
Analisis Perbandingan Metode *Hierarchical* dan *Non-Hierarchical* dalam Pembentukan *Cluster* Provinsi di Indonesia Berdasarkan Indikator *Women Empowerment*

Pikata Aselnino*, Arie Wahyu Wijayanto
Politeknik Statistika STIS

*Corresponding author: pikata.aselnino@bps.go.id

Abstract. *The focus on improving the quality of women's live to lessen discrimination and gender inequality is set in the fifth's goals of SDGs. In Indonesia, the RPJMN 2020-2024 contains measure to improve the contribution of women to equitable development. The Central Bureau of Statistics has developed several indicators related to gender, including Gender Development Index (GDI) an Gender Empowerment Index (GEI), which contain women's improvement on education and health as well as their participation in economic and political fields. The Ministry of Women's Empowerment and Child Protection did a quadrant analysis to split Indonesia's 34 provinces into four categories based solely on GDI and GEI using the national average as a constraint. This study compares the Hierarchical, K-Means, and Fuzzy C-Means method to form number of clusters in Indonesia based on the gender development and empowerment in 2021 in order to complement the quadrant analysis. To choose the number of optimum cluster, Elbow method and Calinski-Harabasz Index were used and the best k value is five. From the validation with Silhoutte Index, K-Means was chosen as the best clustering model.*

Keywords: *clustering; fuzzy; k-means; hierarchical; women empowerment*

1. PENDAHULUAN

Seiring dengan perkembangan zaman, program untuk membangun dan memberdayakan perempuan semakin gencar dilakukan. Mulai tahun 2000, pemerintah Indonesia mengeluarkan instruksi terkait Pengarusutamaan Gender (PUG) sebagai strategi untuk mengurangi kesenjangan antara perempuan dan laki-laki dalam mengakses dan memperoleh manfaat pembangunan. Selanjutnya, pada tahun 2010, pemerintah mengeluarkan instruksi terkait Program Pembangunan dan Berkeadilan yang diintegrasikan pada *Millennium Development Goals* (MDGs) dan *Sustainable Development Goals* (SDGs). Selain itu, dalam Rencana Pembangunan dan Rencana Menengah Nasional (RPJMN) tahun 2020-2024, tercakup kebijakan untuk meningkatkan peran perempuan dalam pembangunan, pendidikan, kesehatan, ekonomi, tenaga kerja, politik, jabatan publik, dan pengambilan keputusan.

Seluruh target capaian yang dicanangkan tentu diterapkan ke seluruh wilayah di Indonesia. Badan Pusat Statistik (BPS) mengeluarkan nilai-nilai indeks menggambarkan perkembangan pembangunan dan pemberdayaan perempuan lewat Indeks Pembangunan Gender (IPG) dan Indeks Pemberdayaan Gender (IDG) hingga level kota/kabupaten setiap provinsi di Indonesia. BPS menyatakan bahwa dalam tingkatan nasional, pada tahun 2021, pembangunan perempuan

Indonesia adalah 91,27% yang mana lebih tinggi dibandingkan tahun 2020, yaitu 91,06%. Angka tersebut menunjukkan bahwa capaian pembangunan perempuan sudah hampir setara dengan pembangunan laki-laki. Skor pemberdayaan perempuan secara nasional juga mengalami peningkatan dari tahun 2020 menuju 2021, yaitu dari 75,57 menjadi 76,26. Secara umum, angka-angka tersebut tinggi. Akan tetapi, ketika ditilik hingga wilayah kota/kabupaten, masih terdapat beragam nilai yang menunjukkan adanya ketimpangan.

Kementerian Pemberdayaan Perempuan dan Perlindungan Anak telah melakukan analisis kuadran untuk membagi provinsi-provinsi di Indonesia menjadi empat kelompok berdasarkan keterkaitan antara IPM dan IPG serta keterkaitan IPG dan IDG pada tahun 2020. Batasan pembuatan gugus yang digunakan adalah nilai IPM, IPG, dan IDG nasional. Uraian kebijakan yang dibuat untuk mengatasi ketimpangan berakhir hanya sebatas pada tingkat provinsi saja. Padahal di dalam provinsi-provinsi itu sendiri, masih juga terjadi ketimpangan pencapaian pembangunan dan pemberdayaan perempuan. Pembentukan *cluster* pada tingkat provinsi dengan memasukkan indikator-indikator tersebut belum banyak dilakukan.

Diana [1] telah melakukan analisis ketimpangan gender di Sumatera Barat berdasarkan tiga variabel pembentuk IDG, yaitu persentase keterlibatan perempuan di parlemen, perempuan sebagai tenaga profesional, dan sumbangan pendapatan perempuan. Perbandingan capaian antar kabupaten dilakukan lewat analisis *biplot* dan *Agglomerative Hierarchical Clustering* dengan membentuk tiga kelompok. Yuliantono et. al. [2] melakukan pengelompokan provinsi di Indonesia menjadi dua *cluster* berdasarkan nilai IPG tahun 2016-2019 menggunakan metode *Fuzzy C-Means*. Nurfadilah [3] melakukan analisis *cluster* longitudinal yang lebih rinci pada satu provinsi saja dengan memanfaatkan metode *K-Means* di Sulawesi Selatan tahun 2010-2020. Dari 24 kota/kabupaten, terbentuk 6 kelompok berdasarkan IPM Perempuan, IPG, dan IDG.

