

Analisis *Cluster* Kemiskinan Provinsi di Indonesia Tahun 2019 dengan Metode *Partitioning* dan *Hierarki*

Natasya Afira^{1*}, Arie Wahyu Wijayanto²

^{1,2}Program Studi Komputasi Statistik Peminatan Sains Data, Politeknik Statistika STIS

Jl. Otto Iskandardinata No. 64C Jakarta 13330

*email: 221810497@stis.ac.id

(Naskah masuk: 26 Desember 2020; diterima untuk diterbitkan: 31 Maret 2021)

ABSTRAK – Kemiskinan merupakan salah satu indikator penting dalam melihat keberhasilan pembangunan ekonomi suatu negara. Tingkat kemiskinan di Indonesia sebesar 9,22 persen dengan jumlah penduduk miskin sebanyak 24,79 juta jiwa. Data kemiskinan di setiap daerah akan berbeda dipengaruhi oleh berbagai indikator. Oleh karena itu penting untuk mengelompokkan daerah-daerah di Indonesia berdasarkan karakteristik kemiskinan sehingga pemerintah dapat dengan tepat mengambil kebijakan terkait penanggulangan kemiskinan. Penelitian ini menggunakan dua metode clustering, yaitu *partitioning* dan *hierarki* untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan karakteristik kemiskinan. Metode *partitioning* yang dipilih adalah K-Means. Data yang digunakan adalah 8 variabel kemiskinan pada 34 provinsi di Indonesia tahun 2019. Penentuan jumlah cluster menggunakan validasi internal dan validasi stabilitas menunjukkan bahwa metode hierarki dengan jumlah cluster 2 menghasilkan cluster yang paling optimal. Perbandingan metode hierarki dinilai berdasarkan *agglomerative coefficient*, dimana metode Ward mampu memberikan hasil pengelompokan terbaik.

Kata Kunci – Kemiskinan, Analisis Klaster, *Partitioning*, Hierarki.

Cluster Analysis with Partitioning and Hierarchical Methods on Provincial Poverty Information Data in Indonesia in 2019

ABSTRACT – Poverty is an important indicator in seeing the success of a country's economic development. The poverty rate in Indonesia is 9.22 percent with a total poor population of 24.79 million people. Poverty data in each region will be different, influenced by various indicators. Therefore, it is important to categorize regions in Indonesia based on poverty characteristics so that the government can make the right policies related to poverty reduction. This study uses two clustering methods, namely *partitioning* and *hierarchy* to group provinces in Indonesia based on poverty characteristics. The *partitioning* method chosen is K-Means. The data used are 8 poverty variables in 34 provinces in Indonesia in 2019. The determination of the number of clusters using internal validation and stability validation shows that the hierarchical method with the optimum number of clusters 2 produces the most optimal clusters. The comparison of the hierarchical method is assessed based on the *agglomerative coefficient*, where the Ward method is able to provide the best grouping results.

Keywords – Poverty, Cluster Analysis, *Partitioning*, Hierarchical.

1. PENDAHULUAN

Menurut Badan Pusat Statistik (BPS), kemiskinan adalah ketidakmampuan dari sisi ekonomi untuk memenuhi kebutuhan dasar makanan dan bukan makanan yang diukur dari sisi pengeluaran perkapita perbulan. Berdasarkan data BPS tahun 2019, persentase kemiskinan di Indonesia sebesar

9,22% (bulan September) dengan jumlah penduduk miskin sebanyak 24,79 juta jiwa [1].

Data kemiskinan di setiap daerah akan berbeda karena dipengaruhi oleh berbagai indikator. Pemerintah perlu mendapat gambaran kemiskinan tiap daerah dengan mengelompokkan provinsi-provinsi di Indonesia berdasarkan karakteristik kemiskinan sehingga pemerintah dapat dengan

tepat mengambil kebijakan penanggulangan kemiskinan.

Analisis *cluster* adalah sebuah pendekatan untuk mencari kesamaan dalam data dan menempatkan data yang sama ke dalam kelompok-kelompok. Analisis *cluster* membagi sekumpulan data ke dalam beberapa kelompok dimana kesamaan dalam sebuah kelompok lebih besar daripada diantara kelompok-kelompok [2].

Dalam analisis *cluster* terdapat dua metode yaitu *partitioning* dan hierarki. Proses pengelompokan metode *partitioning* menentukan jumlah *cluster* dan pemilihan titik pusat terlebih dahulu. Sedangkan metode hierarki dilakukan secara bertahap atau terstruktur tetapi banyaknya *cluster* belum diketahui. Salah satu metode *partitioning* adalah K-Means. Metode hierarki sendiri terdiri dari *single linkage*, *complete linkage*, *average linkage*, dan metode *ward* [3].

