Selanjutnya, data dianalisis menggunakan dua pendekatan klasterisasi, yaitu *Agglomerative Hierarchical Clustering* (*Single, Complete, Average Linkage*, dan *Ward*) dan HDBSCAN untuk membandingkan ketahanan metode terhadap data outlier. Validasi kualitas klaster dilakukan dengan *Silhouette Coefficient*. Hasil menunjukkan bahwa metode Single Linkage memiliki nilai koefisien tertinggi, namun kurang konsisten dalam memisahkan outlier. Sebaliknya, HDBSCAN lebih adaptif terhadap data yang mengandung *noise* dan pencilan dengan *Silhouette Coefficient* sebesar 0.546. Dengan demikian, HDBSCAN dinilai lebih efektif dalam analisis klasterisasi data ketenagakerjaan yang kompleks, sementara metode AHC lebih unggul dalam membentuk klaster yang jelas jika pencilan dapat ditangani secara terpisah.

Kata Kunci: HDBSCAN, *Agglomerative Hierarchical Clustering*, *Outliers*, Ketenagakerjaan.

Abstract

Employment is an important indicator in supporting national economic development. However, the distribution of labor in Indonesia still shows disparities between provinces. Some provinces have high economic contributions and formal employment rates, while others lag behind. This study aims to identify patterns of labor distribution across provinces by applying cluster analysis using eight variables from BPS data. Given the presence of outliers in the data, outlier detection was conducted using the Local Outlier Factor (LOF) method, which identified six provinces as outliers: West Java, Central Java, East Java, Jakarta Special Capital Region, Banten, and North Sumatra. The data was then analyzed using two clustering approaches: Agglomerative Hierarchical Clustering (Single, Complete, Average Linkage, and Ward) and HDBSCAN to compare the robustness of the methods against outlier data. Cluster quality validation was performed using the Silhouette Coefficient. The results showed that the Single Linkage method had the highest coefficient value but was less consistent in separating outliers. Conversely, HDBSCAN is more adaptive to data containing noise and outliers, with a Silhouette Coefficient of 0.546. Thus, HDBSCAN is considered more effective in analyzing complex labor data clustering, while the AHC method is superior in forming clear clusters if outliers can be handled separately.

Keywords: HDBSCAN, Agglomerative Hierarchical Clustering, Outliers, Employment.

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Corresponding Author:
E-mail : 21083010004@student.upnjatim.ac.id

1. PENDAHULUAN

Ketenagakerjaan merupakan salah satu dasar yang mendukung pembangunan ekonomi suatu negara. Dalam proses pembangunan ekonomi, Indonesia menghadapi tantangan meningkatnya jumlah angkatan kerja yang tidak sebanding dengan ketersediaan lapangan pekerjaan [1]. Menurut BPS, tingkat ketenagakerjaan suatu negara dapat dilihat melalui berbagai indikator yaitu penduduk usia kerja, angkatan kerja, pengangguran, lapangan usaha, dan gaji [2].

Namun, distribusi tenaga kerja di provinsi di Indonesia masih belum merata. Pernyataan ini didukung data SAKERNAS yang dipublikasikan oleh BPS tahun 2021, menunjukkan terdapat empat provinsi di Indonesia dengan persentase tenaga kerja formal yang cukup tinggi terdiri dari Provinsi Kepulauan Riau, DKI Jakarta, Banten, dan Kalimantan Timur. Sedangkan beberapa provinsi lainnya masih memiliki persentase tenaga kerja yang masih rendah meliputi Provinsi Papua, Nusa Tenggara Barat, Nusa Tenggara Timur, dan Sulawesi Barat [3].

Ketidakmerataan persebaran tenaga kerja ini menyebabkan banyak tenaga kerja mencari pekerjaan di provinsi yang memiliki lapangan pekerjaan yang lebih luas dan gaji yang lebih tinggi. Hal ini didukung oleh penelitian sebelumnya yang dilakukan oleh Amalia et al., dalam menganalisis variabel ketenagakerjaan terhadap produktivitas pekerja Indonesia tahun 2022 [4]. Hasil penelitian menunjukkan bahwa provinsi DKI Jakarta dan Banten, menjadi pencilan dalam variabel rata-rata upah buruh karena kedua provinsi tersebut berada di pusat perekonomian Indonesia, yang menarik banyak tenaga kerja dari berbagai daerah untuk mencari pekerjaan dengan upah yang lebih tinggi disertai tuntutan pekerjaan yang lebih berat.

Untuk mengatasi ketimpangan distribusi ketenagakerjaan antarprovinsi, diperlukan peningkatan dan pembangunan pada daerah-daerah dengan tingkat ketenagakerjaan yang masih rendah. Salah satu langkah awal yang dapat dilakukan adalah

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

mengidentifikasi provinsi-provinsi yang termasuk dalam kategori rendah. Proses pengelompokan ini dapat dilakukan secara lebih efektif menggunakan metode analisis klaster.

Analisis klaster merupakan metode yang digunakan untuk mengelompokkan data ke dalam kelompok kelompok berdasarkan kemiripannya [5]. Klasterisasi dapat dibagi menjadi dua kategori berdasarkan tingkat kemiripan, yaitu metode hierarki (*hierarchical clustering methods*) dan metode non-hierarki (*non-hierarchical clustering methods*).

Metode hierarki yang paling umum digunakan adalah metode agglomeratif yang memulai prosesnya dengan menganggap setiap objek sebagai klaster terpisah, kemudian secara bertahap mengelompokkan objek-objek tersebut ke dalam klaster yang semakin besar. Namun, metode agglomeratif memiliki kelemahan dalam menangani pencilan yang dapat mempengaruhi hasil pengelompokkan [6]. Untuk menangani pencilan dalam data, metode agglomeratif memperkenalkan pendekatan baru, yaitu metode HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) yang mampu mendeteksi klaster berdasarkan kepadatan dan secara otomatis mengidentifikasi pencilan (outlier) sebagai noise [7].