Berdasarkan uraian di atas, maka perlu dilakukan pembuatan gugus-gugus capaian pembangunan dan pemberdayaan perempuan pada 34 provinsi di Indonesia untuk melihat pola pengelompokan setiap wilayah. Pengelompokan akan dilakukan berdasarkan indikator pembangunan dan pemberdayaan perempuan pada bidang pendidikan, ekonomi, politik, dan kesehatan pada tahun 2021.

2. KAJIAN PUSTAKA

2.1 Clustering

Clustering merupakan teknik dari *unsupervised learning* yang digunakan untuk membentuk kelompok-kelompok dari sekumpulan objek, poin data, atau elemen. *Unsupervised learning* merupakan analisis bersifat deskriptif dan berguna untuk mengategorikan atau mengelompokkan data [4]. Berdasarkan Yuliantono et. al. [2], tujuan utama dari kegiatan *clustering* adalah melakukan pengelompokan data berdasarkan kemiripan karakteristik dan tidak memerlukan label atau target *output*. Maka dari itu, untuk menguji kebaikan suatu metode *clustering*, dapat dilakukan dengan melakukan validasi. *Cluster* yang baik adalah yang memiliki keragaman sekecil mungkin di dalamnya dan keragaman sebesar mungkin dengan *cluster* yang lain [5].

2.2 Ukuran Validasi

Validitas *cluster* internal merupakan faktor utama yang mampu mempengaruhi validasi dalam memilih metode *clustering* terbaik maupun jumlah *cluster* optimal [6]. Menurut Liu et. al.

[7], konsep dasar dari pengukuran validasi internal ada dua, yaitu *compactness* atau kepadatan suatu *cluster* dan *separation* atau perbedaan antar *cluster*. Ada sebelas indeks yang dapat digunakan sebagai pengukuran validasi. Pada penelitian ini, Davies-Bouldin *Index*, Calinski-Harabasz *Index*, Silhouette *Index*, dan Dunn *Index* akan dipilih untuk melakukan validasi.

2.2.1 Metode Elbow

Metode Elbow merupakan salah satu metode yang bisa digunakan untuk menentukan nilai k atau jumlah *cluster* paling optimum ketika hendak melakukan klusterisasi. Pada metode ini, akan memanfaatkan nilai *Sum of Square Error* (SSE) dari setiap *cluster* yang terbentuk. Semakin besar jumlah *cluster*, semakin kecil nilai SSE. Titik paling optimal adalah titik awal ketika semakin bertambahnya jumlah *cluster*, penurunan SSE semakin tidak signifikan perubahannya. Titik tersebut yang kemudian menjadi siku atau elbow [2].

2.2.2 Davies-Bouldin Index

Davies-Bouldin *Index* merupakan salah satu dari teknik validasi yang dapat dimanfaatkan untuk menentukan jumlah *cluster* optimal selain Metode Elbow. Semakin kecil nilai Davies-Bouldin *Index*, jumlah *cluster* semakin optimal. Dengan memperkecil indeks tersebut, maka perbedaan antar *cluster* semakin besar [7]. Nilai Davies-Bouldin *Index* didapatkan dengan menghitung diameter cluster dan jarak antar pusat *cluster* [8]. Davies-Bouldin *Index* dihitung dengan persamaan berikut.

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (1)$$

$$DB(k) = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} R_{ij} \quad (2)$$

Nilai s_i dan s_j merupakan rata-rata jarak tiap elemen dalam *cluster* i dan pusat *cluster* (diameter *cluster*) dan d_{ij} adalah jarak antara pusat *cluster* i dan satu pusat *cluster* lainnya. Jumlah *cluster* yang ingin dihitung nilai indeksnya dilambangkan dengan k .

2.2.3 Calinski-Harabasz Index

Calinski-Harabasz *Index* menghitung perbandingan antara nilai *Sum of Square Between Cluster* (SSB) sebagai *separation* dan nilai *Sum of Square Within Cluster* (SSW) sebagai *compactness* yang selanjutnya dikalikan dengan faktor normalisasi [9]. Faktor normalisasi didapatkan dengan membagi selisih jumlah data dan *cluster* yang terbentuk dengan jumlah *cluster* yang sudah dikurangi satu. Indeks ini dapat digunakan untuk melakukan pemilihan k optimum dengan nama lain *Variance Ratio Criterion* [8]. Jumlah optimal *cluster* ditandai dengan nilai Calinski-Harabasz yang tinggi [7]. Tibshirani *et. al.* [10] menuliskan rumus untuk mendapatkan nilai indeks Calinski-Harabasz sebagai berikut.

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)} \quad (3)$$

$B(k)$ dan $W(k)$ adalah varians antar-*cluster* (*between*) dan di dalam *cluster* (*within*). Nilai k adalah jumlah cluster di mana $k > 1$, sementara n adalah jumlah amatan atau elemen.