Penelitian terdahulu yang pernah dilakukan adalah pengelompokan kabupaten/kota di Provinsi Jawa Timur berdasarkan indikator kemiskinan [4], penerapan metode *complete linkage* dan metode *hierarchical clustering multiscale bootstrap* (kemiskinan di Kalimantan Timur) yang dilakukan oleh Ramadhani tahun 2016 [5], analisis *clustering* K-Means pada data informasi kemiskinan di Jawa Barat yang dilakukan oleh Febianto tahun 2018 [6], dan analisis klaster untuk pengelompokan kabupaten/kota di Provinsi Sulawesi Selatan berdasarkan indikator kesejahteraan rakyat oleh Hasrul tahun 2018 [7].

Berdasarkan penelitian di atas, penulis ingin melakukan analisis klaster dengan metode *partitioning* dan hierarki untuk mengelompokkan

provinsi di seluruh Indonesia menggunakan data informasi kemiskinan tahun 2019.

2. METODE DAN BAHAN

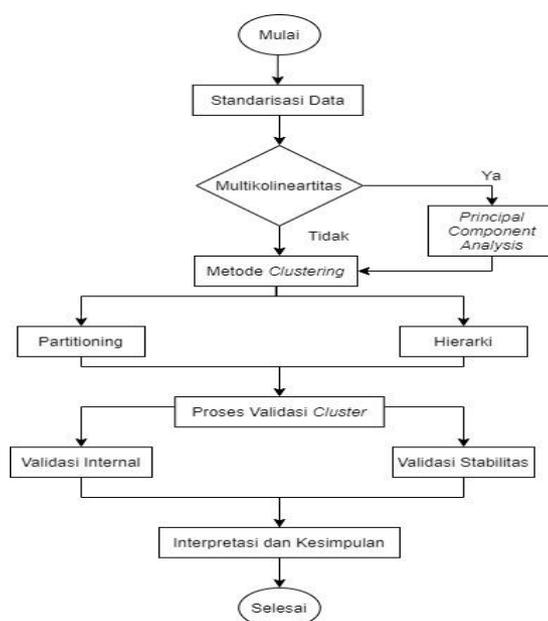
Penelitian ini menggunakan data sekunder yang diperoleh dari website Badan Pusat Statistik. Data yang digunakan adalah data kemiskinan tahun 2019 dengan jumlah 34 provinsi. Dataset dan R code yang digunakan peneliti dapat diakses melalui <https://github.com/NatasyaAfiraa/Paper>. Tahapan-tahapan penelitian dapat dilihat pada gambar 1. Daftar variabel yang digunakan pada penelitian ini dapat dilihat pada tabel 1.

Tabel 1. Daftar Variabel Penelitian

Variabel	Keterangan
X1	Persentase Penduduk Miskin
X2	PDRB (Juta Rupiah)
X3	Rata-Rata Lama Sekolah
X4	Pengeluaran Riil Perkapita (Juta Rupiah)
X5	Tingkat Pengangguran Terbuka
X6	Angka Harapan Hidup
X7	Angka Melek Huruf
X8	Persentase Kepemilikan Rumah Sendiri

2.1 Analisis Cluster

Analisis *cluster* merupakan metode analisis untuk mengelompokkan objek-objek pengamatan menjadi beberapa kelompok, sehingga akan diperoleh kelompok dimana objek-objek dalam satu kelompok memiliki banyak persamaan sedangkan dengan anggota kelompok lain memiliki banyak perbedaan [8].



Gambar 1. Diagram Alur Penelitian

2.2 Analisis Faktor

Analisis faktor yang meliputi *principal component analysis* dan *common factor analysis*, adalah pendekatan statistik yang dapat digunakan untuk menganalisis "inter-relationship" sejumlah (besar) variabel dan untuk menjelaskan dimensi-dimensi apakah yang melandasi variabel-variabel tersebut. Tujuan analisis ini adalah memadatkan sejumlah besar informasi dari sejumlah variabel asli menjadi sejumlah kecil faktor dengan kehilangan informasi minimal [8].

Pada dasarnya *principal component analysis* (PCA) yaitu mereduksi dimensi data multivariat dengan mencari kombinasi linier dari variabel yang menyumbang sebanyak mungkin variasi total dalam data. Kemudian mencari kombinasi kedua yang tidak berkorelasi dengan kombinasi pertama, yang menyumbang sebanyak mungkin variasi yang tersisa dan seterusnya. Jika sebagian besar variasi dihitung oleh sejumlah kecil komponen, maka komponen-komponen tersebut dapat digunakan sebagai pengganti variabel asli [9].

Beberapa asumsi yang harus terpenuhi sebelum melakukan analisis faktor sebagai berikut [10]:

a. Kecukupan sampel

Untuk melihat kecukupan sampel dilakukan uji Kaizer-Meyer-Olkin (KMO). Analisis faktor dapat dilakukan ketika nilai KMO lebih besar 0.5 [10].

