Komputika: Jurnal Sistem Komputer
Volume 13, Nomor 1, April 2024, hlm. 41 - 53
Terakreditasi Peringkat 3, SK No. 148/M/KPT/2020
DOI: **10.34010/komputika.v13i1.10803**

ISSN: 2252-9039 (print)
ISSN: 2655-3198 (online)

Pemodelan *Clustering Ward, K-Means, DIANA*, dan PAM dengan PCA untuk Karakterisasi Kemiskinan Indonesia Tahun 2021

Kautsar Hilmi Izzuddin[1], Arie Wahyu Wijayanto[2*]

[1)] Program Studi D-IV Statistika, Politeknik Statistika STIS, Jakarta, Indonesia
[2)] Program Studi D-IV Komputasi Statistik, Politeknik Statistika STIS, Jakarta, Indonesia
Jl. Otto Iskandardinata No.64C Jakarta 13330
**email: ariewahyu@stis.ac.id*

(Naskah masuk: *16 Agustus 2023*; diterima untuk diterbitkan: *29 Nopember 2023*)

ABSTRAK *– Kemiskinan menjadi permasalahan yang serius dan cukup kompleks. Kemiskinan dipengaruhi secara lintas sektor dari berbagai faktor. Kemiskinan di Indonesia pada tahun 2021 mencapai angka 10,14% yang merupakan tertinggi disebabkan adanya pandemi coronavirus disease. Pengelompokan kemiskinan dapat dilakukan untuk perencanaan dan evaluasi program kemiskinan, Analisis cluster dengan metode partitioning clustering, yaitu k-means dan Partitioning Around Medoids (PAM), dan metode hierarchical clustering, yaitu Ward, dan Divisive Analysis (DIANA) dapat dimanfaatkan dalam pengelompokan provinsi di Indonesia berdasarkan enam indikator kemiskinan yaitu persentase penduduk miskin (P0), indeks kedalaman kemiskinan (P1), indeks keparahan kemiskinan (P2), Tingkat Pengangguran Terbuka (TPT), Angka Melek Huruf (AMH), dan Rata-rata Lama Sekolah (RLS). Berdasarkan evaluasi model, didapatkan model terbaik cluster dengan pendekatan ward dengan analisis Principal Component Analysis (PCA). Cluster 1 mencakup provinsi dengan tingkat pendidikan yang baik, Cluster 2 merupakan golongan pengangguran yang tinggi, Cluster 3 golongan dengan kemiskinan yang paling rendah, Cluster 4 mencakup tingkat kemiskinan sedang, dan Cluster 5 dengan tingkat kemiskinan yang parah. Kontribusi penelitian ini membuktikan metode PCA dapat memaksimalkan performa model clustering dalam menentukan karakterisasi kemiskinan di Indonesia tahun 2021. Model cluster ward membentuk lima cluster yang optimal dengan provinsi tingkat kemiskinan sangat rendah hingga sangat tinggi.*

Kata Kunci *– Analisis Cluster; Kemiskinan; PCA; Cluster Ward; Karakterisasi*

Clustering Ward, K-Means, DIANA, and PAM Modeling with PCA for Characterization of Indonesian Poverty in 2021

ABSTRACT *– Poverty is a serious and quite complex problem. Poverty is influenced across sectors by various factors. Poverty in Indonesia in 2021 reached 10,14%, which is the highest due to the coronavirus disease pandemic. Poverty grouping can be done for planning and evaluating poverty programs. Cluster analysis using partitioning clustering methods, such as k-means and Partitioning Around Medoids (PAM), and hierarchical clustering methods, such as Ward, and Division Analysis (DIANA) can be used to group provinces in Indonesia based on six poverty indicators, namely the percentage of poor people (P0), poverty depth index (P1), poverty severity index (P2), Open Unemployment Rate (TPT), Literacy Rate (AMH), and Average Years of Schooling (RLS). Based on the evaluation of the model, the best cluster model was obtained using the ward approach with Principal Component Analysis (PCA) analysis. Cluster 1 contains provinces with a good level of education, Cluster 2 is a group with high unemployment, Cluster 3 is a group with the lowest poverty, Cluster 4 has a moderate level of poverty, and Cluster 5 has a severe level of poverty. The contribution of this research proves that the PCA method can maximize the performance of the clustering model in determining the characteristics of poverty in Indonesia in 2021. The cluster ward model forms five optimal clusters with provinces with very low to very high poverty rates.*