2.2.4 Silhouette Index

Silhouette *Index* merupakan salah satu indeks yang digunakan untuk melakukan validasi dalam menilai kebaikan suatu model *clustering*. Selain itu, indeks ini juga dapat dimanfaatkan

untuk menentukan nilai k optimum atau jumlah *cluster* optimal dalam proses klusterisasi [11]. Secara umum, indeks validasi ini menghitung rata-rata nilai setiap titik pada himpunan data atau menghitung selisih nilai *separation* dan *compactness* yang dibagi dengan nilai maksimum antara keduanya [9].

Tibshirani *et. al.* [10] menuliskan rumus untuk mendapatkan nilai *Silhouette Index* sebagai berikut.

$$s(i) = \frac{b(i)-a(i)}{\max \{a(i),b(i)\}} \quad (4)$$

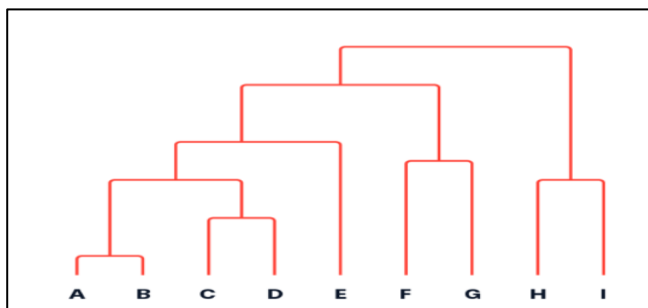
Nilai $a(i)$ didapatkan dengan menghitung rata-rata jarak antarelemen di dalam *cluster*, sementara $b(i)$ adalah rata-rata dari jarak antarelemen dari *cluster* berbeda yang letaknya paling dekat dengan elemen itu sendiri. Rentang nilai *Silhouette Index* adalah [-1,1]. Jumlah *cluster* maupun metode *clustering* akan semakin baik apabila nilai *Silhouette Index* semakin besar atau mendekati 1. Jika nilainya 0, maka objek terletak pada batas *cluster* lain, sementara apabila nilainya adalah negatif, data observasi terletak pada *cluster* yang salah [12], [13].

2.2.5 Dunn Index

Dunn Index menghitung nilai minimum dari perbandingan antara nilai fungsi ketidaksamaan antara dua *cluster* sebagai *separation* dan nilai maksimum dari diameter *cluster* sebagai *compactness* [9]. Jumlah *cluster* optimal akan ditandai dengan nilai *Dunn Index* yang semakin tinggi [7]. *Dunn Index* didapatkan dengan membagi jarak minimum antar titik di dalam *cluster* dengan jarak maksimum diameter antar-*cluster*. Diameter *cluster* adalah jarak terjauh antara dua elemen dari *cluster* yang berbeda.

2.3 Hierarchical Clustering

Metode *clustering* terbagi dalam dua jenis, yaitu *hierarchical clustering* dan *non-hierarchical clustering* [8]. Metode *hierarchical clustering* dilakukan dengan membangun *cluster* yang direpresntasikan dalam bentuk pohon atau dendogram. Akar dari pohon atau dendogram menggambarkan jumlah *cluster*, sementara daunnya akan menggambarkan letak satu elemen atau objek dalam satu *cluster* tertentu [14]. Dalam penelitian ini, algoritma yang digunakan adalah pendekatan *bottom-up manner*, yaitu dimulai dari mengasumsikan bahwa jumlah *cluster* adalah sama dengan jumlah elemen dan pada setiap penambahan *height*, *cluster* akan digabungkan hingga jumlah yang diinginkan atau sesuai dengan penentuan k optimal [15].



Gambar 1. Dendrogram atau Diagram Pohon

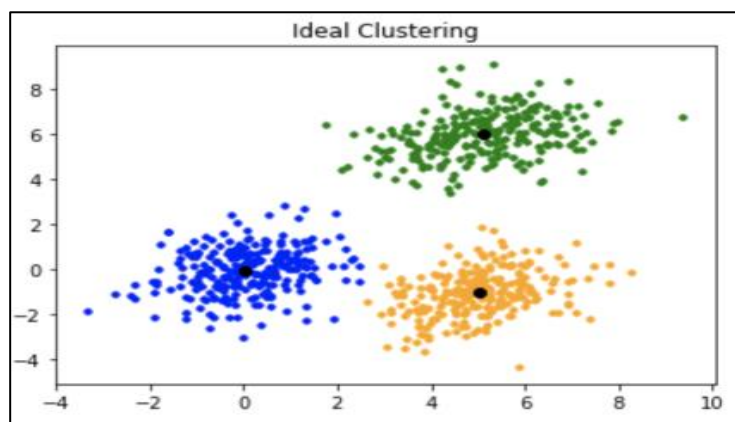
2.4 Non-hierarchical Clustering

Non-hierarchical clustering merupakan metode klusterisasi dengan menggabungkan atau memisahkan elemen-elemen yang ada. Dalam metode *non-hierarchical clustering*, pembentukan

cluster tidak dilakukan dengan membentuk *dendrogram* atau diagram pohon. Klasterisasi dilakukan dengan meminimalisasi atau memaksimalkan suatu nilai sesuai kriteria-kriteria tertentu [16].