Hipotesa:

H_0 : Data layak untuk dianalisis

H_1 : Data tidak layak untuk dianalisis

Statistik Uji :

$$KMO = \frac{\sum_{i \neq j} \sum r_{ij}^2}{\sum_{i \neq j} \sum r_{ij}^2 + \sum_{i \neq j} \sum a_{ij}^2} \quad (1)$$

dengan,

r_{ij}^2 : korelasi antara variabel i dan j

a_{ij}^2 : korelasi parsial antara variabel i dan j

Kriteria pengujian:

H_0 ditolak jika nilai KMO < 0.5. Jadi dapat disimpulkan bahwa data tidak cukup untuk dilakukan analisis faktor.

b. Tidak terdapat multikolinearitas

Pengujian dapat dilakukan dengan uji barlett seperti pada persamaan berikut [10]:

Hipotesa :

H_0 : $R = 1$ (matriks korelasi sama dengan matriks identitas)

H_1 : $R \neq 1$ (matriks korelasi tidak sama dengan matriks identitas)

Statistik Uji :

$$Barlett = -\ln|R| \left(n - 1 - \left(\frac{2p+5}{6} \right) \right) \quad (2)$$

dengan,

$|R|$: nilai determinan dari matriks korelasi

n : banyaknya pengamatan

p : banyaknya variabel

Kriteria pengujian :

Tolak H_0 jika $P\text{-value} < \alpha$ maka variabel-variabel saling berkorelasi, hal ini berarti terdapat hubungan antar variabel (*multikolinearitas*), sehingga layak dilakukan analisis faktor.

2.3 Metode Hierarki

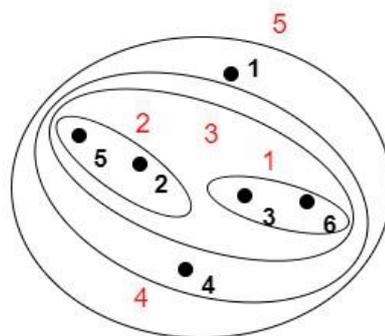
Metode hierarki adalah suatu metode analisis *cluster* yang dilakukan secara bertahap dan bertingkat sehingga membentuk tingkatan seperti pada struktur pohon [5]. Metode ini menghasilkan urutan partisi dengan menggabungkan atau membagi *cluster*. Pada setiap urutan tahapan, partisi baru secara optimal digabungkan atau dipisah dari partisi sebelumnya menurut beberapa kriteria kecukupan [3]. Hasil dari metode ini dapat disajikan dalam bentuk dendogram. *Dendogram* adalah representatif visual dari seluruh tahapan yang menunjukkan bagaimana *cluster* terbentuk. Selain itu juga terdapat nilai koefisien jarak pada setiap tahapan [5].

Nilai koefisien jarak ditentukan antara setiap pasangan poin. Dalam mengukur kesamaan jarak terdapat berbagai macam ukuran pendekatan seperti jarak *Euclidean* dan jarak *Manhattan* [11].

Metode hierarki terbagi dua, yaitu metode *agglomerative* (*bottom up*) dan metode *divisive* (*top down*). Perbedaannya terletak pada arah pengelompokkan. Ada 4 metode *agglomerative* dalam pembentukan *cluster* yaitu :

1. Single Linkage

Jarak dua *cluster* diukur dengan jarak terdekat antara sebuah objek dalam *cluster* yang satu dengan sebuah objek dalam *cluster* yang lain [12]. Ilustrasi *single linkage* dapat dilihat pada gambar 2.



Gambar 2. *Nested Cluster Single Linkage* [13]

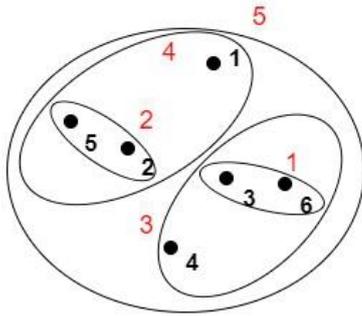
$$d(uv)w = \min(d_{uw}d_{vw}) \quad (3)$$

$d(uv)w$ adalah jarak antara cluster (UV) dan cluster W, d_{uw} dan d_{vw} adalah jarak antara tetangga terdekat *cluster* U dan W, serta *cluster* V dan W [12].

2. Complete Linkage

Jarak dua *cluster* diukur dengan jarak terjauh antara sebuah objek dalam *cluster* yang satu dengan

sebuah objek dalam *cluster* yang lain [12]. Ilustrasi *complete linkage* dapat dilihat pada gambar 3.

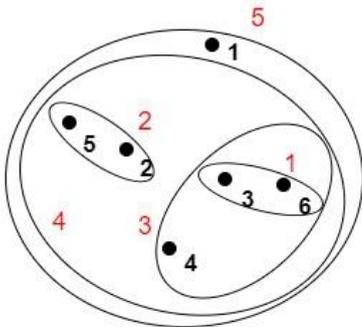


Gambar 3. Nested Cluster Complete Linkage [13]

$$d(uv)w = \max(d_{uw}d_{vw}) \quad (4)$$

3. Average Linkage

Jarak antara dua *cluster* diukur dengan jarak rata-rata antara sebuah objek dalam *cluster* yang satu dengan sebuah objek dalam *cluster* yang lain [12]. Ilustrasi *average linkage* dapat dilihat pada gambar 4.