Ghardapaty et al., melakukan penelitian untuk membandingkan akgoritma Agglomerative Hierarchical Clustering dengan HDBSCAN dalam menangani pencilan menggunakan data Produk Domestik Regional Bruto Atas Dasar Harga Konstan (PDRB ADHK) Provinsi Jawa Timur tahun 2023 [7]. Penelitian tersebut mendapatkan hasil bahwa HDBSCAN memiliki performa terbaik dalam mengelompokkan data PDRB yang kompleks terutama dalam wilayah dengan distribusi yang tidak merata berdasarkan Silhouette Coefficient sebesar 0.624 serta mendeteksi dan mengelola pencilan secara efektif.

Hal ini sejalan dengan tujuan penelitian yang akan membandingkan algoritma HDBSCAN dengan Agglomerative Hierarchical Clustering berdasarkan nilai Silhouette Coefficient dalam mengelompokkan data ketenagakerjaan provinsi Indonesia.

2. METODE PENELITIAN

2.1. Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang bersumber dari Statistika Indonesia atau Indonesia dalam Angka oleh Badan Pusat Statistika pada

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Agustus 2022 meliputi usia produktif (15 hingga 64 tahun), pendidikan terakhir yang ditamatkan, dan jenis lapangan pekerjaan utama. Data yang digunakan berjumlah 34 data sesuai dengan jumlah Provinsi di Indonesia.

2.2. Variabel Penelitian

Variabel penelitian yang digunakan dalam penelitian ini berjumlah 8 variabel dengan keterangan pada Tabel 1.

Tabel 1. Variabel Penelitian

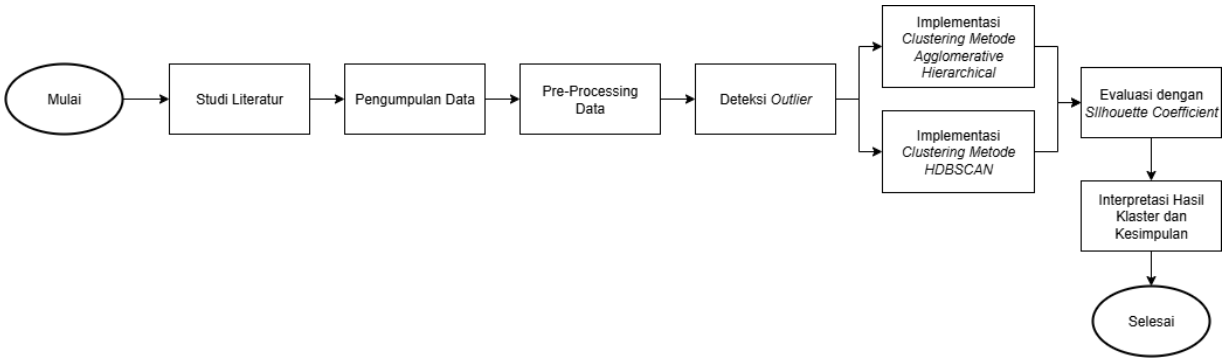
Variabel	Nama Variabel	Keterangan
X1	Usia Produktif	Jumlah penduduk usia produktif yang termasuk angkatan kerja
X2	<= SD/MI	Jumlah penduduk angkatan kerja tamat SD/MI
X3	SMP/MTS	Jumlah penduduk angkatan kerja tamat SMP
X4	SMA/SMK	Jumlah penduduk angkatan kerja tamat SMA/SMK
X5	Perguruan Tinggi	Jumlah penduduk angkatan kerja tamat Sarjana atau Diploma
X6	Primer	Jumlah penduduk yang bekerja di sektor primer
X7	Sekunder	Jumlah penduduk yang bekerja di sektor sekunder
X8	Tersier	Jumlah penduduk yang bekerja di sektor tersier

Berdasarkan variabel yang digunakan pada Tabel 1, memiliki kaitan erat dengan indikator ketenagakerjaan. Usia mempengaruhi tingkat partisipasi kerja, di mana kelompok usia produktif (15-64 tahun) cenderung memiliki tingkat partisipasi lebih tinggi dibandingkan kelompok usia non-produktif [8]. Pendidikan berperan dalam menentukan status pekerjaan, di mana individu dengan pendidikan lebih tinggi lebih cenderung bekerja di sektor formal dengan upah lebih baik dibandingkan mereka yang berpendidikan rendah. Selain itu, ketersediaan lapangan pekerjaan di setiap wilayah memengaruhi tingkat pengangguran dan mobilitas tenaga kerja, terutama di daerah dengan keterbatasan industri dan investasi.

2.3. Tahapan Pengolahan dan Analisis Data

Penelitian ini berfokus pada perbandingan evaluasi algoritma HDBSCAN dengan *Agglomerative Hierarchical Clustering* yang mencakup metode *Single Linkage, Average Linkage, Complete Linkage,* dan Ward dalam mengelompokkan data provinsi di Indonesia berdasarkan indikator ketenagakerjaan. Secara umum tahapan penelitian ini dapat dilihat pada Gambar 1.

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno



Gambar 1. Alir Diagram Penelitian

Berdasarkan alir diagram pada Gambar 1, langkah-langkah dalam penelitian ini dilakukan sebagai berikut.

2.3.1 Studi Literatur

Studi literatur adalah kajian secara mendalam berbagai sumber ilmiah yang relevan dengan menggunakan beragam pendekatan dan metode untuk mendukung topik penelitian [9]. Pada penelitian ini, studi literatur dilakukan dengan mengkaji berbagai penelitian terkait yang membahas metode klasterisasi dalam analisis ketenagakerjaan, penerapan algoritma *Agglomerative Hierarchical Clustering* dan HDBSCAN termasuk kelebihan dan kelemahannya dalam menangani data dengan pencilan.

