

COMPARISON OF K-MEANS++ AND AGGLOMERATIVE HIERARCHICAL METHODS IN CLUSTERING HEALTHCARE WORKERS

KOMPARASI METODE K-MEANS++ DAN AGGLOMERATIVE HIERARCHICAL DALAM PENGELOMPOKAN JUMLAH TENAGA KESEHATAN

Citra Tjipta Nur Handayani¹, Melkior N. N. Sitokdana²

^{1,2}Universitas Kristen Satya Wacana, Salatiga

Politeknik Negeri Bengkalis, Jl. Bathin Alam, Sei. Alam, Bengkalis

*citranh14@gmail.com*¹, *melkior.sitokdana@uksw.edu*²

Abstrack - As an archipelagic country, Indonesia faces disparities in the distribution of healthcare workers, influenced by its diverse geographical conditions. These disparities impact the equitable access to healthcare services across the country. This study aims to compare the effectiveness of two clustering methods, namely K-Means++ and Agglomerative Hierarchical Clustering, using secondary data from Statistics Indonesia (BPS) on the Number of Healthcare Workers by Province in 2023, covering 38 provinces and 13 categories of healthcare professions. The evaluation was conducted using three metrics: Silhouette Score to measure cluster cohesion, Davies-Bouldin Index to assess inter-cluster separation, and Calinski-Harabasz Index to compare inter-cluster variance. The results show that Agglomerative Hierarchical outperformed K-Means++ in Silhouette Score (0.550) and Davies-Bouldin Index (0.457), while K-Means++ performed better in the Calinski-Harabasz Index (63.630). A 2D PCA visualization further illustrates the structural differences between the clusters formed by each method. These findings provide insights into selecting the most appropriate clustering method for analyzing the distribution of healthcare workers and can support data-driven decision-making by policymakers

Keywords - Clustering, K-Means++, Agglomerative Hierarchical, Healthcare Workforce, Python.

Abstrak- Sebagai negara kepulauan, Indonesia menghadapi ketimpangan distribusi tenaga kesehatan yang dipengaruhi oleh kondisi geografis. Ketimpangan ini berdampak pada akses layanan kesehatan yang belum merata di seluruh wilayah. Penelitian ini bertujuan untuk membandingkan efektifitas dua metode *clustering*, yaitu K-Means++ dan Agglomerative Hierarchical dengan menggunakan data sekunder dari Badan Pusat Statistik (BPS) yaitu Jumlah Tenaga Kesehatan menurut Provinsi 2023 yang mencakup 38 provinsi dan 13 kategori jenis tenaga kesehatan. Evaluasi dilakukan menggunakan 3 metrik, yaitu Silhouette Score untuk mengukur kekompakan *cluster*, Davies-Bouldin Index untuk menilai keterpisahan antar *cluster*, dan Calinski-Harabasz Index untuk membandingkan variansi inter-cluster. Hasil menunjukkan bahwa Agglomerative Hierarchical unggul pada Silhouette Score (0,550) dan Davies-Bouldin Index (0,457), sedangkan K-Means++ lebih unggul pada Calinski-Harabasz Index (63,630). Visualisasi PCA 2D akan memperjelas perbedaan struktur *cluster* yang dihasilkan oleh masing-masing metode. Temuan ini memberikan gambaran dalam memilih metode *clustering* yang paling sesuai untuk melakukan *clustering* distribusi tenaga kesehatan, guna mendukung pengambilan keputusan pembuat kebijakan.

Kata Kunci - Clustering, K-Means++, Agglomerative Hierarchical, Tenaga Kesehatan, Python.

I. PENDAHULUAN

Indonesia merupakan negara kepulauan dengan wilayah yang sangat luas dan jumlah penduduk mencapai 279 juta jiwa pada Desember 2023[1]. Sebanyak 43% penduduk tinggal di pedesaan dan 57% di perkotaan[2]. Data Badan Pusat Statistik (BPS) menunjukkan bahwa tenaga kesehatan cenderung terkonsentrasi di wilayah perkotaan, sementara daerah terpencil dan pedesaan masih kekurangan tenaga kesehatan dikarenakan akses yang sulit dijangkau[3]. Provinsi dengan kepadatan penduduk tinggi seperti DKI Jakarta memiliki rasio sebesar 3,2 tenaga kesehatan per 1.000 penduduk, jauh di atas provinsi seperti Papua yang hanya memiliki rasio 1,2. Provinsi lain seperti Jawa Barat memiliki rasio sebesar 1,8, Jawa Tengah (2,0), dan Jawa Timur (2,1), sedangkan Kalimantan Timur berada di angka 2,5[1]. Meskipun provinsi tersebut memiliki kepadatan penduduk yang tinggi, rasio tenaga kesehatannya masih berada di bawah standar ideal yang direkomendasikan oleh World Health Organization (WHO), yaitu minimal 2,3 tenaga kesehatan per 1.000 penduduk untuk menjamin cakupan layanan kesehatan dasar secara universal[4].

Ketidakmerataan persebaran tenaga kesehatan di Indonesia masih menjadi tantangan serius yang berkontribusi pada tingginya angka kematian di beberapa daerah akibat minimnya tenaga kesehatan dan keterbatasan obat-obatan[5]. Permasalahan ini menjadi fokus pemerintah dalam memperbaiki sistem kesehatan nasional dengan cara penambahan tenaga medis melalui program penugasan khusus yaitu program Nusantara Sehat[6]. Program ini dirancang untuk mendistribusikan tenaga kesehatan ke daerah yang membutuhkan melalui penugasan khusus guna mengisi kekosongan tenaga kesehatan di rumah sakit maupun puskesmas di wilayah terpencil.