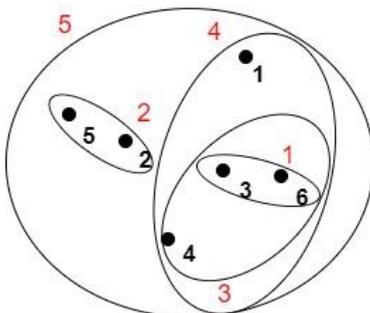
Gambar 4. Nested Cluster Average Linkage [13]

$$d(uv)w = \frac{\sum_i \sum_k d_{ik}}{N_{(uv)}N_w} \quad (5)$$

d_{ik} adalah jarak antara objek ke- i dalam *cluster* (UV) dan objek ke- k dalam *cluster* ke W, dan $N_{(uv)}$ dan N_w adalah jumlah objek dalam *cluster* (UV) dan W.

4. Ward's Method

Metode *agglomerative* yang sering digunakan pada data kontinu adalah metode Ward. Pendekatan ini menggabungkan *cluster* jika total ketidaksamaan kuadrat dengan pusat *cluster* minimum disemua kemungkinan pilihan penggabungan [3]. Ilustrasi *ward's method* dapat dilihat pada gambar 5.



Gambar 5. Nested Cluster Ward's Method [13]

$$ESS = \sum_{j=1}^n x_j^2 - \frac{1}{n} (\sum_{j=1}^n x_j)^2 \quad (6)$$

2.4 Metode Partitioning

Metode *partitioning* menghasilkan suatu partisi data, bukan struktur pengelompokan seperti yang dihasilkan oleh metode hierarki. Kelebihan dari metode ini yaitu dapat mengelompokkan data yang cukup besar dengan waktu komputasi yang efisien [14].

Metode *partitioning* dapat dibagi menjadi dua yaitu *k-clustering* dan *self-determining*. Salah satu metode *k-clustering* adalah K-Means [15]. Prosedur algoritma ini sederhana dan mudah dalam mengklasifikasikan sekumpulan data kedalam sejumlah kluster (asumsi k kluster) yang telah ditetapkan sebelumnya [16].

Langkah-langkah algoritma K-Means sebagai berikut [17]:

1. Tentukan titik pusat dan jumlah kluster.
2. Tentukan titik pusat terdekat dari masing-masing data kemudian masukkan data tersebut ke kluster terdekat.
3. Tentukan ulang letak titik pusat berdasarkan anggota kluster yang baru terbentuk.
4. Jika belum konvergen, ulangi langkah 2 dan 3.

Kriteria konvergen adalah ketika suatu iterasi tidak ada lagi perubahan anggota pada tiap kluster dibandingkan iterasi sebelumnya, atau nilai jumlah *squared error* telah mencapai nilai minimum [17].

2.5 Validasi Cluster

Indeks validitas dihitung berdasarkan dua hal, yaitu [18]:

1. *Compactness*: kesamaan objek pada kluster yang sama.
2. *Separation*: perbedaan objek pada kluster yang berbeda.

Validasi *cluster* yang digunakan untuk menentukan jumlah *cluster* optimal terdiri dari 3 macam yaitu validasi internal, validasi stabilitas, dan validasi biologis [19]. Validasi *cluster* yang digunakan pada penelitian ini adalah validasi internal dan validasi stabilitas sebagai metode untuk memvalidasi hasil *cluster*.

2.5.1 Indeks Connectivity

Indeks *Connectivity* memiliki nilai diantara 0 sampai tak hingga. Semakin kecil nilai Indeks *Connectivity* semakin baik *cluster* yang terbentuk.

$$\text{Conn}(C) = \sum_{i=1}^N \sum_{j=1}^L X_{i,nni(j)} \quad (7)$$

$nn_{i(j)}$ = pengamatan tetangga terdekat i ke j

L = parameter yang menentukan jumlah tetangga yang berkontribusi untuk nilai indeks.

2.5.2 Indeks Silhouette

Indeks *Silhouette* mengukur derajat kepercayaan dalam proses *clustering* pada pengamatan tertentu dengan *cluster* yang dikatakan terbentuk baik bila nilai indeks mendekati 1 dan kondisi sebaliknya jika nilai indeks mendekati angka -1 [20].

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (8)$$

$a(i)$ = jarak rata-rata antara i dan seluruh pengamatan lainnya pada *cluster* yang sama.

$b(i)$ = jarak rata-rata antara i dengan pengamatan *cluster* terdekat.

2.5.3 Indeks Dunn

Indeks *Dunn* adalah rasio jarak terkecil antara observasi pada *cluster* yang berbeda dengan jarak terbesar pada masing-masing *cluster* data [20].

$$C = \frac{d_{\min}}{d_{\max}} \quad (9)$$

d_{\min} = jarak terkecil antara observasi pada *cluster* yang berbeda

d_{\max} = jarak terbesar pada masing-masing *cluster* data

Semakin tinggi nilai indeks *Dunn*, semakin baik *cluster* yang dihasilkan. Karena perhitungan hanya menggunakan minimum dan maksimum, maka indeks ini rentan terhadap *outlier* dan *noise* [21].