2.3.2 Pengumpulan Data

Pengumpulan merupakan tahapan dalam memperoleh data [10]. Data pada penelitian dikumpulkan dengan memilih tabel yang berfokus pada kelompok usia, tingkat pendidikan yang ditamatkan, dan jenis lapangan pekerjaan dengan status tenaga kerja sebagai angkatan kerja di setiap provinsi di Indonesia.

2.3.3 Pre-Processing Data

Setelah data dikumpulkan, tahap selanjutnya adalah pre-processing data untuk menyaring dan mendapatkan informasi yang relevan. Tahap ini mencakup: (1) Pemilihan variabel yang digunakan (*data selection*), (2) Standarisasi atau normalisasi data, dan (3) Pengecekan atau deteksi outlier pada data. Untuk mengatasi perbedaan skala antar variabel, standarisasi data dilakukan menggunakan metode *Standard Scaler*.

Terdapat asumsi dalam analisis *cluster*, yaitu sampel representatif yang diuji menggunakan KMO untuk melihat apakah sampel yang diambil sudah mewakili populasi. Nilai KMO berkisar antara 0 hingga 1 dengan kriteria jika nilai KMO < 0,5 maka

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

dapat dikatakan bahwa sampel yang diambil tidak representatif sedangkan jika nilai KMO ≥ 0,5 maka sampel yang diambil sudah representatif. Untuk menghitung nilai KMO menggunakan persamaan 1.

$$KMO = \frac{\sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p} {r_{jk}}^2}{\sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p} {r_{jk}}^2 + \sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p}\sum_{l=1}^{p} {\rho_{j_{(kl)}}}^2} \quad (1)$$

dengan p adalah banyaknya variabel, r_{jk} adalah koefisien korelasi variabel ke-j dan variabel ke-k, serta $\rho_{j(kl)}$ adalah koefisien korelasi parsial antara variabel ke-j, ke-k, dan ke-l.

2.3.4 Deteksi Outlier

Dalam melakukan pre-processing data, terdapat langkah melakukan identifikasi *outlier*. Dalam proses klastering, *outlier* dalam residual yang menyebabkan ketidaknormalan dapat diidentifikasi sebagai data yang menyimpang dari pola umum yang terbentuk dalam klaster utama, sehingga memerlukan penanganan khusus agar tidak mengganggu validitas hasil klasterisasi [11]. Cara yang digunakan pada penelitian ini adalah metode *Local Outlier Factor* (LOF).

Local Outlier Factor (LOF) adalah algoritma unsupervised learning yang digunakan untuk mendeteksi anomali dalam data, dengan tujuan mengidentifikasi titik-titik data yang memiliki karakteristik berbeda dari mayoritas [12]. LOF bekerja dengan mengukur tingkat keanehan suatu data berdasarkan kedekatannya dengan tetangga terdekat. Prosesnya dimulai dengan menghitung jarak antar titik data dan tetangganya, lalu menghitung *local density* dari setiap titik sebagai rata-rata jarak ke tetangga-tetangganya tersebut. Nilai LOF suatu data kemudian diperoleh dengan membandingkan *local density*-nya dengan *density* milik tetangganya. Titik data yang memiliki nilai LOF tinggi, artinya jauh lebih jarang atau tidak seragam dibandingkan lingkungan sekitarnya, akan dianggap sebagai anomali.

2.3.5 HDBSCAN

Penelitian ini menggunakan HDBSCAN sebagai metode pertama pengelompokan data ketenagakerjaan. Percobaan pengujian dengan metode HBDSCAN menggunakan data hasil standarisasi yang mengandung pencilan dengan 8 variabel ketenagakerjaan dan menggunakan variasi nilai *min_cluster_size* = 2, 3, 4, 5, dan 6. Penggunaan variasi nilai *min_cluster_size* didasarkan pada penelitian sebelumnya [13]. Adapun langkah-

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

langkah yang dilakukan penulis dalam mengelompokan provinsi berdasarkan data ketenagakerjaan menggunakan metode HDBSCAN adalah sebagai berikut:

1. Membuat parameter awal ukuran minimum kluster $(M_{pts}) = 2, 3, 4, 5, \ dan\ 6$.
2. Menghitung *Core Distance* yaitu jarak setiap titik data ke tetangga ke k (M_{pts}) terdekat. Jika titik memiliki jumlah tetangga yang kurang dari M_{pts}, maka titik tersebut dianggap sebagai *noise* dan tidak akan digunakan sebagai pusat kluster.
3. Menghitung *Mutual Raechability Distance* antara semua pasangan titik yang didefinisikan sebagai berikut:

   $$MRD(a,b) = \max\big(coreDistance_k(a), coreDistance_k(b), d(a,b)\big) \quad (2)$$

   dengan,

   $d(a,b)$: jarak antara titik a dan b

   $coreDistance$: jarak dari suatu titik ke tetangga terdekat ke-k
4. Membuat *Minimum Spanning Tree* (MST) dari semua titik berdasarkan *Mutual Raechability Distance* untuk menyusun titik-titik data menjadi pohon yang menghubungkan semua titik dengan jarak terkecil.
5. Membangun struktur hierarki kluster atau *dendogram* yang berfungsi mengidentifikasi kluster stabil dengan menghapus tepi-tepi dari MST yang memiliki *mutual reachability distance* tinggi kemudian ditandai sebagai batas antara kluster.
6. Memadatkan pohon hierarki kluster dan menyisakan kluster yang stabil pada kepadatan tertentu sementara kluster yang tidak stabil akan dihilangkan dari hierarki dan dianggap sebagai *noise* (biasanya diberi label -1).
7. Menghitung nilai SC (*Silhoutte Coefficient*) untuk setiap kombinasi parameter M_{pts} yang diberikan.