Pendekatan dalam metode *clustering* secara umum terbagi menjadi dua, yaitu hierarchical clustering dan partitional clustering[7]. Hierarchical clustering membentuk *cluster* secara bertingkat (hierarki) tanpa perlu menentukan jumlah *cluster* di awal, sedangkan partitional clustering seperti K-Means++ mengharuskan menentukan jumlah *cluster* di awal, karena partitional clustering membagi data ke dalam *cluster* yang telah ditentukan sebelumnya. K-Means++ diperkenalkan oleh David Arthur dan Sergei Vassilvitskii dalam makalah mereka yang berjudul "k-means++: *The Advantages of Careful Seeding*" pada tahun 2007[8]. K-Means++ merupakan pengembangan dari K-Means dengan inialisasi *centroid* awal yang lebih baik, sehingga meningkatkan efisiensi komputasi. Metode ini menentukan *centroid* di awal proses berdasarkan probabilitas yang proporsional terhadap jarak kuadrat ke *centroid* terdekat yang sudah ditentukan. Sementara itu, Agglomerative Hierarchical merupakan metode *bottom-up* yang memulai proses *clustering* dengan menganggap setiap data sebagai satu *cluster* individu, lalu menggabungkan *cluster* yang paling mirip secara bertahap hingga terbentuk satu *cluster* besar dan hasilnya divisualisasikan dalam bentuk diagram dendrogram. Metode ini telah dikembangkan sejak 1950-an, dengan berbagai algoritma seperti *single linkage*, *complete linkage*, *average linkage*, hingga *Ward's method*.

Penelitian-penelitian sebelumnya menunjukkan bahwa penggunaan metode K-Means++ dan Agglomerative Hierarchical masih jarang diterapkan dalam konteks analisis penyebaran tenaga kesehatan di Indonesia. Berdasarkan hal tersebut, penelitian ini berupaya menjawab pertanyaan: Metode *clustering* mana yang lebih efektif dalam mengelompokkan provinsi berdasarkan penyebaran tenaga kesehatan di Indonesia? Hasil temuan ini diharapkan dapat menjawab pertanyaan dengan memberikan analisis komparatif yang untuk menentukan metode yang paling sesuai agar mendukung pengambilan kebijakan terkait distribusi tenaga kesehatan yang lebih tepat sasaran.

II. SIGNIFIKANSI STUDI

A. Tinjauan Literatur

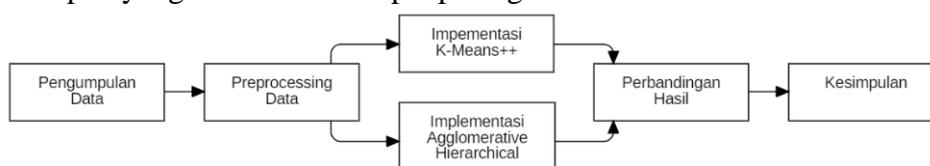
Penelitian yang dilakukan oleh Azzahra dan Wijayanto membandingkan algoritma Agglomerative Hierarchical Clustering dan K-Means dalam mengelompokkan 34 provinsi di Indonesia berdasarkan indikator pelayanan kesehatan maternal tahun 2020[9]. Hasil terbaik diperoleh dari algoritma Agglomerative Hierarchical dengan metode *average linkage*, menghasilkan lima *cluster*: *cluster* 1 terdiri dari 14 provinsi dengan kategori baik, *cluster* 2 terdiri dari 15 provinsi dengan kategori sangat baik, *cluster* 3 terdiri dari 2 provinsi dengan kategori buruk, *cluster* 4 terdiri dari 1 provinsi dengan kategori cukup, dan *cluster* 5 terdiri dari 2 provinsi dengan kategori sangat buruk. Penelitian ini membandingkan dua metode *clustering* serta mengevaluasi hasilnya menggunakan metrik Dunn Index dan Silhouette Index. Hasil dari *clustering* akan membantu pemerintah dalam memetakan wilayah prioritas sehingga dapat menjadi dasar kebijakan kesehatan yang lebih tepat sasaran.

Penelitian lain yang dilakukan oleh Ghanim Ramadhan dan Yuli Astuti membandingkan kinerja algoritma K-Means dan Agglomerative Clustering dalam segmentasi pasar untuk penjualan *online* [10]. Penelitian ini menggunakan dua metode algoritma untuk menganalisis segmentasi data pelanggan retail yang berjumlah 541.909 data, kemudian dikelompokkan menjadi tiga *cluster* berdasarkan pembobotan variabel RFM (*Recency, Frequency, Monetary*). Hasil penelitian menunjukkan bahwa K-Means menghasilkan nilai Silhouette Score sebesar 0,5087, sementara Agglomerative menghasilkan nilai lebih tinggi yaitu 0,6363, yang menunjukkan bahwa Agglomerative lebih efektif dalam menentukan segmentasi pasar. penelitian ini memberikan kontribusi terhadap pengembangan studi komparatif dalam algoritma *unsupervised learning*, khususnya dalam konteks segmentasi pasar berbasis *big data*. Penelitian ini memberikan panduan bagi pelaku bisnis *online* dalam menentukan strategi pemasaran yang lebih tepat sasaran, yaitu dengan memanfaatkan metode *clustering* yang lebih akurat dalam mengelompokkan konsumen berdasarkan perilaku belanja.

Berdasarkan penelitian-penelitian terdahulu, dapat disimpulkan bahwa teknik *clustering* menggunakan K-Means++ belum tentu efektif dalam memetakan *cluster*. Hal itu juga berlaku untuk Agglomerative Hierarchical. K-Means++ cenderung lebih efisien dalam menangani dataset besar dan dapat menghasilkan *cluster* yang lebih baik, namun terkadang menjadi tidak efektif jika jumlah *cluster* yang diinginkan tidak sesuai dengan distribusi data. Sedangkan, Agglomerative Hierarchical unggul dalam pemahaman terhadap struktur hierarki data, meskipun cenderung lebih lambat pada dataset yang besar. Selain itu, faktor-faktor lain seperti pemilihan jumlah *cluster*, skala data, dan jenis fitur juga mempengaruhi keakuratan hasil dari metode *clustering*. Oleh karena itu, pemilihan metode *clustering* terbaik sangat bergantung pada karakteristik dataset dan tujuan analisis yang akan dilakukan.

B. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan membandingkan metode K-Means++ dan Agglomerative Hierarchical untuk mengidentifikasi wilayah dengan keterbatasan tenaga kesehatan. Tahapan yang dilakukan terdapat pada gambar dibawah ini.



Gambar 1. Metode Penelitian

1. Pengumpulan Data

Pada penelitian ini data yang digunakan adalah data Jumlah Tenaga Kesehatan Menurut Provinsi 2023 yang merupakan data sekunder BPS (Badan Pusat Statistik)[11]. Data tersebut mencakup 38 Provinsi di Indonesia dengan jumlah tenaga kesehatan di berbagai kategori yaitu perawat, bidan, tenaga kefarmasian, tenaga kesehatan masyarakat, tenaga kesehatan lingkungan, tenaga gizi, ahli teknologi laboratorium medik, tenaga medis, kesehatan psikologi klinis, keterampilan fisik, keteknisan medis, tenaga teknik biomedika, dan tenaga kesehatan tradisional.

2. Preprocessing Data

Tahap pertama dalam penelitian ini adalah *data cleansing*, yang mencakup beberapa langkah penting untuk memastikan kualitas, kelengkapan, dan konsistensi data sebelum dilakukan analisis lebih lanjut. Langkah awal pada tahap ini adalah menangani *missing value*, *missing value* dalam dataset ini diartikan sebagai tidak tersedianya tenaga kesehatan di provinsi tertentu. Oleh karena itu, nilai *null* diisi dengan angka 0 karena merepresentasikan ketiadaan tenaga kesehatan di suatu provinsi, bukan akibat kesalahan pencatatan. Selain itu, kolom "Ahli Teknologi Laboratorium Medik" dihapus karena tidak mengandung data dan tidak memberikan kontribusi terhadap hasil analisis dalam penelitian ini. Pendekatan ini dipilih untuk mempertahankan makna asli dari data dan menghindari potensi distorsi dalam proses analisis yang sensitif terhadap keberadaan nilai *null*. Setelah itu, dilakukan proses standarisasi data yaitu dengan menyamakan skala seluruh variabel agar setiap fitur memiliki kontribusi yang setara dalam proses perhitungan jarak, terutama pada algoritma K-Means++ dan Agglomerative Hierarchical yang sensitif terhadap perbedaan skala antar variabel. Tahapan *data cleansing* kemudian dilanjutkan dengan analisis statistik deskriptif, yang bertujuan untuk memahami karakteristik awal dari data dengan menghitung nilai-nilai penting seperti mean, median, dan standar deviasi, sehingga memberikan gambaran umum mengenai sebaran dan distribusi data Jumlah Tenaga Kesehatan menurut Provinsi di Indonesia.

3. Implementasi K-Means++

K-Means++ dikembangkan untuk meningkatkan kualitas proses *clustering* dengan menentukan *centroid* awal secara lebih optimal dibandingkan metode inisialisasi acak pada K-Means. Pemilihan *centroid* yang lebih strategis mampu mempercepat proses serta meningkatkan akurasi hasil *clustering*. Penentuan jumlah *cluster* yang optimal dilakukan menggunakan Elbow Method dan Silhouette Score. Implementasi algoritma ini juga dilakukan menggunakan Python, yaitu dengan menerapkan *library* Scikit-learn. Pada tahap ini, penentuan *centroid* tidak dilakukan secara manual melainkan secara otomatis dibantu oleh *library* Python menggunakan pendekatan *machine learning*.

Penentuan jumlah *cluster* yang optimal dilakukan menggunakan Elbow Method dan Silhouette Score. Elbow Method membantu menentukan jumlah *cluster* dengan melihat titik di mana penurunan nilai error mulai melambat secara signifikan dan membentuk pola seperti siku yang menandakan jumlah *cluster* ideal. Sementara itu, Silhouette Score juga digunakan untuk mengukur sejauh mana titik data berada dalam *cluster* yang sesuai.

4. Implementasi Agglomerative Hierarchical

Agglomerative Hierarchical adalah sebuah pendekatan luas dalam metode *clustering* yang membentuk hirarki *cluster* dengan menempatkan setiap data sebagai *cluster* individu, kemudian secara bertahap menggabungkan *cluster-cluster* yang memiliki kemiripan tinggi hingga terbentuk *cluster* yang lebih besar dalam suatu struktur hierarki[7]. Agglomerative Hierarchical adalah *subcategory* dari Hierarchical Clustering yang menggunakan pendekatan *bottom up* dengan komputasi yang lebih sederhana dan *linkage method* yang digunakan antara lain yaitu *single*, *complete*, *ward*, and *centroid*.

Agglomerative Hierarchical merupakan metode *clustering* yang cocok digunakan untuk dataset berukuran kecil hingga sedang, terutama ketika terdapat *outlier* ekstrem seperti yang terdapat pada dataset Jumlah Tenaga Kesehatan menurut Provinsi [12], di mana terdapat provinsi-provinsi dengan jumlah tenaga kesehatan yang jauh lebih tinggi atau lebih rendah jika dibandingkan dengan provinsi lainnya. Kelebihan utama metode ini terletak pada fleksibilitasnya, jumlah *cluster* tidak perlu ditentukan di awal dan dapat memilih jumlah *cluster* yang diinginkan yaitu dengan memotong dendrogram pada level tertentu sesuai kebutuhan analisis. Metode ini juga tidak bergantung pada inisialisasi *centroid* sehingga menghasilkan pengelompokan yang lebih stabil dan konsisten. Agglomerative Hierarchical memiliki kompleksitas waktu dan ruang yang tinggi sekitar $O(m^2)$, sehingga kurang cocok untuk dataset yang berukuran besar. Agglomerative Hierarchical diterapkan dengan menghitung jarak antar provinsi menggunakan *euclidean distance*. Proses *clustering* dilakukan dengan metode *Ward's minimum variance* untuk meminimalkan variansi intra-*cluster*. Proses ini dimulai dari setiap provinsi sebagai *cluster* individu, kemudian menggabungkan *cluster* terdekat secara bertahap hingga seluruh data tergabung dalam satu hierarki. Hasil dari proses *clustering* divisualisasikan dalam bentuk dendrogram yang menampilkan struktur *cluster* secara hierarkis.