2.4.1 K-Means Clustering

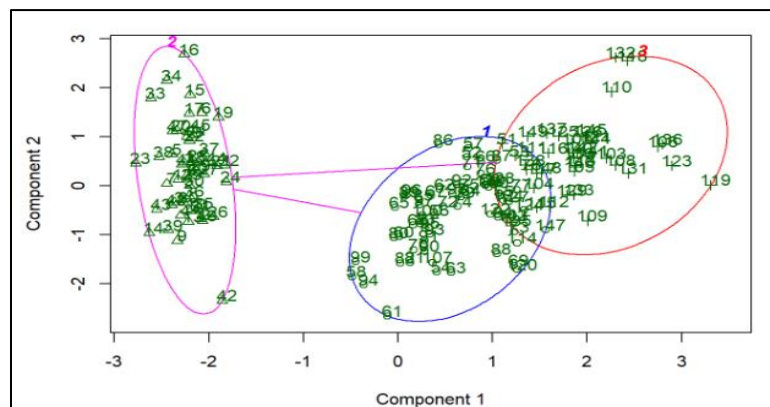
Metode *K-Means clustering* merupakan salah satu metode *non-hierarchical clustering* yang sensitif terhadap pemilihan titik pusat. Khairati *et al.* [9] menyebutkan bahwa kelebihan dari pemanfaatan *K-Means* antara lain adalah implementasi yang mudah, tingkat konvergensi tinggi, dan kemampuannya dalam menghasilkan *cluster* yang lebih padat dibandingkan metode hierarki. Pada teknik ini, titik pusat atau *centroid* yang digunakan untuk mewakili *cluster* adalah rata-rata dari setiap *cluster* yang terbentuk. Oleh karena itu, algoritma *K-Means* akan sangat sensitif terhadap *outlier* atau pencilan [8].



Gambar 2. *K-Means Clustering*

2.4.2 Fuzzy C-Means

Metode *Fuzzy C-Means clustering* yang memanfaatkan pendekatan *fuzzy* pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981. Sesuai dengan konsep *fuzzy*, yaitu ketidakberaturan, maka elemen data bisa tergabung dalam lebih dari satu *cluster*. *Fuzzy C-Means* merupakan teknik klasterisasi yang menggunakan derajat keanggotaan untuk menentukan penempatan suatu data dalam sebuah *cluster* [2].



Gambar 3. *Fuzzy C-Means Clustering*

3. METODE PENELITIAN

Data yang digunakan dalam penelitian ini adalah data sekunder yang didapatkan dari *official website* Badan Pusat Statistik (BPS) pada kategori Gender. Periode yang dipilih adalah tahun 2021 dengan lokus 34 provinsi di Indonesia. Detail dari proksi-proksi variabel yang menggambarkan pembangun dan pemberdayaan perempuan dijelaskan pada tabel di bawah ini.

Tabel 1. Keterangan Variabel

No	Variabel	Proksi	Satuan
1	Pendidikan	Rata-rata Lama Sekolah Perempuan	Tahun
2	Ekonomi	Sumbangan Pendapatan Perempuan	Persen
3	Politik	Keterlibatan Perempuan di Parlemen	Persen
4	Kesehatan	Angka Harapan Hidup Perempuan	Tahun

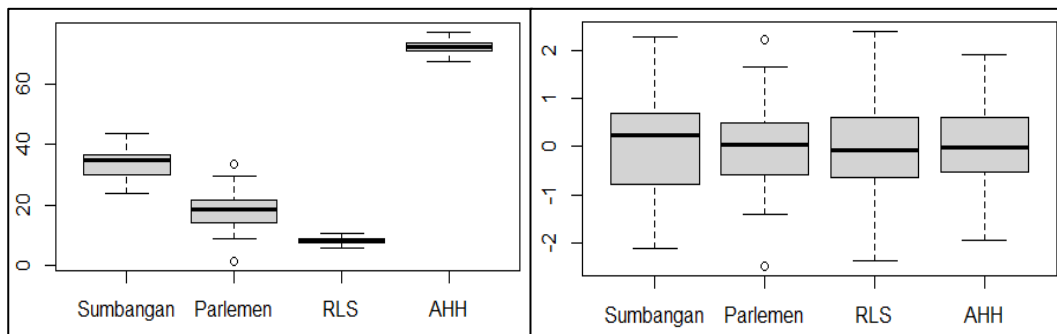
Langkah-langkah untuk memilih metode *clustering* terbaik di antara metode *Hierarchical*, *K-Means*, dan *Fuzzy C-Means* adalah sebagai berikut.