2.3.6 Agglomerative Hierarchical Clustering (AHC)

Metode kedua yang digunakan untuk membandingkan hasil pengelompokan adalah *Agglomerative Hierarchical Clustering* (AHC). AHC merupakan metode klasterisasi hirarkis yang mengelompokkan data berdasarkan kemiripan karakteristiknya, membentuk hierarki klaster secara bertahap dari individu hingga terbentuk kelompok yang lebih besar menggunakan jarak *Euclidean*. Pada penelitian ini, digunakan empat pendekatan dalam AHC, yaitu:

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

1. Single Linkage

   Single linkage method melakukan pengelompokan yang didasarkan pada jarak terdekat antar objeknya [14]. Berikut cara kerja metode single linkage:

   - Misalkan bentuk matriks jarak Euclidean untuk matriks data sampel

$$D(1)_{mxm} = \begin{bmatrix} d_{11} & d_{12} & \dots d_{13} \\ d_{21} & d_{22} & \dots d_{23} \\ d_{m1} & d_{m2} & \dots d_{m3} \end{bmatrix} \quad (3)$$

   - Mengasumsikan bahwa tiap data dianggap sebagai kelompok atau klaster, kemudian menentukan klaster yang memiliki jarak terdekat. Contohnya klaster XY merupakan gabungan dari klaster X dan klaster Y yang memiliki jarak terdekat. Jarak terdekat diukur menggunakan jarak Euclidean dengan rumus berikut:

$$d_{jl} = \left\{ (x_l - x_j)'(x_l - x_j) \right\}^{\frac{1}{2}} = \sqrt{\sum_{k=1}^{i} (x_{lk} - x_{jk})^2} \quad (4)$$

     Dengan, $i = 1, 2, \dots, p$ atau $l = 1, 2, \dots, n$.

   - Dari hasil gabungan klaster XY akan dicari jarak minimum antar klaster XY dengan klaster lainnya yang belum bergabung.
   - Mengulangi langkah 2 sampai semua objek bergabung menjadi kelompok.

2. Average Linkage

   Metode Average Linkage melakukan pengelompokan dengan menghitung jarak rata-rata antara setiap pasangan objek dalam dua klaster yang akan digabungkan. Perbedaan utama metode ini dibandingkan metode linkage lainnya terletak pada langkah ketiga, yaitu proses pembentukan klaster yang didasarkan pada rata-rata jarak antar klaster dengan objek lain yang belum bergabung.

3. Complete Linkage

   Metode Complete Linkage melakukan pengelompokan dengan menghitung jarak terjauh antara pasangan objek dalam dua klaster yang akan digabungkan. Jika dua objek memiliki jarak terbesar di antara klaster yang sedang dipertimbangkan, maka penggabungan akan dilakukan berdasarkan jarak tersebut.

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

4. Ward

Metode Ward memiliki perhitungan yang berbeda dari metode agglomeratif lainnya. Perhitungan jarak antar klaster pada metode ward merupakan total jumlah kuadrat dua klaster pada masing-masing variabel. Berikut tahapan metod ward:

- Mengasumsikan setiap data sebagai kelompok atau klaster
- Membentuk klaster yang terdiri dari pasangan dua objek sehingga klaster sebanyak C_2^n, kemudian menghitung jumlah kuadrat (SSE) dari semua pasangan klaster menggunakan rumus:

$$SSE = \sum_{j=1}^{n} {x_j}^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2 \quad (5)$$

- Memilih nilai SSE terkecil kemudian menggabungkan pasangan dari klaster membentuk kelompok atau klaster baru.
- Mengulangi langkah 2 hingga membentuk satu klaster.

2.3.7 Evaluasi

Evaluasi hasil pengelompokan metode *Agglomerative Hierarchical Clustering* dan HDBSCAN dilakukan untuk menilai kinerja masing-masing metode dalam mengelompokkan data provinsi di Indonesia yang mengandung pencilan. Evaluasi ini divalidasi menggunakan nilai *Silhouette Coefficient* dari hasil klaster untuk menentukan metode yang paling efektif. Nilai *Silhouette Coefficient* dapat dihitung menggunakan persamaan 6.

$$silhouette(x) = \frac{b(x) - a(x)}{max\{a(x), b(x)\}} \quad (6)$$

dengan, $a(x)$ adalah rata-rata jarak dari suatu objek dengan objek lain yang berada dalam satu cluster dan $b(x)$ adalah nilai jarak yang terkecil. Interpretasi nilai *Silhouette Coefficient* yang digunakan mengacu pada penjelasan yang dibuat oleh Kaufman dan Rousseeuw, yang dapat dilihat pada Tabel 2.

Tabel 2. Interpretasi Nilai Silhouette Coefficient

No.	Rentang Nilai	Interpretasi
1.	0.7 < SC ≤ 1	Susunan Kelompok Data Sangat Baik
2.	0.5 < SC ≤ 0.7	Susunan Kelompok Data Baik
3.	0.25 < SC ≤ 0.5	Susunan Kelompok Data Lemah
4.	SC ≤ 0.25	Susunan Kelompok Data Buruk

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Tahap terakhir dalam penelitian ini adalah menyusun kesimpulan berdasarkan perbandingan kedua metode dalam membentuk klaster pada data *outlier*, serta menganalisis hasil klasterisasi yang berkaitan dengan kondisi angkatan kerja di setiap provinsi di Indonesia.