5. Perbandingan Hasil

Tahap ini adalah membandingkan hasil *clustering* dari kedua metode untuk menentukan mana yang lebih efektif dalam mengelompokkan data berdasarkan karakteristik jumlah tenaga kesehatan di setiap provinsi. Proses evaluasi ini dilakukan dengan menganalisis sejauh mana masing-masing metode mampu membentuk *cluster* yang optimal dan mewakili pola distribusi data secara akurat, serta menggunakan beberapa metrik validasi untuk menilai kualitas *clustering*. Silhouette Score digunakan untuk melihat sejauh mana setiap variabel cocok dengan *cluster* tempatnya berada dibandingkan dengan *cluster* lain, nilai yang mendekati 1 menunjukkan bahwa hasil *clustering* semakin baik. Calinski-Harabasz Index mengukur rasio antara variansi antar *cluster* dengan variansi di dalam *cluster*, dan semakin tinggi nilainya akan semakin jelas pemisahan antar *cluster*. Terakhir adalah Davies-Bouldin Index untuk mengukur rata-rata kemiripan antar *cluster*, dan nilai yang lebih rendah menunjukkan bahwa *cluster* yang terbentuk lebih terpisah dan kompak. Kombinasi dari ketiga metrik ini memberikan gambaran menyeluruh mengenai performa dan efektivitas metode *clustering* yang digunakan [13].

6. Kesimpulan

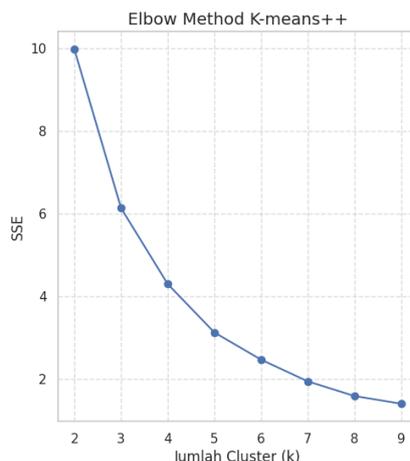
Tahap ini merupakan tahap menyimpulkan hasil dari perbandingan *clustering* K-Means++ dan Agglomerative Hierarchical dengan mengkaji performa masing-masing algoritma berdasarkan 3 metrik evaluasi.

III. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk membandingkan dua algoritma *clustering*, yaitu K-Means++ dan Agglomerative Hierarchical guna mengetahui algoritma mana yang paling efektif dalam mengelompokkan data. Setelah melalui tahapan *preprocessing data*, selanjutnya dilakukan implementasi kedua algoritma tersebut untuk dianalisis dan dibandingkan hasilnya.

1. K-Means++

Penentuan jumlah *centroid* pada algoritma K-Means++ dilakukan menggunakan Elbow Method yang didasarkan pada titik tekukan (*elbow point*), yaitu titik di mana penurunan nilai *Sum of Squared Errors* (SSE) mulai melambat secara signifikan.



Gambar 1. Elbow Method K-Means++

Visualisasi tersebut menunjukkan grafik Elbow untuk menentukan jumlah *cluster* optimal dalam algoritma K-Means++, dengan menggunakan *Sum of Squared Errors* (SSE) terhadap jumlah *cluster*. Sumbu x berisikan jumlah *cluster* (k) dan sumbu Y berisi SSE, artinya jika nilainya semakin rendah maka semakin kecil varians dalam *cluster*. Terlihat bahwa terjadi penurunan tajam dari k=2 ke k=3 dan masih cukup signifikan hingga k=4. Setelah k=5, penurunan SSE mulai landai yang artinya penambahan *cluster* tidak memberikan peningkatan signifikan dalam mengurangi varians internal. Sehingga bisa dikatakan jumlah *cluster* yang ideal adalah 4 *cluster*.

Evaluasi performa hasil *clustering* dilakukan menggunakan tiga metrik, yaitu Silhouette Score, Calinski-Harabasz, dan Davies-Bouldin. Hasil evaluasi menunjukkan nilai Silhouette Score sebesar 0,495; Calinski-Harabasz sebesar 63,630; dan Davies-Bouldin sebesar 0,585. Nilai yang lebih tinggi pada Silhouette Score dan Calinski-Harabasz menunjukkan kualitas *clustering* yang lebih baik, sedangkan nilai yang lebih rendah pada Davies-Bouldin Index menandakan hasil *clustering* yang lebih optimal. Perbedaan performa antar ketiga metrik ini disebabkan oleh aspek yang diukur serta sensitivitas terhadap karakteristik data yang berbeda. Setelah evaluasi dengan tiga metrik, selanjutnya adalah implemmentasi algoritma K-Means++ menggunakan *library* Scikit-learn. Jumlah *cluster* optimal sebanyak empat, yaitu: *cluster* 0 (sangat rendah), *cluster* 1 (rendah), *cluster* 2 (sedang), dan *cluster* 3 (tinggi). Rincian hasil pengelompokan K-Means++ disajikan pada tabel berikut.