2.5.4 Indeks Davies-Bouldin

Indeks *Davies-Bouldin* memaksimalkan jarak *inter cluster* sekaligus meminimalkan jarak *intra cluster*. Semakin kecil nilai indeks menunjukkan semakin baik *cluster* yang dihasilkan [22].

2.5.5 Validasi Stabilitas

Validasi stabilitas membandingkan hasil dari analisis *cluster* berdasarkan penghapusan setiap kolom pada data, satu per satu. Pada validasi ini digunakan nilai *Average proportion of non-overlap* (APN), *average distance* (AD), *average distance between means* (ADM), dan *figure of merit* (FOM). Semakin kecil nilai APN, AD, ADM dan FOM, semakin baik *cluster* yang dihasilkan [19].

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Statistik

Berikut adalah deskripsi mengenai variabel kemiskinan seluruh provinsi di Indonesia berdasarkan ukuran pemusatan dan ukuran penyebarannya.

Tabel 2. Deskripsi Statistik

Variabel	Mean	Minimum	Maximum
X1	10,455	3,47	27,53
X2	4,729E+08	3,972E+07	2,841E+09

X3	8,544	6,65	11,06
X4	10916	7336	18527
X5	4,712	1,57	8,11
X6	69,43	64,38	74,76
X7	96,32	78	99,8
X8	78,66	47,12	89,04

Tabel 2 menunjukkan deskripsi statistik dari kedelapan variabel penelitian. Terlihat perbedaan satuan dari tiap variabel sehingga perlu dilakukan standarisasi untuk semua variabel.

3.2 Analisis Faktor

Sebelum mengelompokkan tiap provinsi berdasarkan karakteristik kemiskinannya, perlu dilakukan pengecekan asumsi bahwa data cukup secara statistik dan tidak terdapat multikolinearitas. Jika terdapat multikolinearitas maka harus dilakukan analisis faktor terlebih dahulu. Uji KMO dilakukan untuk memeriksa kecukupan sampel.

Hipotesa:

H_0 : Data layak untuk dianalisis

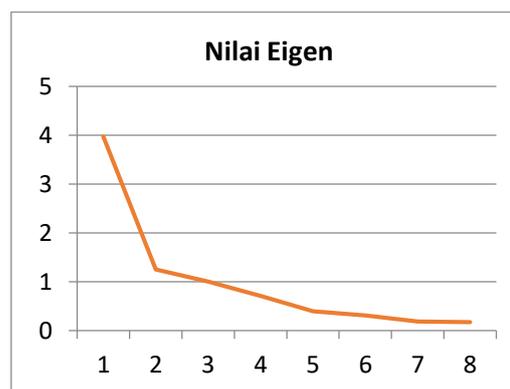
H_1 : Data tidak layak untuk dianalisis

Tabel 3. Nilai KMO dan Barlett Test

Uji	Nilai
<i>Kaiser Meyer Olkin (KMO)</i>	0,7445
<i>Barlett Test</i>	Approx. 126,405
Chi-Square	
df	28
sig	2.2e-16

Berdasarkan tabel di atas, nilai KMO yang dihasilkan adalah sebesar 0,7445 dan lebih dari 0,5 sehingga keputusannya gagal tolak H_0 . Maka dapat disimpulkan bahwa data cukup untuk dianalisis lebih lanjut.

Kemudian dilakukan uji *Barlett* untuk melihat ada atau tidaknya multikolinearitas. Berdasarkan tabel 3 terlihat bahwa nilai *p-value* signifikan. Oleh karena itu sudah cukup bukti untuk menyatakan bahwa variabel-variabel saling berkorelasi (multikolinearitas) sehingga perlu dilakukan *principal component analysis*.



Gambar 6. Diagram Nilai Eigen

Dalam analisis faktor dilakukan reduksi variabel dengan metode *principal component analysis*. Nilai eigen menunjukkan besarnya total varian yang dapat dijelaskan oleh komponen. Syarat dalam pemilihan komponen utama yaitu nilai eigen lebih besar atau sama dengan 1. Berdasarkan gambar 6, dapat diketahui tiga nilai eigen yang lebih besar dari 1. Oleh karena itu ketiga komponen secara bersama-sama menjelaskan keragaman dari 8 variabel sebesar 77,8%. Nilai *loading factor* dari ketiga komponen utama dapat dilihat pada tabel 4.

Tabel 4. *Loading Factor*

Variabel	PC1	PC2	PC3
X1	-0,373	0,177	-0,436
X2	0,273	-0,414	-0,585

X3	0,415	0,298	0,089
X4	0,410	-0,311	0,040
X5	0,269	0,426	-0,507
X6	0,339	-0,442	0,163
X7	0,327	0,442	0,346
X8	-0,386	-0,191	0,229

3.3 Validasi Cluster

Sebelum melakukan analisis *cluster*, dilakukan uji validitas *cluster* terlebih dahulu untuk mengetahui jumlah *cluster* terbaik. Uji validitas internal menggunakan indeks *Connectivity*, *Sillhoutte*, dan *Dunn*. Semakin kecil nilai *Connectivity* semakin baik *cluster* yang didapatkan. Jumlah *cluster* terbaik ditunjukkan dengan nilai indeks *Sillhouette* dan *Dunn* yang semakin mendekati 1. Nilai indeks validitas internal dapat dilihat pada tabel 5.