3. HASIL DAN PEMBAHASAN

Penelitian ini membandingkan metode pengelompokan klaster pada data ketenagakerjaan tiap provinsi di Indonesia yang mengandung pencilan (*outlier*). Metode yang dibandingkan meliputi *Agglomerative Hierarchical Clustering* (AHC) dengan empat pendekatan, yaitu *Single Linkage, Average Linkage, Complete Linkage*, dan Ward, serta HDBSCAN. Perbandingan difokuskan pada ketahanan masing-masing metode terhadap keberadaan pencilan serta kemampuan metode dalam mengelompokkan provinsi-provinsi *outlier* secara tepat. Untuk melihat perbandingan antara HDBSCAN dan AHC secara lebih rinci disajikan pada Tabel 3.1.

Tabel 3. Perbandingan HDBSCAN vs AHC

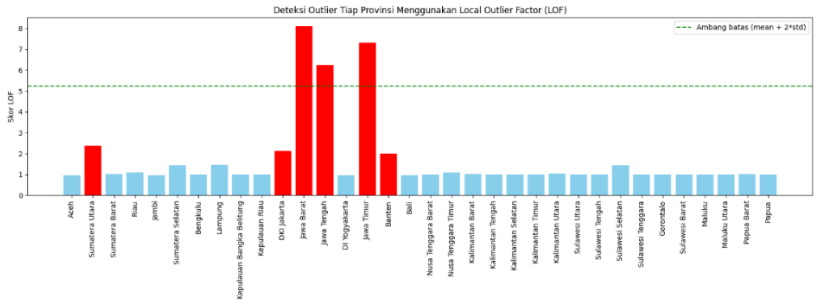
Aspek	**HDBSCAN**	***Agglomerative Hierarchical Clustering* (AHC)**
Penentuan Jumlah Klaster	Tidak memerlukan penentuan jumlah klaster di awal. Klaster ditentukan otomatis berdasarkan kepadatan.	Jumlah klaster ditentukan dengan cara memotong dendrogram pada tingkat tertentu berdasarkan jarak antar objek.
Penanganan Pencilan	Mampu mengidentifikasi pencilan (outlier) dan otomatis memberikan kategori sebagai noise	Tidak dapat mendeteksi pencilan sehingga semua data dipaksa masuk ke dalam suatu klaster
Pembentukan Klaster	Mampu membentuk klaster dengan densitas atau kepadatan data yang berbeda-beda (density-based clustering).	Cenderung membentuk klaster berbasis jarak (distance-based) tanpa memperhatikan kepadatan
Kompleksitas	Lebih kompleks secara komputasi, cocok untuk data berukuran besar dan tidak berstruktur.	Relatif lebih sederhana dan efisien untuk data berukuran sedang hingga kecil

Selanjutnya, validasi klaster dilakukan menggunakan *Silhouette Coefficient* untuk membandingkan kualitas pengelompokan yang dihasilkan oleh masing-masing metode.

3.1 Deteksi Outlier

Data ketenagakerjaan yang digunakan dalam penelitian ini mengandung pencilan atau *outlier*. Hal tersebut disebabkan oleh perbedaan distribusi penduduk dan karakteristik wilayah di setiap provinsi di Indonesia. Deteksi pencilan pada penelitian ini menggunakan metode LOF yang dapat dilihat pada Gambar 2.

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno



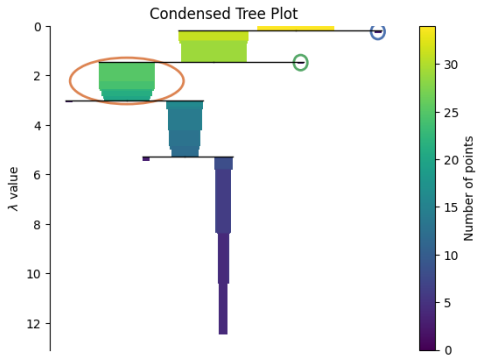
Gambar 2. Visualisasi Outlier dengan Local Outlier Factor (LOF)

Berdasarkan visualisasi pada Gambar 2, terdapat enam provinsi yang terdeteksi sebagai outlier oleh algoritma *Local Outlier Factor*, yaitu Jawa Timur, Jawa Barat, Jawa Tengah, Sumatera Utara, DKI Jakarta, dan Banten. Keenam provinsi ini memiliki skor LOF yang tinggi yang menunjukkan bahwa kerapatan lokal data mereka jauh lebih rendah dibandingkan dengan provinsi-provinsi sekitarnya.

3.2 Hasil Klasterisasi dengan HDBSCAN

Metode pertama yang digunakan untuk mengelompokkan data ketenagakerjaan adalah HDBSCAN. Proses klasterisasi dengan HDBSCAN dilakukan menggunakan data ketenagakerjaan yang telah dinormalisasi. Pemilihan nilai *min_cluster_size* terbaik ditentukan berdasarkan hasil nilai *Silhouette Coefficient* tertinggi dari masing-masing percobaan, sementara parameter *min_samples* ditetapkan sebesar 2. Berdasarkan hasil evaluasi, nilai terbaik diperoleh pada *min_cluster_size* = 2 dengan nilai *Silhouette Coefficient* sebesar 0.501.

Selain melalui nilai *Silhouette Coefficient*, pembentukan klaster metode HDBSCAN juga dapat divisualisasikan melalui *Condensed Tree Plot* seperti yang terlihat pada Gambar 3.

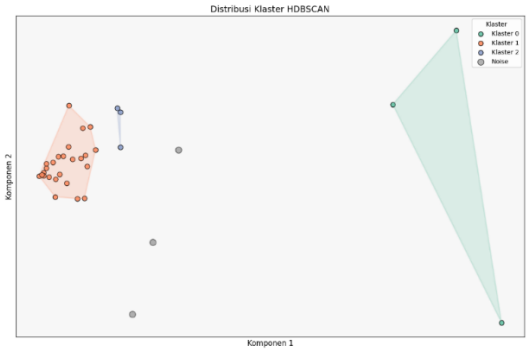


Gambar 3. Condesed Tree Plot klaster HDBSCAN

Berdasarkan hasil visualisasi Condensed Tree Plot, tampak bahwa struktur klaster yang terbentuk cukup stabil dan terpisah jelas ketika *nilai min_cluster_size* = 2 digunakan.