TABEL I
HASIL CLUSTERING K-MEANS++

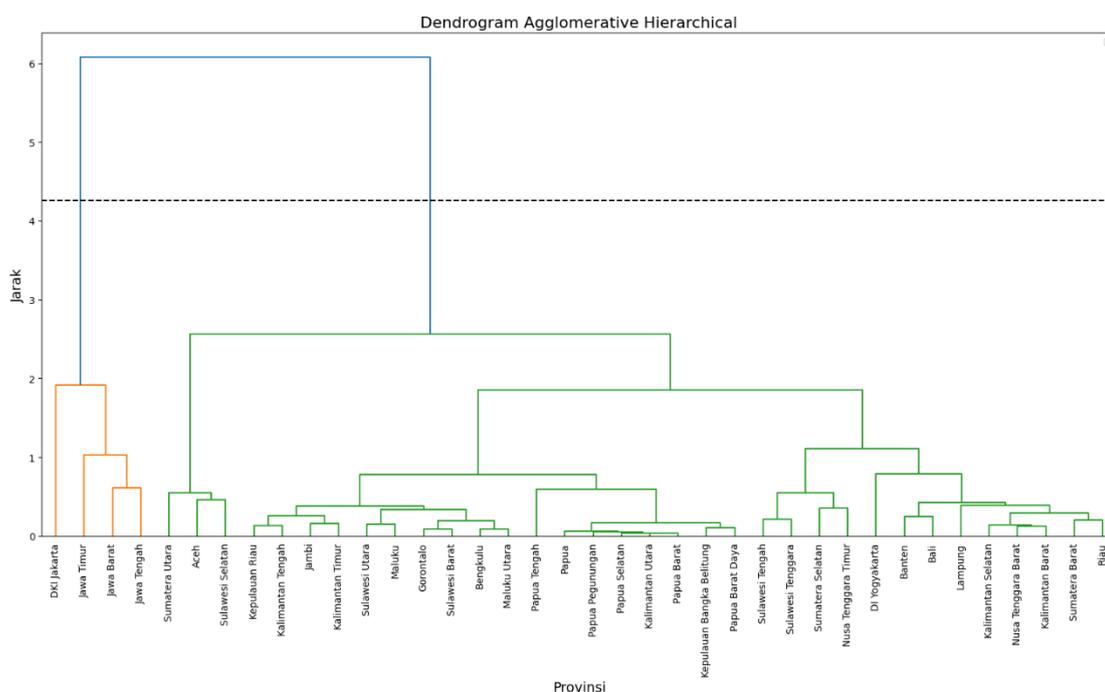
| DI | Cluster | Keterangan | Provinsi | Jumlah |
|----|---------|---------------|---|--------|
| | 0 | Sangat Rendah | Sumatera Barat, Riau, Jambi, Bengkulu, Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, DI Yogyakarta, Banten, Bali, Nusa Tenggara Barat, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Papua Barat Daya, Papua, Papua Selatan, Papua Tengah, Papua Pegunungan. | 27 |
| | 1 | Rendah | Jawa Barat, Jawa Tengah, Jawa Timur. | 3 |
| | 2 | Sedang | Aceh, Sumatera Utara, Sumatera Selatan, Nusa Tenggara Timur, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara. | 7 |
| | 3 | Tinggi | DKI Jakarta | 1 |

Yogyakarta termasuk dalam *cluster* 0 (sangat rendah), meskipun secara geografis provinsi tersebut memiliki akses yang relatif mudah dijangkau. Kemungkinan hal ini bisa terjadi disebabkan oleh

ketidakseimbangan distribusi tenaga kesehatan berdasarkan kategori. Ketidakseimbangan ini dapat mencakup distribusi tenaga medis, perawat, atau bidan yang tidak proporsional, sehingga meskipun secara total jumlah tenaga kesehatan mencukupi, kekurangan pada kategori tertentu dapat memengaruhi hasil clustering.

2. Agglomerative Hierarchical

Penentuan jumlah *cluster* optimal pada algoritma Agglomerative Hierarchical dilakukan dengan pendekatan yang berbeda dari Elbow Method yang digunakan pada K-Means++, yaitu menggunakan diagram dendrogram. Dendrogram menampilkan struktur hierarki penggabungan antar *cluster* dan jumlah *cluster* optimal ditentukan dengan mengamati lompatan jarak terbesar (*distance jump*) yang menunjukkan titik pemisahan *cluster* yang paling signifikan[14].



Gambar 2. Diagram Dendrogram Agglomerative Hierarchical

Pada diagram dendrogram, garis vertikal mewakili masing-masing provinsi. Sumbu Y (jarak) merepresentasikan ukuran ketidakmiripan antar *cluster*, di mana nilai yang lebih tinggi menandakan perbedaan yang lebih besar. Garis berwarna biru, oranye, dan hijau menunjukkan kelompok utama yang terbentuk pada tingkat hierarki tertentu.

Selanjutnya mengevaluasi performa hasil *clustering* dengan menggunakan tiga metrik, yaitu Silhouette Score sebesar 0,550; Calinski-Harabasz Index sebesar 55,111; dan Davies-Bouldin Index sebesar 0,457. Nilai yang lebih tinggi menunjukkan kualitas *clustering* yang lebih baik untuk metrik Silhouette dan Calinski-Harabasz, sedangkan nilai yang lebih rendah lebih baik untuk Davies-Bouldin. Setelah menentukan jumlah *cluster* optimal, algoritma Agglomerative Hierarchical diimplementasikan dan rincian hasil *clustering* Agglomerative Hierarchical dapat dilihat pada tabel berikut.

TABEL II
HASIL CLUSTERING AGGLOMERATIVE HIERARCHICAL

| Cluster | Keterangan | Provinsi | Jumlah |
|---------|---------------|--|--------|
| 0 | Sangat Rendah | Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, 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 Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Papua Barat Daya, Papua, Papua Selatan, Papua Tengah, Papua Pegunungan. | 31 |
| 1 | Rendah | Jawa Barat, Jawa Tengah, Jawa Timur. | 3 |
| 2 | Sedang | Aceh, Sumatera Utara, Sulawesi Selatan. | 3 |
| 3 | Tinggi | DKI Jakarta | 1 |

Hasil *clustering* menunjukkan adanya pola geografis yang kuat, di mana sebagian besar provinsi dalam *cluster* 0 (Sangat Rendah) berasal dari wilayah Indonesia Timur dan luar Jawa, mencerminkan keterbatasan akses geografis terhadap tenaga kesehatan. Sebaliknya, *cluster* 1 (Rendah) berisi provinsi dengan populasi besar seperti Jawa Barat, Jawa Tengah, dan Jawa Timur, yang meskipun infrastrukturnya memadai, masih menghadapi tantangan dalam rasio tenaga kesehatan akibat tingginya jumlah penduduk. DI Yogyakarta dan Bali yang masuk dalam *cluster* 0 meskipun dikenal memiliki layanan kesehatan yang baik. Hal ini mengindikasikan adanya ketimpangan distribusi internal. DKI Jakarta, satu-satunya provinsi di *cluster* 3 (Tinggi), mencerminkan konsentrasi tenaga kesehatan di ibu kota. Pola ini dapat menjadi dasar kebijakan distribusi tenaga kesehatan yang lebih tepat sasaran.