Tabel 5. Nilai Indeks Validitas

\sum Klaster		2	3	4	5	6
Hierarki	Connectivity	2,929	5,8579	15,9111	16.4290	20.4119
	Dunn	0.5130	0.5160	0.2202	0.2202	0.2314
	Silhouette	0.5696	0.4552	0.2925	0.2639	0.2718
	Davies Bouldin	0.2802	0.2975	0.7633	0.7953	0.7752
K-Means	Connectivity	13.1913	12.6988	17.7726	21.6345	26.6869
	Dunn	0.0649	0.2056	0.1624	0.1892	0.2054
	Silhouette	0.2821	0.3691	0.2602	0.2607	0.3010
	Davies Bouldin	1.1671	1.0145	1.0513	0.8966	0.7450

Tabel 6. Nilai Optimal

	Score	Method	Clusters
Connectivity	2.929	Hierarki	2
Dunn	0.516	Hierarki	3
Silhouette	0.569	Hierarki	2
Davies Bouldin	0.2802	Hierarki	2

Berdasarkan tabel 6 terlihat bahwa metode hierarki dengan jumlah *cluster* 2 yang paling optimal. Nilai *Connectivity* sebesar 2.929, nilai *Sillhoutte* sebesar 0.569 dan nilai Indeks *Davies*

Bouldin sebesar 0.2802. Selain itu untuk metode yang sama dengan jumlah *cluster* 3 nilai Indeks *Dunn* yang paling mendekati 1. Perbedaan nilai Indeks *Dunn* untuk metode hierarki antara *cluster* 2 dan 3 hanya sebesar 0.003.

Tabel 7. Nilai Indeks Validitas

\sum Klaster		2	3	4	5	6
Hierarki	APN	0.0689	0.1397	0.1165	0.2938	0.2621
	AD	2.7302	2.5769	2.3016	2.1977	2.0202

	ADM	0.2455	0.5169	0.5958	0.9675	0.8606
	FOM	1.3610	1.3563	1.3457	1.3551	1.3283
K-Means	APN	0.0964	0.2620	0.3235	0.4524	0.4335
	AD	2.5745	2.4976	2.3323	2.2223	2.0128
	ADM	0.5490	0.8596	1.1607	1.3873	1.2944
	FOM	1.3924	1.3718	1.3680	1.3732	1.3611

Tabel 8. Nilai Optimal

	Score	Method	Clusters
APN	0.0689	hierarchical	2
AD	2.0127	kmeans	6
ADM	0.2454	hierarchical	2
FOM	1.3283	hierarchical	6

Setelah itu dilakukan uji validitas stabilitas menggunakan nilai APN, AD, ADM dan FOM. Nilai indeks validitas stabilitas dapat dilihat pada tabel 7. Berdasarkan tabel 8 nilai APN dan ADM menunjukkan metode hierarki dengan jumlah *cluster* 2 adalah metode terbaik. Sedangkan hanya nilai AD dan FOM yang menunjukkan metode K-Means dengan jumlah *cluster* 6 merupakan metode yang terbaik.

3.4 Analisis Cluster

Hasil dari uji validitas internal dan stabilitas menunjukkan bahwa metode terbaik yang digunakan adalah hierarki dengan jumlah *cluster* 2. Hal ini dikarenakan nilai indeks *Connectivity* dan *Davies Bouldin* yang lebih kecil. Selain itu nilai indeks *Sillhoutte* dan *Dunn* juga mendekati 1. Nilai APN dan ADM juga lebih kecil ketimbang metode dengan jumlah *cluster* lainnya. Sehingga peneliti menggunakan metode hierarki dengan jumlah *cluster* 2 pada data kemiskinan tahun 2019.

3.4.1 Single Linkage, Complete Linkage, Average Linkage

Berikut hasil pengelompokan provinsi-provinsi di Indonesia menggunakan metode *single linkage*, *complete linkage* dan *average linkage*.

Kelompok 1 : Aceh, Sumatera Utara, Sumatera Barat, Riau, Kepulauan Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kep.Bangka Belitung, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, Banten, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Selatan,

Kalimantan Timur, Kalimantan Tengah, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Papua.
 Kelompok 2 : DKI Jakarta.

3.4.2 Ward's Method

Berikut hasil pengelompokan provinsi-provinsi di Indonesia menggunakan metode *Ward*.

Kelompok 1 : Aceh, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kep.Bangka Belitung, Jawa Tengah, DI Yogyakarta, Jawa Timur, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Selatan, Kalimantan Tengah, Kalimantan Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Papua.

Kelompok 2 : Sumatera Utara, Sumatera Barat, Riau, Kepulauan Riau, DKI Jakarta, Jawa Barat, Banten, Kalimantan Timur, Sulawesi Utara.