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Tahap selanjutnya adalah melakukan proses klasterisasi menggunakan metode HDBSCAN dengan data ketenagakerjaan yang telah dinormalisasi. Berdasarkan hasil klasterisasi tersebut, HDBSCAN berhasil membentuk tiga klaster utama dari data yang dianalisis. Hasil klasterisasi menggunakan HDBSCAN dapat dilihat pada visualisasi berikut.



Gambar 4. Visualisasi Hasil Klaster Metod HDBSCAN

Berdasarkan visualisasi pada Gambar 4, data terbagi ke dalam tiga klaster utama yang ditandai dengan warna area yang berbeda, serta terdapat beberapa provinsi yang diklasifikasikan sebagai *noise*, yang ditampilkan dalam warna abu-abu.

Provinsi-provinsi yang termasuk dalam kategori noise pada hasil klasterisasi HDBSCAN dapat dianggap sebagai *outlier* karena tidak memiliki kepadatan data yang cukup untuk membentuk klaster tersendiri ataupun tidak sesuai dengan pola kepadatan dari klaster yang sudah terbentuk.

3.3 Hasil Klasterisasi dengan Agglomerative Hierarchical Clustering

Metode kedua yang digunakan untuk perbandingan adalah *Agglomerative Hierarchical Clustering* (AHC), yang mencakup *Single Linkage, Average Linkage, Complete Linkage,* dan Ward. Sebelum melakukan klastering, dilakukan pengecekan sampel representatif menggunakan uji KMO. Hasil uji menunjukkan bahwa nilai KMO sebesar 0.598 yang berada pada kategori cukup layak untuk analisis klaster. Tahap selanjutnya yaitu menentukan *k* terbaik untuk jumlah klaster yang akan dibentuk dengan metode AHC. Penentuan klaster terbaik menggunakan nilai *Silhouette Coefficient* dapat dilihat pada Tabel 4.

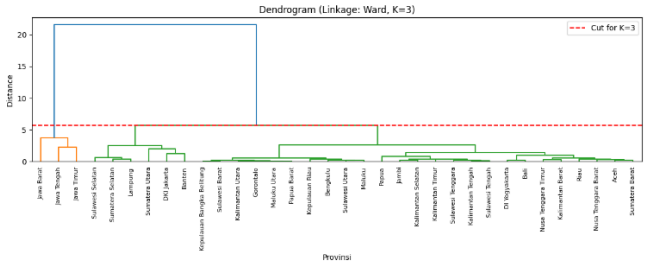
Tabel 4. *k* Terbaik Metode AHC

Metode	Banyaknya Cluster	Nilai Silhouette Coefficient
Single Linkage	4	0.776
Average Linkage	4	0.557
Complete Linkage	4	0.589

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Ward	3	0.591

Berdasarkan kriteria evaluasi menggunakan *Silhouette Coefficient*, metode Ward memberikan hasil yang paling optimal karena berada dalam rentang kualitas klaster yang baik (0,51–0,70) dan relatif stabil. Meskipun metode Single Linkage menghasilkan skor tertinggi, nilai ini melampaui rentang ideal dan dapat mengindikasikan adanya pembentukan klaster yang terlalu longgar atau tidak seimbang. Oleh karena itu, metode Ward linkage dengan 3 klaster dipilih sebagai konfigurasi terbaik untuk pembentukan klaster yang mempertimbangkan keseimbangan antara kualitas klaster dan jumlah kelompok yang cukup untuk membedakan karakteristik provinsi secara lebih spesifik. Pengelompokkan metode dapat dilihat melalui dendogram Gambar 5.



Gambar 5. Dendogram Metode Ward k=3

Langkah selanjutnya adalah pembentukan klaster dengan metode Agglomerative. Setiap metode Agglomerative, memiliki proporsi pengelompokkan data yang berbeda-beda. Hal tersebut dapat dilihat pada Tabel 5.

Tabel 5. Proporsi Hasil Klaster Metode Agglomerative Hierarchical Clustering

Klaster	Single Linkage	Complete Linkage	Average Linkage	Ward
1	31 (91.18%)	25 (73.53%)	28 (82.35%)	2 (5.88%)
2	1 (2.94%)	6 (17.65%)	3 (8.82%)	1 (2.94%)
3	1 (2.94%)	2 (5.88%)	2 (5.88%)	6 (17.65%)
4	1 (2.94%)	1 (2.94%)	1 (2.94%)	25 (73.53%)

Dari Tabel 5 terlihat bahwa keempat metode Agglomerative menghasilkan kelompok klaster yang berbeda-beda dan masing-masing metode memiliki setidaknya satu klaster yang hanya terdiri dari satu provinsi. Klaster tunggal ini dapat dianggap sebagai indikasi adanya *outlier*, yaitu provinsi yang memiliki karakteristik sangat berbeda dibandingkan dengan provinsi lainnya. Namun, jika dilihat lebih lanjut, metode-metode tersebut secara keseluruhan hanya mengidentifikasi tiga provinsi sebagai *outlier*

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

sedangkan 3 provinsi lain yang termasuk *outlier* masih menjadi satu klaster dengan provinsi lainnya.

Perbedaan ini menunjukkan bahwa tidak semua provinsi yang teridentifikasi sebagai *outlier* juga dikenali sebagai klaster terpisah dalam pendekatan klasterisasi hierarkis. Hal ini menunjukkan bahwa metode Agglomerative mampu mengelompokkan data berdasarkan kemiripan namun memiliki keterbatasan dalam mendeteksi semua *outlier* yang tersebar jika perbedaan karakteristiknya tidak cukup signifikan untuk membentuk klaster tersendiri.