Keywords *– Cluster Analysis; Poverty; PCA; Ward Clustering; Characterization*

1. PENDAHULUAN

Kemiskinan menjadi sebuah permasalahan cukup serius dari generasi ke generasi terutama pada negara berkembang. Kondisi kemiskinan terjadi Ketika pada suatu negara keadaannya berada dibawah standar minimum baik makanan maupun bukan makanan atau garis kemiskinan [1]. Berdasarkan buku Indikator Kesejahteraan Rakyat dari BPS, masalah kemiskinan menjadi persoalan pokok bangsa Indonesia yang menjadi prioritas pemerintah. Masalah ini menjadi masalah yang sangat kompleks sehingga pengentasan kemiskinan perlu dilakukan secara komprehensif dengan melingkupi berbagai aspek kehidupan di masyarakat [2]. Kondisi masyarakat disebut miskin ketika akses terhadap sarana dan prasana tidak memadai dan kualitas tempat tinggal jauh dibawah standar kelayakan dengan pekerjaan yang tidak menentu. Standar penggolongan kemiskinan didasari pada tingkat pendapatan seseorang untuk dapat memenuhi kebutuhan pokoknya [3]. Indonesia dengan tercakup 34 provinsi didalamnya per 2021 memiliki jumlah penduduk miskin mencapai angka 10,14%. Angka tersebut merupakan angka yang paling tinggi daripada periode 3 tahun kebelakang yang disebabkan adanya krisis pandemi *coronavirus disease*.

Kemiskinan adalah masalah multidimensi yang dipengaruhi secara lintas sektor dan berbagai faktor baik tingkat pendidikan, pendapatan hingga kondisi lingkungan dan sosial masyarakat. Menurut *World Bank* terdapat tiga faktor yang dapat menyebabkan kemiskinan yaitu rendahnya aset dan pendapatan dalam kebutuhan dasar, ketidakmampuan dalam bersuara dan menunjukkan dirinya, serta rentan terhadap guncangan ekonomi yang digambarkan dalam indikator [4]. Indikator yang dapat dijadikan dalam mengukur tingkat kemiskinan salah satunya adalah Angka Melek Huruf (AMH). Semakin besar nilai AMH maka semakin tinggi nilai dan kualitas SDM yang ada di masyarakat karena dengan kemampuan baca tulis, masyarakat dapat memiliki kemampuan dan keterampilan untuk menyerap informasi. Jika dilihat dari tingkat ketenagakerjaan dapat digunakan pendekatan Tingkat Pengangguran Terbuka (TPT) yang mengukur keadaan angkatan kerja dan tenaga kerja. Jika dilihat dari pendidikan sebagai tujuan dasar pembangunan, Rata-rata Lama Sekolah (RLS) dapat dijadikan indikasi semakin tingginya pendidikan formal yang diraih oleh masyarakat di sebuah provinsi.

Ukuran dalam menggambarkan kemiskinan yang dihitung BPS didekati dengan tiga indikator. Indikator tersebut adalah *Head Count Index* (P0), *Poverty Gap Index* (P1), dan *Distributionally Sensitive Index* (P2). Nilai P0 menunjukkan semakin kecilnya nilai tersebut, maka akan semakin menurun jumlah penduduk yang berada di bawah garis kemiskinan atau *poverty line*. Nilai P1 menunjukkan semakin kecilnya nilai tersebut, maka secara rata-rata pendapatan penduduk miskin memiliki nilai yang semakin mendekati garis kemiskinan. Nilai P2 menunjukkan semakin kecil angkanya, maka distribusi pendapatan masyarakat akan semakin merata. Indikator kemiskinan ini menggambarkan bagaimana sebaran kemiskinan di berbagai provinsi sehingga dapat dijadikan indikator pengelompokan berdasarkan jumlah persentase kemiskinan, rata-rata kesenjangan dan penyebaran pengeluaran.