1. Melakukan standarisasi data untuk menyeragamkan satuan yang berbeda-beda.
2. Memilih perkiraan nilai *k* optimum atau jumlah *cluster* optimal dengan metode Elbow.
3. Membandingkan indeks-indeks validasi antar-*cluster*, yaitu dengan *Davies-Bouldin Index* dan *Calinski-Harabasz Index*. Jumlah *cluster* dengan nilai *Davies-Bouldin Index* terendah dan *Calinski-Harabasz Index* tertinggi merupakan *k* optimum.
4. Melakukan validasi *k* optimum dari metode *Hierarchical*, *K-Means*, dan *Fuzzy C-Means* dengan memeriksa *Silhouette Index*, *Dunn Index*, dan *Calinski-Harabasz Index*.
5. Memeriksa *missclustering* dengan memerhatikan nilai *Silhouette Index* bernilai negatif pada ketiga metode *clustering*.
6. Memilih metode *clustering* terbaik.

4. HASIL DAN PEMBAHASAN

4.1 Gambaran Umum

Variabel yang digunakan dalam penelitian ini memiliki satuan yang berbeda, sehingga sebelum melakukan *clustering*, perlu dilakukan standarisasi pada seluruh data yang digunakan. Hal tersebut bertujuan untuk menyeragamkan *dataset* agar rata-rata dari setiap variabel akan mendekati 0 dengan standar deviasi mendekati 1 atau *dataset* mengikuti asumsi normalitas.



Gambar 4. Output Boxplot Sebelum Standarisasi (kanan) dan Sesudah Standarisasi (kiri)

Gambar 4 menunjukkan perbandingan visualisasi data sebelum dan sesudah distandarisasi. Dari *boxplot* kanan, variabel yang mencakup elemen-elemen bernilai besar adalah Angka

Harapan Hidup dalam tahun sebagai proksi dari bidang kesehatan. Variabel yang memiliki *outlier* di dalamnya adalah persentase dari Keterlibatan Perempuan di Parlemen sebagai proksi dari bidang politik.

Tabel 2. Persentase Keterlibatan Perempuan di Parlemen tahun 2021

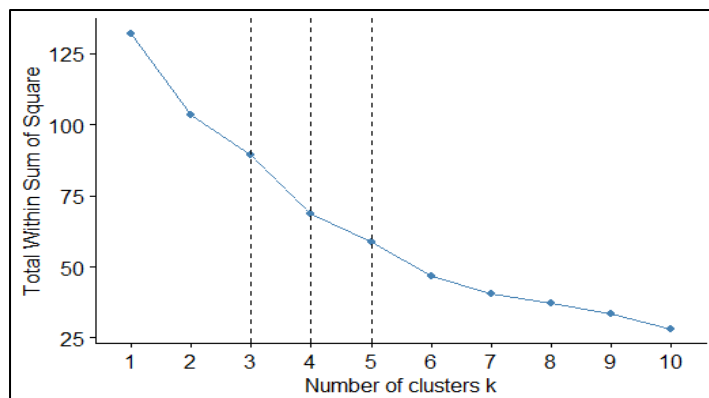
No	Provinsi	%	No	Provinsi	%	No	Provinsi	%
1	Aceh	11.11	13	Jawa Tengah	18.33	25	Sulawesi Utara	29.55
2	Sumatera Utara	13.13	14	Di Yogyakarta	20.00	26	Sulawesi Tengah	27.27
3	Sumatera Barat	10.77	15	Jawa Timur	17.50	27	Sulawesi Selatan	25.88
4	Riau	21.54	16	Banten	17.65	28	Sulawesi Tenggara	20.00
5	Jambi	14.55	17	Bali	16.36	29	Gorontalo	26.67
6	Sumatera Selatan	21.33	18	Nusa Tenggara Barat	1.59	30	Sulawesi Barat	11.11
7	Bengkulu	15.91	19	Nusa Tenggara Timur	20.00	31	Maluku	23.26
8	Lampung	18.82	20	Kalimantan Barat	16.92	32	Maluku Utara	28.89
9	Kep. Bangka Belitung	8.89	21	Kalimantan Tengah	33.33	33	Papua Barat	14.29
10	Kep. Riau	8.89	22	Kalimantan Selatan	20.00	34	Papua	14.55
11	Dki Jakarta	21.70	23	Kalimantan Timur	20.37			
12	Jawa Barat	20.83	24	Kalimantan Utara	11.43			

Berdasarkan Tabel 2 persentase terendah adalah Provinsi Nusa Tenggara Barat dengan angka 1,59% saja, sementara tertinggi adalah 33,33% dari Provinsi Kalimantan Tengah. Akibat adanya perbedaan satuan yang menyebabkan nilai *mean* keempat variabel berbeda-beda serta adanya *outlier*, maka perlu dilakukan proses standarisasi data.

Boxplot kiri pada Gambar 4 menunjukkan bahwa sekalipun proses standarisasi belum mampu menghilangkan *outlier* pada variabel partisipasi perempuan dalam parlemen, nilai *mean* sudah seragam dengan ketiga indikator yang lain. *Dataset* yang akan digunakan untuk melakukan *clustering* pada tahap-tahap selanjutnya adalah *dataset* hasil standarisasi.

4.2 Penentuan *k* atau Jumlah *Cluster* Optimal

Metode paling umum yang biasa digunakan untuk memilih jumlah kluster (*k*) optimum adalah Metode Elbow. Maka dari itu, dalam menentukan *k* optimal pertama kali, akan digunakan Metode Elbow dengan *output* seperti gambar di bawah ini.