3. Perbandingan Hasil

Tahap ini adalah membandingkan hasil *clustering* yang diperoleh dari algoritma K-Means++ dan Agglomerative Hierarchical. Perbandingan ini dilakukan berdasarkan tiga metrik evaluasi, yaitu Silhouette Score, Calinski-Harabasz Index, dan Davies-Bouldin Index, untuk menilai kualitas dan efektivitas masing-masing metode dalam *clustering*.

TABEL III
HASIL EVALUASI CLUSTERING

| Metode | Silhouette Score | Calinski Harabasz | Davies-Bouldin |
|----------------------------|------------------|-------------------|----------------|
| K-Means++ | 0.495 | 63.630 | 0.585 |
| Agglomerative Hierarchical | 0.550 | 55.111 | 0.457 |

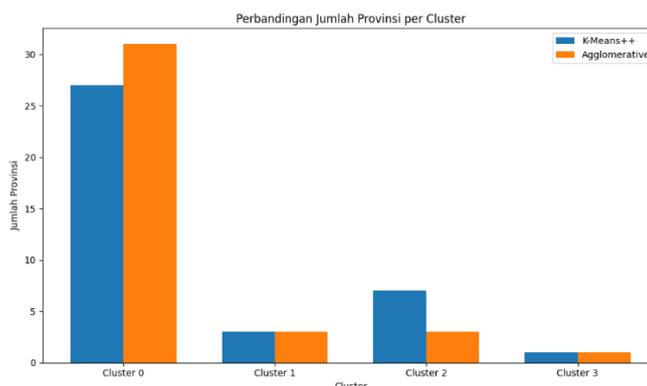
Berdasarkan hasil evaluasi *clustering* pada Tabel 3, metrik Silhouette Score pada Agglomerative Hierarchical menunjukkan performa yang lebih baik dengan skor 0,550, dibandingkan dengan K-Means++ yang memperoleh skor 0,495. Hal ini menunjukkan bahwa *cluster* yang dihasilkan oleh Agglomerative Hierarchical lebih kompak dan lebih terpisah dengan baik dibandingkan dengan K-Means++. Selanjutnya, pada metrik Calinski-Harabasz Index, K-Means++ unggul dengan nilai 63,630, lebih tinggi dari Agglomerative Hierarchical yang hanya mencapai 55,111. Ini mengindikasikan bahwa K-Means++ lebih baik dalam memisahkan *cluster* secara global berdasarkan sebaran data. Namun, pada metrik Davies-Bouldin Index, Agglomerative Hierarchical kembali menunjukkan performa yang lebih baik dengan nilai 0,457, yang lebih rendah dibandingkan dengan K-Means++ yang memiliki nilai 0,585. Karena nilai yang lebih rendah menunjukkan kualitas *cluster* yang lebih baik. Secara keseluruhan dapat disimpulkan bahwa, Agglomerative Hierarchical lebih unggul dalam membentuk *cluster* yang lebih representatif dan stabil pada dataset Jumlah Tenaga Kesehatan Menurut Provinsi.

Perbedaan performa antar tiga metrik disebabkan oleh perbedaan pendekatan penilaian masing-masing metrik. Silhouette Score menilai seberapa mirip suatu objek dengan klusternya dibandingkan dengan kluster lain, sehingga sensitif terhadap jarak antar objek[15]. Calinski-Harabasz berfokus pada rasio dispersi antar-*cluster* dan dalam-*cluster*, yang berarti dipengaruhi oleh kompaknya *cluster* dan pemisahan yang jelas antar *cluster*. Sementara itu, Davies-Bouldin mempertimbangkan dengan baik penyebaran internal tiap *cluster* maupun kedekatan antar *cluster*, sehingga lebih sensitif terhadap tumpang tindih antar kelompok data. Oleh karena itu, perbedaan struktur data seperti jumlah *cluster*, distribusi data, dan sebaran dalam *cluster* dapat memengaruhi nilai dari ketiga metrik.

TABEL IIV
PERBEDAAN KLASIFIKASI HASIL CLUSTERING

| Provinsi | K-Means++ | Keterangan | Agglomerative Hierarchical | Keterangan |
|---------------------------|-----------|---------------|----------------------------|---------------|
| Aceh | Cluster 2 | Sedang | Cluster 2 | Sedang |
| Sumatera Utara | Cluster 2 | Sedang | Cluster 2 | Sedang |
| Sumatera Barat | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Riau | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Jambi | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Sumatera Selatan | Cluster 2 | Sedang | Cluster 0 | Sangat Rendah |
| Bengkulu | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Lampung | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Kepulauan Bangka Belitung | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Kepulauan Riau | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| DKI Jakarta | Cluster 3 | Tinggi | Cluster 3 | Tinggi |
| Jawa Barat | Cluster 1 | Rendah | Cluster 1 | Rendah |
| Jawa Tengah | Cluster 1 | Rendah | Cluster 1 | Rendah |
| DI Yogyakarta | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Jawa Timur | Cluster 1 | Rendah | Cluster 1 | Rendah |
| Banten | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Bali | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Nusa Tenggara Barat | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Nusa Tenggara Timur | Cluster 2 | Sedang | Cluster 0 | Sangat Rendah |
| Kalimantan Barat | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Kalimantan Tengah | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Kalimantan Selatan | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Kalimantan Timur | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Kalimantan Utara | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Sulawesi Utara | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Sulawesi Tengah | Cluster 2 | Sedang | Cluster 0 | Sangat Rendah |
| Sulawesi Selatan | Cluster 2 | Sedang | Cluster 2 | Sedang |
| Sulawesi Tenggara | Cluster 2 | Sedang | Cluster 0 | Sangat Rendah |
| Gorontalo | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Sulawesi Barat | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Maluku | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Maluku Utara | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Papua Barat | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Papua Barat Daya | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Papua | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Papua Selatan | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Papua Tengah | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |
| Papua Pegunungan | Cluster 0 | Sangat Rendah | Cluster 0 | Sangat Rendah |