Analisis *cluster* dapat digunakan dalam pengelompokan objek atau data menjadi beberapa kelompok yang disesuaikan berdasar kemiripan variabel yang diamati. Pengelompokan provinsi berdasarkan indikator kemiskinan dapat dijadikan analisis untuk perencanaan dan evaluasi sasaran program kemiskinan pemerintah. Pemanfaatan algoritma *clustering* baik *Hierarchial Agnes* maupun *Diana*, *k-means*, dan PAM digunakan dengan meninjau enam indikator kemiskinan di daerah. Maka pada penelitian ini akan dilakukan perbandingan analisis menggunakan beberapa metode *cluster* dan menambahkan metode *Principal Component Analysis* (PCA) untuk karakterisasi tingkat kemiskinan di setiap provinsi di Indonesia dengan menggunakan indikator kemiskinan yang disusun oleh BPS. Penggunaan PCA dapat dilakukan untuk memilih fitur yang sekiranya relevan sehingga pengelompokan *cluster* dapat bekerja dengan baik pada data dengan dimensi yang rendah [5]. PCA dapat digunakan sebagai indeks yang mencerminkan status sosial ekonomi [6].

Penelitian yang pernah dilakukan sebelumnya adalah analisis menggunakan *k-means* pada karakteristik kemiskinan di Jawa Barat pada tahun 2018 [7] serta pemodelan *cluster* kemiskinan di provinsi Indonesia pada 2019 oleh Afira [8]. Penelitian juga dilakukan oleh Repollo menggunakan *k-means* dalam membagi kelompok di negara Filipina berdasarkan kemiskinan [9]. Analisis *cluster* digunakan untuk membagi kelompok pada setiap negara di dunia yang termasuk kedalam kelompok OECD [10]. Selain itu juga terdapat penelitian berupa analisis *cluster* menggunakan metode *hard* dan *soft clustering* dalam mengelompokkan kesejahteraan kabupaten/kota di Jawa oleh Thamrin [11] dan membandingkan *cluster* metode *hierarchical* dan *partitioning* terhadap Indeks Pembangunan Manusia tahun 2019 oleh Sikana [12]. Penelitian Fauziyah dan Achmad menunjukkan *cluster hybrid k-means* dan *ward* memiliki hasil yang optimal dalam mengelompokkan kemiskinan di Jawa Barat [13]. Pada tahun 2021, kemiskinan Indonesia dapat dilihat yang paling terdampak

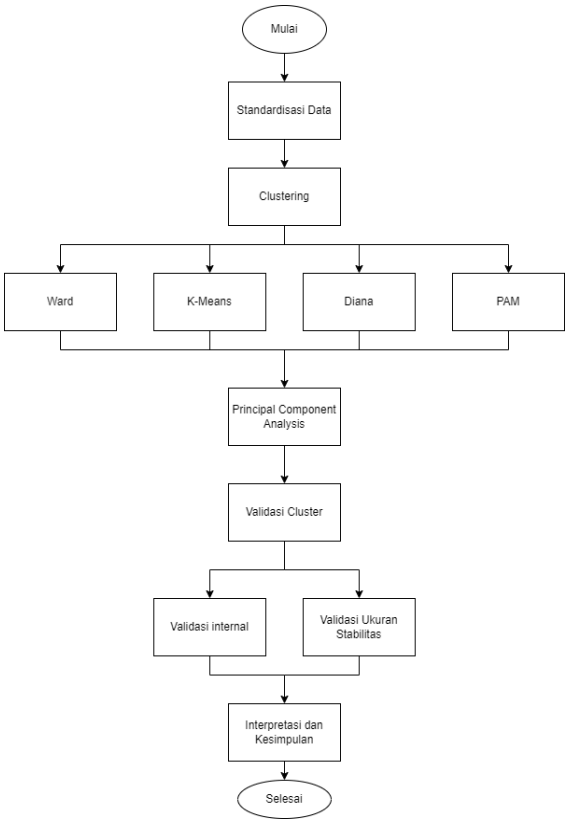
adalah provinsi Papua karena sektor utama yang berada di sektor pertambangan [14].