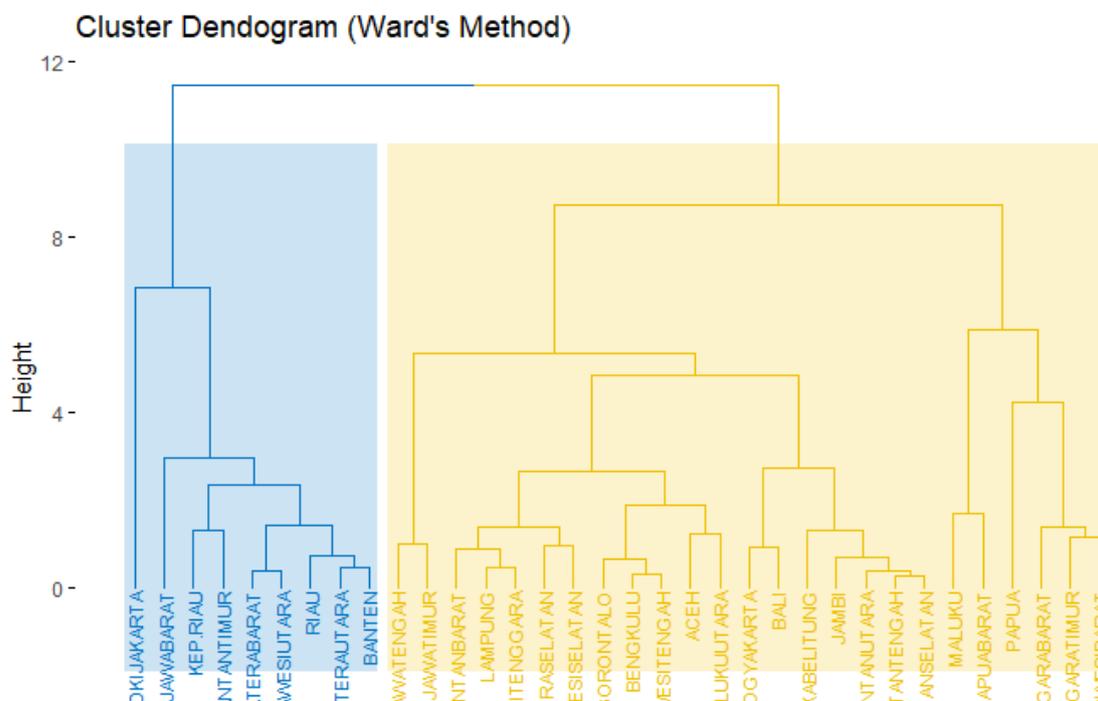
Tabel 9. Agglomerative Coefficient Metode Hierarki

Metode	Agglomerative Coefficient
Single	0.7759
Complete	0.8721618
Average	0.8303338
Ward	0.8969488

Berdasarkan tabel 9 didapatkan bahwa metode hierarki yang menghasilkan *cluster* terbaik yaitu

metode Ward karena nilai *agglomerative coefficient* lebih besar ketimbang metode hierarki lainnya.

Dendrogram yang dihasilkan dari metode Ward seperti pada gambar 7.



Gambar 7. Dendrogram Metode Ward

3.5 Karakteristik Tiap Cluster

Cluster yang terbentuk oleh metode Ward, perlu diketahui karakteristik dari setiap cluster berdasarkan variabel-variabel kemiskinan.

Tabel 10. Deskripsi Statistik tiap Cluster

Variabel	Cluster 1	Cluster 2
Persentase Kemiskinan	0.2590	-0.7194
PDRB	-0.2466	0.6849
Lama Sekolah	-0.3355	0.9321
Pengeluaran TPT	-0.26158	0.72660
Angka Harapan Hidup	-0.4174	1.1594
Angka Melek Huruf	-0.18495	0.51374
Persentase Rumah Tangga dengan Status Kepemilikan Sendiri	-0.2259	0.6276
	0.37061	-1.0295

1. Kelompok 1 terdiri dari 25 provinsi yang memiliki rata-rata lebih tinggi pada variabel persentase kemiskinan, dan persentase rumah tangga dengan status kepemilikan sendiri.

2. Kelompok 2 terdiri dari 9 provinsi yang memiliki rata-rata lebih tinggi pada variabel PDRB, rata-rata lama sekolah, pengeluaran perkapita, tingkat pengangguran terbuka, angka harapan hidup, dan angka melek huruf.

4. KESIMPULAN

Berdasarkan hasil validasi internal dan validasi stabilitas didapat metode terbaik yaitu hierarki dengan jumlah cluster 2. Hasil akhir diperoleh untuk cluster 1 berjumlah 25 provinsi sedangkan cluster 2 berjumlah 9 provinsi. Karakteristik cluster 1 yaitu tingkat kemiskinan yang lebih tinggi.

Dengan mengetahui karakteristik kemiskinan tiap provinsi, pemerintah dapat dengan tepat dalam membuat kebijakan untuk mengatasi masalah kemiskinan. Untuk penelitian berikutnya akan lebih baik jika menambahkan metode clustering lainnya.