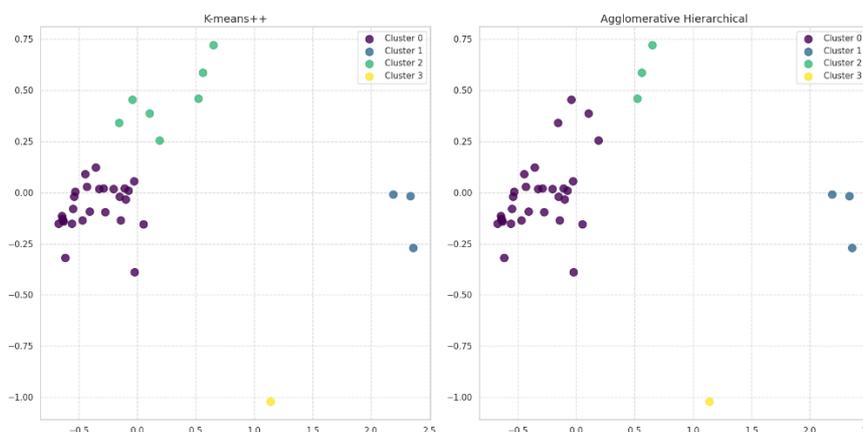
Persentase kecocokan hasil *cluster* antar metode diperoleh nilai sebesar 89,47%, menunjukkan adanya tingkat kesamaan yang cukup tinggi. Provinsi DI Yogyakarta masuk dalam *cluster* 0 (sangat rendah), meskipun secara geografis provinsi tersebut memiliki akses yang relatif mudah dijangkau. Faktor yang menyebabkan hal ini terjadi adalah ketidakseimbangan distribusi tenaga kesehatan antar kategori. Ketidakseimbangan ini dapat mencakup distribusi tenaga medis, perawat, atau bidan yang tidak proporsional, meskipun secara total jumlah tenaga kesehatan mencukupi, ketimpangan distribusi dalam satu kategori tertentu dapat berpengaruh terhadap hasil *clustering*. Faktor penyebab perbedaan hasil antar metode juga berasal dari pendekatan dasar algoritma yang digunakan.



Gambar 3. Bar Chart Perbandingan Jumlah Provinsi

Grafik tersebut menunjukkan perbandingan jumlah provinsi setiap *cluster* dalam bentuk *bar chart* antara K-Means++ dan Agglomerative Hierarchical. Mayoritas wilayah dikelompokkan ke dalam *cluster* 0 oleh kedua metode, yang mengindikasikan bahwa distribusi tenaga kesehatan di sebagian besar daerah masih tergolong sangat rendah. Hal ini banyak terjadi pada wilayah dengan cakupan geografis luas dengan keterbatasan akses, yang menyebabkan persebaran tenaga kesehatan tidak merata.

Perbedaan mencolok terlihat pada *cluster* 2, di mana K-Means++ mengelompokkan lebih banyak wilayah dibandingkan Agglomerative. Ini dapat disebabkan oleh sensitivitas K-Means++ terhadap *centroid*, sehingga beberapa wilayah yang memiliki jumlah tenaga kesehatan mendekati rata-rata nasional masuk ke dalam *cluster* 2, meskipun distribusinya belum merata. Sebaliknya, metode Agglomerative mempertimbangkan kedekatan spasial dan struktur hierarki akan lebih ketat dalam menyaring provinsi yang memiliki ketimpangan signifikan, sehingga jumlah wilayah dalam *cluster* 2 lebih sedikit.



Gambar 4. Scatter Plot K-Means++ dan Agglomerative Hierarchical

Gambar tersebut adalah visualisasi scatter plot perbandingan hasil *clustering* antara K-Means++ dan Agglomerative Hierarchical. Kedua metode sama-sama mengelompokkan mayoritas wilayah

ke dalam satu *cluster* besar, yang mencerminkan rendahnya distribusi tenaga kesehatan secara nasional. K-Means++ menghasilkan lebih banyak variasi *cluster* karena sensitif terhadap *centroid*, sehingga wilayah dengan jumlah total tenaga kesehatan yang mendekati rata-rata nasional dikelompokkan secara terpisah. Sementara itu, Agglomerative Hierarchical lebih selektif dan mempertimbangkan kedekatan struktur data, sehingga hanya wilayah dengan perbedaan yang sangat mencolok akan dikelompokkan ke dalam *cluster* berbeda. Salah satu wilayah yang secara geografis memiliki akses baik tetap masuk dalam *cluster* dengan kategori sangat rendah, ketimpangan tersebut kemungkinan karena distribusi antar jenis tenaga kesehatan. Hal ini menunjukkan bahwa kedua metode memiliki pendekatan berbeda dalam mengenali karakteristik wilayah, dan perbedaan hasil *clustering* dapat memberikan gambaran penting dalam merancang kebijakan yang tepat sasaran.