3.4 Evaluasi dan Pemilihan Metode Klaster Terbaik

Tahap selanjutnya dalam penelitian ini adalah melakukan evaluasi terhadap hasil klasterisasi untuk menentukan metode paling optimal dalam mengelompokkan data ketenagakerjaan yang mengandung pencilan (outlier). Evaluasi dilakukan menggunakan metrik Silhouette Coefficient dari setiap metode klaster yang digunakan. Hasil evaluasi ini memberikan gambaran sejauh mana masing-masing metode mampu membentuk struktur klaster yang jelas dan terpisah secara optimal. Hasil evaluasi metode klaster menggunakan Silhouette Coefficient untuk semua metode dapat dilihat pada Tabel 6.

Tabel 6. Hasil Evaluasi Metode Klasterisasi dengan Silhouette Coefficient

No.	Metode	Silhouette Coefficient
1	HDBSCAN	0.546
2	Single Linkage	0.776
3	Average Linkage	0.557
4	Complete Linkage	0.589
5	Ward Method	0.557

Berdasarkan Tabel 6, pengelompokan menggunakan metode Single Linkage memberikan hasil yang lebih baik dengan nilai Silhouette sebesar 0.776, dibandingkan dengan HDBSCAN yang hanya mencapai 0.546. Meskipun nilai Silhouette Coefficient dari HDBSCAN hanya mencapai 0.546 namun nilai tersebut masih termasuk dalam susunan kelompok yang baik. Hal ini perlu dipertimbangkan dalam konteks kompleksitas data terutama terkait keberadaan noise dan outlier yang memengaruhi hasil klasterisasi. Pendekatan HDBSCAN tetap relevan karena memiliki kemampuan untuk menangani data secara menyeluruh, termasuk mengidentifikasi provinsi-provinsi yang dianggap sebagai noise dan outlier.

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

3.5 Interpretasi Hasil Klaster dengan Metode Terbaik

Metode klasterisasi terbaik menurut nilai *Silhouette Coefficient* adalah *Single Linkage* dengan skor tertinggi yang menunjukkan kualitas pemisahan klaster yang baik secara umum. Namun, jika mempertimbangkan keberadaan *noise* atau *outlier* dalam data yang sebelumnya telah diidentifikasi dengan *Local Outlier Factor*, maka metode HDBSCAN menjadi sangat relevan dengan nilai *Silhouette Coefficient* yang termasuk susunan kelompok baik. Berdasarkan Tabel 7, hasil pengelompokan menggunakan metode HDBSCAN membagi provinsi di Indonesia ke dalam empat klaster, termasuk satu klaster noise.

Tabel 7. Hasil Klaster Metode HDBSCAN

Cluster	Jumlah Anggota	Provinsi
0	3	Jawa Barat, Jawa Tengah, dan Jawa Timur.
1	25	Aceh, Sumatera Barat, Riau, Jambi, Bengkulu, Kepulauan Bangka Belitung, Kepulauan Riau, DI Yogyakarta, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Papua.
2	3	Sumatera Selatan, Lampung, Sulawesi Selatan.
-1	3	Sumatera Utara, DKI Jakarta, Banten.

Berdasarkan Tabel 7, karakteristik tiap kelompok provinsi sebagai berikut:

- Cluster 0 terdiri dari provinsi dengan populasi usia produktif terbesar di Indonesia, akses pendidikan menengah dan perguruan tinggi yang lebih tinggi dan merata, serta menunjukkan aktivitas ekonomi yang padat, terutama di sektor sekunder dan tersier. Cluster ini teridentifikasi sebagai outlier berdasarkan hasil metode Local Outlier Factor. Meskipun tergolong outlier, ketiga provinsi tersebut masih dapat dikelompokkan bersama karena memiliki karakteristik ketenagakerjaan yang serupa dan secara signifikan berbeda dari mayoritas provinsi lainnya.
- Cluster 1 terdiri dari provinsi dengan populasi dari usia produktif yang lebih rendah, proporsi penduduk berpendidikan dasar hingga menengah yang masih tinggi, serta ketergantungan yang besar terhadap sektor primer seperti pertanian dan perikanan.
- Cluster 2 terdiri dari provinsi-provinsi dengan jumlah penduduk usia produktif yang cukup besar, tingkat pendidikan menengah yang cukup merata, serta

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

memiliki struktur ketenagakerjaan yang seimbang antara sektor primer dan sekunder.

- Cluster -1 merupakan kelompok noise, yang mencerminkan provinsi-provinsi dengan distribusi indikator yang tidak seimbang. Meskipun memiliki jumlah penduduk usia produktif yang tinggi, provinsi dalam klaster ini menunjukkan ketimpangan yang mencolok antara pendidikan rendah dan tinggi, serta distribusi sektor pekerjaan yang berbeda signifikan dari pola umum. Cluster ini juga terdiri dari anggota yang teridentifikasi sebagai outlier. Namun, karakteristik ketenagakerjaan dari masing-masing provinsi dalam kelompok ini sangat beragam dan tidak memiliki kesamaan yang cukup kuat untuk dikelompokkan ke dalam klaster utama mana pun.

4. SIMPULAN

Berdasarkan hasil pengumpulan dan pengolahan data, diperoleh beberapa kesimpulan penting. Hasil klasterisasi terhadap data provinsi di Indonesia berdasarkan indikator ketenagakerjaan dengan 8 variabel menunjukkan keberadaan *outlier*, di mana terdapat enam provinsi yang teridentifikasi sebagai pencilan menggunakan metode *Local Outlier Factor*, yaitu Jawa Barat, Jawa Tengah, Jawa Timur, DKI Jakarta, Banten, dan Sumatera Utara. Keberadaan pencilan ini disebabkan oleh perbedaan signifikan dalam jumlah penduduk usia produktif serta karakteristik wilayah antarprovinsi yang memengaruhi struktur ketenagakerjaan.