DAFTAR PUSTAKA

- [1] Badan Pusat Statistik, "Berita Resmi Statistik" *Berita Resmi Statistik No.08/01/Th. XXII*, Januari 2020.
- [2] R. Xu and D. Wunsch, "Clustering", *John Wiley & Sons*, vol. 10., 2008.

- [3] S. Landau and I. C. Ster, "Cluster analysis: overview". *Elsevier Ltd.*, vol. 11, pp. 1-12, 2010.
- [4] Z. Abidin, "Pengelompokan Kabupaten/Kota Di Jawa Timur Berdasarkan Indikator Kemiskinan Dengan Menggunakan Analisis Cluster Hierarki", Doctoral dissertation, Surabaya : Institut Teknologi Sepuluh Nopember, 2018.
- [5] L. Ramadhani, I. Purnamasari, and F. D. T. Amijaya, "Penerapan Metode Complete Linkage dan Metode Hierarchical Clustering Multiscale Bootstrap", *J. Eksponensial*, vol. 9, no. 1, pp. 1-10, 2018.
- [6] N. I. Febianto and N. Palasara, "Analisa Clustering KMeans Pada Data Informasi Kemiskinan Di Jawa Barat Tahun 2018", *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 8, no. 2, pp. 130-140, 2019.
- [7] W. Alwi and M. Hasrul, "Analisis Klaster Untuk Pengelompokan Kabupaten/Kota Di Provinsi Sulawesi Selatan Berdasarkan Indikator Kesejahteraan Rakyat", *J. MSA (Matematika dan Stat. serta Apl.)*, vol. 6, no. 1, p. 35, 2018.
- [8] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*, USA: Pearson Education, Inc , 2007.
- [9] D. Bartholomew, M. Knott, & I. Moustaki, "Latent variable models and factor analysis: A unified approach (Vol. 904)", John Wiley & Sons, 2011
- [10] S. Machfudhoh and N. Wahyuningsih, "Analisis Cluster Kabupaten/Kota Berdasarkan Pertumbuhan Ekonomi Jawa Timur," *Jurnal Sains dan Pomits*, vol. 2, no. 1, pp. 1-7, 2013.
- [11] K. Yeturu, "Machine learning algorithms, applications, and practices in data science," in *Handbook of Statistics*, vol. 43, Elsevier, 2020, pp. 81-206, 2020.
- [12] C. Suhaeni, A. Kurnia, and R. Ristiyanti, "Perbandingan Hasil Pengelompokan menggunakan Analisis Cluster Berhierarchy, K-Means Cluster, dan Cluster Ensemble (Studi Kasus Data Indikator Pelayanan Kesehatan Ibu Hamil)", *J. Media Infotama*, vol. 14, no. 1, 2018.
- [13] Castillo, "Hierarchical Clustering C ", [Online]. <https://www.slideshare.net/ChaToX/hierarchical-clustering-56364612> [Diakses 22 Desember 2015]
- [14] R. Pramudita, "Pengujian Black Box pada Aplikasi Ecampus Menggunakan Metode Equivalence Partitioning", *INFORMATICS FOR EDUCATORS AND PROFESSIONAL: Journal of Informatics*, vol 4, no. 2, pp. 193-202, 2020
- [15] D. Zhang, K. Lee, K., & I. Lee, "Hierarchical trajectory clustering for spatio-temporal periodic pattern mining". *Expert Systems with Applications*, no. 92, pp. 1-11, 2018
- [16] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering", *International Journal of Advanced Research in Computer Science and Management Studie*, vol. 1, no. 6, pp. 90-95, 2013.
- [17] A. K. Jain, M. N. Nurty, and P. J. Flynn, 1999, Data Clustering: A Review, *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264-323.
- [18] Y. Liu, Z. Li, H. Xiong, X. Gao and J. Wu, "Understanding of Internal Clustering Validation Measures", in 2010 *IEEE International Conference on Data Mining*, pp. 911-916, 2010.
- [19] M.A. Nahdliyah, T. Widiharah, & A. Prahutama, "Metode K-Medoids Clustering dengan Validasi Silhouette Index dan C-Index (Studi Kasus Jumlah Kriminalitas Kabupaten/Kota di Jawa Tengah Tahun 2018)". *Jurnal Gaussian*, vol. 8, no. 2, pp. 161-170, 2019.
- [20] N. N. Halim and E. Widodo, "Clustering Dampak Gempa Bumi di Indonesia Menggunakan Kohonen Self Organizing Maps", in *Prosiding SI MaNIs (Seminar Nasional Integrasi Matematika dan Nilai-Nilai Islami)*, vol. 1, no. 1, pp. 188-194, 2017.
- [21] J. Baarsh and M. E. Celebi, "Investigation of Internal Validity Measures for K-Means Clustering", in *Proceedings of the international multiconference of engineers and computer scientists*, vol. 1, pp. 14-16, 2012.
- [22] B. J. D. Sitompul, O. S. Sitompul, and P. Sihombing, "Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm," in *Journal of Physics: Conference Series*, vol. 1235, no. 1, 2019.