Gambar 5. *Output* Metode Elbow untuk Penentuan *k* Optimum

Berdasarkan Gambar 5, ada beberapa titik *elbow* atau siku yang dapat dikatakan sebagai jumlah *cluster* terbaik. Perubahan nilai SSE akan semakin tidak terlihat perbedaannya pada jumlah *k* optimum 3. *Cluster* 4 dan 5 juga dapat dipilih menjadi titik optimum karena grafik *Elbow* yang tidak membentuk siku-siku, sehingga tidak dapat memperlihatkan satu titik atau angka *k* yang menunjukkan jumlah *cluster* paling optimum. Pertimbangan untuk memilih *k* optimum mana yang paling baik dapat dilakukan dengan memeriksa nilai indeks-indeks yang lain, yaitu *Davies-Bouldin Index* dan *Calinski-Harabasz Index*. Perbandingannya diringkas dalam tabel berikut.

Tabel 3. Nilai *Davies-Bouldin Index* dan *Calinski-Harabasz Index*

<i>k</i>	<i>Davies-Bouldin Index</i>	<i>Calinski-Harabasz Index</i>
3	1.4053	9.5990
4	1.4028	9.9182
5	1.2398	10.1118

Selain memiliki nilai SSE sekecil mungkin, *cluster* optimum juga ditandai dengan nilai *Davies-Bouldin Index* sekecil mungkin dan *Calinski-Harabasz Index* setinggi mungkin. Berdasarkan Tabel 3, jumlah *cluster* lima memenuhi syarat tersebut. Artinya, *cluster* paling optimal yang dapat mengelompokkan provinsi-provinsi di Indonesia menurut indikator pemberdayaan perempuan adalah lima.

4.3 Pembentukan *Cluster* dan Pemilihan Metode Terbaik

Validasi merupakan cara yang digunakan untuk menilai apakah metode *clustering* yang digunakan merupakan metode yang paling tepat atau tidak karena tidak ada ukuran yang pasti untuk membuktikan kebenaran dalam *unsupervised learning*. Indeks validasi juga dapat dimanfaatkan untuk membanding suatu metode dengan metode yang lain. Ada beberapa nilai indeks yang biasa dipakai untuk validasi, antara lain *Silhouette Index*, *Dunn Index*, dan *Calinski-Harabasz Index*. Semakin besar nilai ketiga indeks tersebut, maka metode *clustering* yang digunakan akan semakin optimal.

Tabel 4. Indeks Validasi untuk Pemilihan Model *Clustering* Terbaik

Metode <i>Clustering</i>	<i>Silhouette Index</i>	<i>Dunn Index</i>	<i>Calinski-Harabasz Index</i>
<i>Hierarchical</i>	0.2254	0.2109	10.2234
<i>K-Means</i>	0.2434	0.2116	10.6696
<i>Fuzzy C-Means</i>	0.2272	0.2140	10.1118

Berdasarkan Tabel 4, Metode *K-Means* telah memenuhi dua syarat validasi, dilihat dari nilai *Silhouette Index* dan *Calinski-Harabasz Index* yang paling besar dibandingkan dua metode yang lain. Artinya, berdasarkan indeks-indeks validasi, dapat dinyatakan bahwa metode *K-Means* adalah metode yang paling baik untuk membentuk *cluster*. Saat melakukan *clustering*, ada kemungkinan *missclustering*, yaitu kondisi ketika suatu poin atau elemen salah diikutkan dalam suatu *cluster*. Hal tersebut dapat diidentifikasi dengan *Silhouette Index*. Indeks yang bernilai negatif mengindikasikan bahwa ada *missclustering* elemen dari kelompok-kelompok yang telah dibentuk dan memperkirakan *cluster* tetangga (*neighbor*) yang sebenarnya paling cocok.

Tabel 5. *Missclustering* Provinsi berdasarkan Metode *Clustering* yang Digunakan

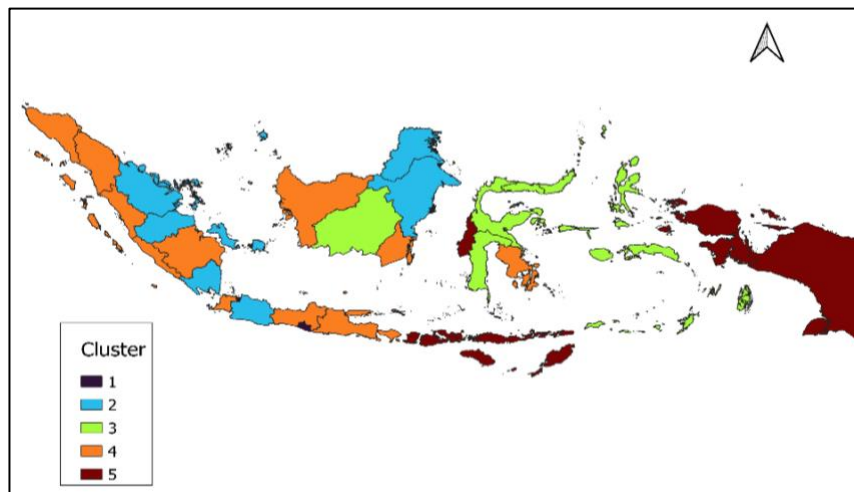
Metode <i>Clustering</i>	Provinsi	<i>Negative Silhouette Index</i>	<i>Cluster</i>	<i>Neighbor</i>
<i>Hierarchical</i>	DIY	-0,0353	1	3
	Maluku	-0,0423	1	5
	Bengkulu	-0,1159	1	3
<i>K-Means</i>	NTT	-0,0675	5	4
<i>Fuzzy C-Means</i>	Maluku	-0,0139	1	5
	Kep. Riau	-0,0376	1	2
	Bengkulu	-0,0624	1	3
	NTT	-0,0680	4	3