Berdasarkan Soemartini (2017) digunakan indikator mengenai angkatan kerja, pertumbuhan penduduk, angka harapan hidup (AHH), rata-rata lama sekolah (RLS) sebagai pengukur dalam *clustering* masyarakat di Jawa Barat. [15]. Astuti (2022) menyebutkan bahwa analisis multivariate melalui *cluster* dapat membagi distribusi pendidikan di daerah Sidoarjo Jawa Timur dan dapat melihat ketimpangannya [16]. Indikator yang digunakan dalam membagi *cluster* negara menurut kemakmurannya adalah kesehatan, kelayakan hidup dan pendidikan serta lingkungan dan manajemen [17].

Tujuan dari penelitian ini adalah membandingkan dan mengevaluasi model *cluster* seluruh indikator kemiskinan dengan *cluster* menggunakan pendekatan *Principal Component Analysis* (PCA), mengetahui pengelompokan tingkat kemiskinan tiap provinsi di Indonesia menggunakan metode *cluster* terbaik, dan mendeskripsikan karakterisasi provinsi miskin di Indonesia berdasarkan indicator kemiskinannya. Pada penelitian ini dilakukan komparasi beberapa metode *clustering* berupa ward, *k-means*, Diana, dan PAM. Kontribusi yang diberikan dalam penelitian ini adalah menentukan hasil yang optimal dalam pengelompokkan kemiskinan di Indonesia dengan membandingkan metode *clustering* mana yang terbaik dengan menggunakan pendekatan PCA. Metode *cluster* yang memberikan hasil model optimal dapat digunakan dalam karakterisasi indikator kesejahteraan masyarakat. Dengan itu, penggunaan metode karakterisasi yang terbaik dapat memberikan hasil yang sesuai dalam penentuan kebijakan pemerintah secara efektif.

2. METODE DAN BAHAN

Data yang digunakan dalam penelitian ini bersumber dari website Badan Pusat Statistik dalam statistik Sosial dan Kependudukan yang tercantum pada Tabel 1. Dalam hal ini, variabel persentase penduduk miskin (P0) menggambarkan penduduk yang mempunyai rata-rata pengeluaran dibawah garis kemiskinan. Variabel indeks kedalaman (P1) dan keparahan kemiskinan (P2) menggambarkan rata-rata dan sebaran pengeluaran pada penduduk miskin. Variabel Tingkat Pengangguran Terbuka (TPT) menunjukkan banyaknya pengangguran terhadap jumlah penduduk. Variabel Angka Melek Huruf (AMH) dan Rata-rata Lama Sekolah (RLS) menunjukkan kualitas penduduk dalam pendidikan. Variabel-variabel tersebut digunakan sebagai indikator kemiskinan di Indonesia. Data-data tersebut diolah menggunakan aplikasi R-Studio. Gambar 1 menunjukkan alur penelitian yang menggunakan *software* R-Studio dalam melakukan pemodelan *cluster*.



Gambar 1. Diagram Alir Metode Penelitian

Metode Cluster

Clustering merupakan cara dalam mengelompokkan data menjadi beberapa *cluster* atau grup [18]. Analisis *cluster* menjadi salah satu metode yang dapat mengidentifikasi sekelompok objek dengan karakteristik tertentu yang serupa sehingga dapat dipisahkan dari objek yang termasuk *cluster* lainnya. *Clustering* dilakukan dengan mengelompokkan objek sesuai dengan kesamaan jaraknya (*dissimilarities*) [19]. Dalam hal ini, objek yang berada di dalam *cluster* yang sama memliki sifat yang homogen dibandingkan dengan objek yang berada di *cluster* lainnya. Adapun *cluster* yang optimal adalah *cluster* dengan homogenitas yang tinggi antar anggota di dalam *cluster* (within) serta heterogenitas yang tinggi antar *cluster (between*). Dengan didapatkan sekumpulan pola pada obyek dalam *cluster*, akan membentuk kesamaan yang dapat dijadikan sebagai kesimpulan terbaik [20]. Secara umum, metode *cluster* terdiri atas dua tipe yaitu metode hierarki dan non hirearki.