IV. KESIMPULAN

Persentase kecocokan hasil *cluster* antar metode diperoleh nilai sebesar 89,47%, menunjukkan adanya tingkat kesamaan yang cukup tinggi, perbedaan mendasar antara kedua metode terletak pada pendekatannya. Agglomerative Hierarchical membentuk *cluster* secara bertahap melalui proses penggabungan antar *cluster* terdekat dalam struktur hierarki, sedangkan K-Means++ berupaya menemukan *centroid* awal yang optimal untuk meningkatkan kualitas *cluster* dari iterasi awal. Distribusi hasil pengelompokan juga menunjukkan perbedaan. Pada *cluster* 0, Agglomerative mencakup 31 provinsi sementara K-Means++ mencakup 27 provinsi. Beberapa provinsi seperti Sumatera Selatan, NTT, dan Sulawesi Tengah masuk dalam *cluster* berbeda antara kedua metode. Namun, terdapat kesamaan dalam *cluster* 1 dan *cluster* 3 yang masing-masing berisi provinsi yang sama yaitu Jawa Barat, Jawa Tengah, Jawa Timur di *cluster* 1 dan DKI Jakarta di *cluster* 3. Berdasarkan hasil *clustering*, penggunaan metode Agglomerative Hierarchical disarankan ketika tujuan utamanya adalah mendapatkan struktur *cluster* yang lebih stabil secara hierarki, terutama untuk analisis kebijakan berbasis wilayah. K-Means++ tetap menjadi pilihan yang efisien terutama untuk data berukuran besar atau dalam proses eksploratif awal. Selain itu, disarankan untuk memperluas variabel yang digunakan dalam *clustering*, seperti rasio tenaga kesehatan terhadap jumlah penduduk, ketersediaan fasilitas kesehatan, atau indikator pelayanan kesehatan lainnya agar hasil pengelompokan semakin akurat dan relevan untuk mendukung perencanaan kebijakan distribusi tenaga kesehatan yang lebih tepat sasaran.

REFERENSI

- [1] Badan Pusat Statistik, “Jumlah Penduduk Pertengahan Tahun (Ribu Jiwa), 2022-2024,” bps.go.id. [Online]. Available: <https://www.bps.go.id/id/statistics-table/2/MTk3NSMy/jumlah-penduduk-pertengahan-tahun--ribu-jiwa-.html>
- [2] Indonesia Investments, “Penduduk Indonesia,” 2024. [Online]. Available: <https://www.indonesia-investments.com/id/budaya/penduduk/item67>
- [3] N. Ahmad, “Ketidakmerataan Fasilitas Kesehatan dan Tenaga Kesehatan di Indonesia Konten ini telah tayang di Kompasiana.com dengan judul ‘Ketidakmerataan Fasilitas Kesehatan dan Tenaga Kesehatan di Indonesia’, Klik untuk baca: <https://www.kompasiana.com/nabhan2104/64>,” Kompasiana.
- [4] Health Workforce (HWF), *The world health report 2006: working together for health*. 2006.
- [5] Kemenkes, *Profil Kesehatan Indonesia 2023*. 2023.
- [6] Rokom, “Kemenkes Penuhi Kebutuhan Nakes di Puskesmas Lewat Penugasan Khusus,” Redaksi Sehat Negeriku. [Online]. Available: <https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20230214/0242408/kemenkes-penuhi-kebutuhan-nakes-di-puskesmas-lewat-penugasan-khusus/>
- [7] P. Asehnino and A. W. Wijayanto, “Analisis Perbandingan Metode Hierarchical dan Non-Hierarchical dalam Pembentukan Cluster Provinsi di Indonesia Berdasarkan Indikator Women Empowerment,” *Indones. J. Appl. Stat.*, vol. 6, no. 1, p. 57, 2024, doi: 10.13057/ijas.v6i1.68876.
- [8] D. Arthur and S. Vassilvitskii, “SODA ’07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms,” in *K-means++: The Advantages of Careful Seeding*, 2007, pp. 1027–1035.
- [9] A. Azzahra and A. W. Wijayanto, “Perbandingan Agglomerative Hierarchical dan K-Means Dalam Pengelompokan Provinsi Berdasarkan Pelayanan Kesehatan Maternal,” *Sist. J. Sist. Inf.*, vol. 11, no. 2, pp. 481–495, 2022, [Online]. Available: [http://download.garuda.kemdikbud.go.id/article.php?article=2929100%5C&val=10006%5C&title=Comparison of Agglomerative Hierarchical and K-Means in Grouping Provinces Based on Maternal Health Services](http://download.garuda.kemdikbud.go.id/article.php?article=2929100%5C&val=10006%5C&title=Comparison%20of%20Agglomerative%20Hierarchical%20and%20K-Means%20in%20Grouping%20Provinces%20Based%20on%20Maternal%20Health%20Services)
- [10] G. Ramadhan and Y. Astuti, “Perbandingan Kinerja Algoritma K-means dan Agglomerative Clustering Untuk Segmentasi Penjualan Online Pada Customer Retail,” *J. Pengemb. IT*, vol. 9, no. 1, pp. 92–96, 2024.
- [11] Badan Pusat Statistik, “Jumlah Tenaga Kesehatan Menurut Provinsi, 2023,” bps.go.id.
- [12] L. Sulistina, S. Syaharuddin, M. Ibrahim, and ..., “Perbedaan tingkat akurasi metode k-means dan hierarchical clustering di bidang peramalan dan klasifikasi,” *Semin. Nas. ...*, vol. 1, 2022, [Online]. Available: <http://journal.ummat.ac.id/index.php/semnaslppm/article/view/9142%0Ahttp://journal.ummat.ac.id/index.php/semnaslppm/article/download/9142/4771>
- [13] U. F. Laili, C. Umatin, and M. U. Ridwanulloh, “Analisis Potensial Drop Out Mahasiswa Dengan K-Means++ Clustering Dalam Upaya Peningkatan Kualitas IAIN Kediri,” *Paedagogia J. Kajian, Penelit. dan Pengemb. Kependidikan*, vol. 14, no. 2, pp. 145–153, 2023, doi: 10.31764.
- [14] I. Indra, N. Nur, M. Iqram, and N. Inayah, “Perbandingan K-Means dan Hierarchical Clustering dalam Pengelompokan Daerah Beresiko Stunting,” *INOVTEK Polbeng - Seri Inform.*, vol. 8, no. 2, p. 356, 2023, doi: 10.35314/isi.v8i2.3612.
- [15] M. R. Syahkur and D. Hartama, “Evaluasi Jumlah Cluster pada Algoritma K-Means ++ Menggunakan Silhouette dan Elbow dengan Validasi Nilai DBI dalam Mengelompokkan Gizi Balita,” vol. 13, no. 3, pp. 487–496, 2024.