Tabel 5 menunjukkan bahwa masih terdapat *missclustering* dalam penggunaan tiga metode *clustering*. Di antara ketiga metode yang digunakan, *K-Means* memiliki jumlah kesalahan paling sedikit. Oleh karena itu, disimpulkan bahwa dengan metode *K-Means*, pengelompokan 34 Provinsi di Indonesia menggunakan indikator pembangunan dan pemberdayaan perempuan dalam bidang pendidikan, ekonomi, politik, dan kesehatan menjadi lima *cluster* akan optimal. Dengan jumlah optimal lima *cluster*, maka hasil akhir dari pembentukan kelompok berdasarkan indikator pembangunan dan pemberdayaan perempuan dari 34 provinsi di Indonesia diringkas pada Tabel 6.

Tabel 6. Pembagian *Cluster* dengan Metode *K-Means*

<i>Cluster</i>	Jumlah Anggota	Provinsi Anggota
1	2	DIY dan DKI Jakarta
2	8	Jambi, Jawa Barat, Kalimantan Timur, Kalimantan Utara, Kep. Bangka Belitung, Kep. Riau, Lampung, dan Riau
3	7	Gorontalo, Kalimantan Tengah, Maluku, Maluku Utara, Sulawesi Selatan, Sulawesi Tengah, dan Sulawesi Utara
4	12	Aceh, Bali, Banten, Bengkulu, Jawa Tengah, Jawa Timur, Kalimantan Barat, Kalimantan Selatan, Sulawesi Tenggara, Sumatera Barat, Sumatera Selatan, dan Sumatera Utara
5	5	NTB, NTT, Papua Barat, Papua, dan Sulawesi Barat

Gambar 6 merupakan visualisasi sebaran wilayah hasil *clustering* pada Tabel 6. Berdasarkan peta dari Gambar 6, terlihat bahwa anggota dari sebuah *cluster* terdiri dari wilayah yang berdekatan, baik provinsi maupun pulau. *Cluster* 1 terdiri dari DKI Jakarta dan DI Yogyakarta yang walaupun berjauhan, masih terletak di satau pulau, yaitu Pulau Jawa. *Cluster* 2 dan 3 banyak diisi oleh wilayah bagian barat Indonesia, sementara *cluster* 4 dan 5 banyak diisi oleh provinsi dari daerah timur Indonesia. Selain kedekatan antar wilayah, *cluster* yang terbentuk juga dapat dijelaskan sesuai dengan kondisi pembangunan perempuan.



Gambar 6. Klasterisasi Provinsi di Indonesia berdasarkan Indikator Pembangunan dan Pemberdayaan Gender

Tabel 7 menunjukkan nilai rata-rata seluruh variabel dari *cluster* yang terbentuk. Wilayah pada *Cluster 1* merupakan wilayah yang dapat dikatakan memiliki tingkat pembangunan dan pemberdayaan perempuan terbaik. *Cluster 2* merupakan provinsi yang capaian pendidikan dan partisipasi perempuan dalam parlemennya tinggi, akan tetapi masih butuh pemajuan dalam bidang kelayakan hidup serta kesehatan. *Cluster 3* merupakan kelompok daerah yang tingkat pembangunan perempuan cukup baik, akan tetapi masih kurang terlihat peran perempuan dalam kegiatan perekonomian dan politik. *Cluster 4* dan *5* memiliki kemiripan, yaitu kualitas pendidikan dan angka partisipasi perempuan dalam politik masih rendah. Kualitas hidup perempuan dalam *Cluster 4* sudah cukup baik, begitu juga dengan *Cluster 5*. Akan tetapi, kualitas kesehatan perempuan pada *Cluster 4* perlu menjadi catatan pembenahan kebijakan dalam pembangunan perempuan.