Dari lima metode klasterisasi yang digunakan, metode *Single Linkage* menghasilkan nilai *Silhouettes Coefficient* tertinggi yang menandakan struktur klaster yang paling jelas namun belum sepenuhnya mampu memisahkan provinsi-provinsi outlier ke dalam klaster yang terpisah secara konsisten. Sedangkan metode HDBSCAN menghasilkan klaster yang stabil dan mengelompokkan pencilan dengan jelas dengan nilai *Silhouette Coefficient* sebesar 0.546 yang menunjukkan susunan kelompok yang baik. Hal ini menunjukkan keunggulan metode HDBSCAN dalam mengidentifikasi struktur klaster secara lebih menyeluruh, termasuk dalam menangani keberadaan *outlier* dan *noise* yang belum sepenuhnya dapat ditangani oleh metode klasterisasi tradisional. Dengan demikian, HDBSCAN menjadi metode yang lebih adaptif terhadap pencilan dan efektif dalam mengelompokkan karakteristik ketenagakerjaan antarprovinsi di Indonesia.

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

REFERENCES

[1] A. T. Damaliana and Setiawan, "Pemodelan Penyerapan Tenaga Kerja Sektor Industri di Indonesia Dengan Pendekatan Regresi Data Panel Dinamis," 2016.

[2] Badan Pusat Statistik, "Tenaga Kerja - Tabel Statistik - Badan Pusat Statistik Indonesia," Https://Www.Bps.Go.Id/Id/Statistics-Table?Subject=520.

[3] Muhammad Azis Azra N, "Penggunaan Davies Bouldin Index Dalam Perbandingan Algoritma K-Means Dan K-Medoids Untuk Klasterisasi Provinsi di Indonesia Berdasarkan Tingkat Indikator Ketenagakerjaan," Jambi, Jambi, 2023.

[4] A. I. Widiyasari, I. A'mal, N. Fatma, P. Yunardi, and F. Kartiasih, "Analisis Variabel Ketenagakerjaan Terhadap Produktivitas Pekerja Indonesia Tahun 2022 (Analysis of Employment Variables on Indonesian Labor Productivity In 2022)."

[5] P. Ryan Harnanda, T. Maulana Fahrudin, and N. Damastuti, "Gis Implementation and Classterization of Potential Blood Donors Using the Agglomerative Hierarchical Clustering Method," *International Journal of Electrical Engineering and Information Technology*, Vol. 03, 2020.

[6] S. Hidayati, A. T. Darmaliana, and R. Riski, "Comparison Of K-Means, Fuzzy C-Means, Fuzzy Gustafson Kessel, And Dbscan for Village Grouping iIn Surabaya Based n Poverty Indicators," *Jurnal Pendidikan Matematika (Kudus)*, Vol. 5, No. 2, P. 185, Dec. 2022, Doi: 10.21043/Jpmk.V5i2.16552.

[7] G. Ghaly Ghiffary, K. Alifviansyah, A. Fitrianto, L. M. Risman, and D. Jumansyah, "Perbandingan Algoritma HDBSCAN Dan Agglomerative Hierarchical Clustering dalam Klasterisasi pada Data yang Mengandung Pencilan," *Jurnal Riset Dan Aplikasi Matematika*, Vol. 08, No. 02, Pp. 122–135, 2024.

[8] Rena Armita Sari and Rr Retno Sugiharti, "Analisis Faktor-Faktor yang Mempengaruhi Tingkat Partisipasi Angkatan Kerja Di Indonesia Tahun 2001-2020 ," *Jiep: Jurnal Ilmu Ekonomi Dan Pembangunan*, Vol. 5, No. 2, Pp. 603–616, 2022.

[9] M. Idhom, D. A. Prasetya, P. A. Riyantoko, T. M. Fahrudin, and A. P. Sari, "Pneumonia Classification Utilizing Vgg-16 Architecture and Convolutional Neural Network Algorithm for Imbalanced Datasets," *Tiers Information Technology Journal*, Vol. 4, No. 1, Pp. 73–82, Jun. 2023, Doi: 10.38043/Tiers.V4i1.4380.

[10] A. T. Damaliana, K. M. Hindrayani, and T. M. Fahrudin, "Hybrid Holt Winter-Prophet Method to Forecast the Num-Ber of Foreign Tourist Arrivals Through Bali's Ngurah Rai Airport", Doi: 10.3390/Xxxxx.

[11] M. Idhom, A. Fauzi, T. Trimono, and P. Riyantoko, "Time Series Regression: Prediction of Electricity Consumption Based on Number of Consumers at National Electricity Supply Company," *Tem Journal*, Vol. 12, No. 3, Pp. 1575–1581, Aug. 2023, Doi: 10.18421/Tem123-39.

[12] W. Yustanti, "Studi Komparasi Local Outlier Factor (Lof) Dan Isolation Forest (If) Pada Analisis Anomali Kinerja Dosen," *Journal of Informatics and Computer Science*, Vol. 06, 2024.

[13] N. A. Wahyuni, M. N. Hayati, and N. A. Rizki, "Metode Hierarchical Density-Based Spatial Clustering of Application with Noise (HDBSCAN) Pada Wilayah Desa/Kelurahan Tertinggal Di Kabupaten Kutai Kartanegara," *Eksponensial*, Vol. 12, No. 1, P. 47, Jun. 2021, Doi: 10.30872/Eksponensial.V12i1.758.

[14] M. Alexander Justin Audison Sibarani, I. Gede Susrama Mas Diyasa, P. Studi Sains Data, F. Ilmu Komputer, and U. Pembangunan, "Penggunaan K-Means dan Hierarchical Clustering Single Linkage dalam Pengelompokkan Stok Obat," Vol. 5, No. 2, 2024, Doi: 10.46306/Lb.V5i2.