Tabel 7. Rata-rata Setiap *Cluster*

<i>Cluster</i>	Sumbangan	Parlemen	RLS	AHH
1	1.3933	0.3775	1.6442	1.5157
2	-0.1545	1.4144	0.3922	-0.4051
3	-1.2146	-0.3921	0.2779	0.6793
4	0.4871	-0.8903	-1.3514	-1.5761
5	0.4647	-0.2556	-0.1250	0.1875

5. KESIMPULAN

Pemilihan nilai *k* atau jumlah *cluster* minimum menggunakan Metode Elbow masih bergantung pada subjektivitas, terutama jika grafik *output* tidak menampilkan titik siku atau *elbow* yang jelas. Nilai Davies-Bouldin *Index* dan Calinski-Harabasz *Index* dapat digunakan sebagai metode tambahan dalam pemilihan *k* paling optimal. Berdasarkan kedua indeks tersebut, jumlah pembentukan *cluster* yang terpilih adalah lima. Selanjutnya, untuk memilih satu dari tiga metode *clustering* yang paling baik, digunakan validasi dengan Silhouette Indeks, Davies-Bouldin *Index*, dan Calinski-Harabasz *Index*. Metode *K-Means* memiliki nilai Silhouette *Index* dan Calinski-Harabasz *Index* terbesar dibanding Hierarchical maupun *Fuzzy C-Means*. *K-Means* juga memiliki kesalahan pengklasteran provinsi paling minimum di antara kedua metode yang lain.

Artinya, metode *K-Means* adalah metode yang paling baik untuk membentuk lima *cluster* dari 34 provinsi di Indonesia menurut indikator pembangunan dan pemberdayaan perempuan dalam bidang pendidikan, ekonomi, politik, dan kesehatan pada tahun 2021.

DAFTAR PUSTAKA

- [1] R. Diana, “Analisis Ketimpangan Gender Di Provinsi Sumatera Barat,” *J. Kependud. Indones.*, vol. 13, no. Juni, pp. 55–66, 2018.
- [2] D. A. Yuliantono *et al.*, “Algoritma Fuzzy C-Means Dengan Metode Elbow Untuk Mengelompokkan Provinsi Di Indonesia Berdasarkan Indeks Pembangunan Gender,” Universitas Muhammadiyah Jember, 2019. [Online]. Available: <http://repository.unmuhjember.ac.id/4708/10/J. ARTIKEL.pdf>
- [3] K. Nurfadilah, “Analisis Cluster Longitudinal Pada Pembangunan Manusia Di Sulawesi Selatan Berbasis Gender,” *J. MSA (Matematika dan Stat. serta Apl.*, vol. 9, no. 1, 2021, doi: 10.24252/msa.v9i1.21180.
- [4] Badan Pusat Statistik, *Sensus Ekonomi 2006 Evaluasi Terhadap Kriteria UMK dan UMB Hasil SE2006-SS*. Jakarta: Badan Pusat Statistik, 2009.
- [5] R. T. Vuldari, S. Siswanti, A. K. Kusumawijaya, and K. Sandradewi, “Classification of Human Development Index Using K-Means,” *Indones. J. Appl. Stat.*, vol. 2, no. 1, pp. 1, 2019, doi: 10.13057/ijas.v2i1.28566.
- [6] R. P. Prayogo and J. L. Buliali, “Penentuan Jumlah Cluster Optimum Pada Segmen Rute Penerbangan Menggunakan Data Automatic Dependent Surveillance-Broadcast,” *JUTI J. Ilm. Teknol. Inf.*, vol. 18, no. 1, pp. 48, 2020, doi: 10.12962/j24068535.v18i1.a902.
- [7] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, “Understanding of internal clustering validation measures,” in *International Conference on Data Mining*, 2010, pp. 911–916. doi: 10.1109/ICDM.2010.35.
- [8] A. M. Sikana and A. W. Wijayanto, “Analisis Perbandingan Pengelompokan Indeks Pembangunan Manusia Indonesia Tahun 2019 dengan Metode Partitioning dan Hierarchical Clustering,” *J. Ilmu Komput.*, vol. 14, no. 2, pp. 66, 2021, doi: 10.24843/jik.2021.v14.i02.p01.
- [9] A. F. Khairati, A. . Adlina, G. . Hertono, and B. . Handari, “Kajian Indeks Validitas pada Algoritma K-Means Enhanced dan K-Means MMCA,” in *PRISMA, Prosiding Seminar Nasional Matematika*, 2019, vol. 2, pp. 161–170. [Online]. Available: <https://journal.unnes.ac.id/sju/index.php/prisma/article/view/28906>
- [10] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the Number of Clusters in a Data Set Via the Gap Statistic,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 63, no. 2, pp. 411–423, Jul. 2001, doi: 10.1111/1467-9868.00293.
- [11] Scikit Learn, “Selecting the number of clusters with silhouette analysis on KMeans clustering,” https://scikitlearn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html, 2017.
- [12] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.
- [13] L.Fonseca, “Clustering Analysis in R using K-means,”

- <https://towardsdatascience.com/clustering-analysis-in-r-using-k-means-73eca4fb7967>, 2019.
- [14] University of Cincinnati, “Hierarchical Cluster Analysis,” <http://uc-r.github.io/descriptive>, 2018.
- [15] I. Mokris and L. Skovajsova, “Comparison of Document Clustering Techniques,” Twin Cities, 2008.
- [16] Kapilparshi, “Difference between Hierarchical and Non Hierarchical Clustering,” <https://www.geeksforgeeks.org/difference-between-hierarchical-and-non-hierarchical-clustering/>, 2022.