Cluster Hierarki

Metode hierarki dilakukan dengan membagi obyek atau data menjadi dua kelompok yang memiliki kesamaan yang secara berturut-turut

dilanjutkan kepada obyek yang memiliki kesamaan kedua. Klasifikasi ini akan menghhasilkan pohon yang menggambarkan tingkatan dari kesamaan obyek yang ada [21]. Gambaran klasifikasi berupa pohon diilustrasikan berupa diagram yang disebut dendogram [22]. *Clustering* metode hierarki terdiri atas dua metode yaitu *agglomerative nesting* (AGNES) dan *divisive analysis* (DIANA). Metode *agglomerative* mengklasifikasikan obyek dari yang paling mirip hingga seterusnya membentuk sebuah *cluster.* Namun metode *divisive* merupakan kebalikan dari metode *agglomerative* dimana obyek disusun menjadi sebuah *cluster* yang melingkupi seluruh obyek hingga secara berturut-turut dilakukan pemisahan obyek berdasarkan kesamaan.

Terdapat beberapa metode yang digunakan dalam *clustering agglomerative* yaitu *single linkage, complete linkage, average linkage, centroid* dan *Ward* [18]. Perbedaan dari masing-masing metode tersebut adalah penghitungkan kesamaan (*similarity*) antar *cluster*.

1. *Single linkage*

Metode ini mendefinisikan kesamaan *cluster* berdasarkan jarak terdekat. Secara matematis perhitungan kesamaan dapat dirumuskan sesuai persamaan 1.

$$d(ab)c = \min\{dac, dbc\} \quad (1)$$

Dimana dac dan dbc menunjukkan jarak terdekat antara *cluster a* dengan *cluster b* dan *c.*

2. *Complete linkage*

Metode ini menghitung kesamaan berdasarkan jarak maksimum antar obyek setiap *cluster*. Secara matematis jarak maksimum dapat dirumuskan sesuai persamaan 2.

$$d(ab)c = \max\{dac, dbc\} \quad (2)$$

Persamaan (2) menunjukkan dac dan dbc merupakan jarak terjauh antar objek antara *cluster a* dengan *c*, serta *cluster b* dengan *c.*

3. *Average linkage*

Metode ini menentukan kesamaan berrdasarkan rata-rata antar seluruh individu. Perhitungan rata-rata tersebut sesuai dengan persamaan 3. Metode ini sering digunakan ketika metode *single* dan *complete* memiliki keterbatasan dalam menghitung jarak pada obyek yang memiliki kedekatan [23].

$$d(ab)c = \frac{n_a}{n_a+n_b} d_{ac} + \frac{n_b}{n_a+n_b} d_{bc} \quad (3)$$

Keterangan:
d_{ac} = jarak antar *cluster a* dan *c*
d_{bc} = jarak antar *cluster b* dan *c*
n_a = jumlah obyek pada *cluster a*
n_b = jumlah obyek pada *cluster b*

4. Metode *centroid*

Metode ini menghitung kesamaan didasarkan atas jarak obyek dengan titik pusat *cluster*. Penggunaan metode ini cukup baik dalam mengatasi outlier. Titik *centroid* yang terbentuk didapatkan sesuai persamaan 4.

$$\bar{x} = \frac{n_1\overline{x_1}+n_2\overline{x_2}}{n_1} \quad (4)$$

Dengan n_1 dan n_2 adalah jumlah obyek.

5. Metode *ward*

Metode *ward* adalah metode yang berbeda daripada metode sebelumnya, metode ini menghitung kesamaan berdasarkan jumlah kuadrat dalam *cluster* yang dijumlahkan dari seluruh variabel. [Metode ini merupakan metode yang efektif karena perhitungan jarak disesuaikan dengan nilai jumlah kuadrat errornya sehingga sering digunakan dalam menentukan metode *cluster* dengan metode lainnya [22]. Perhitungan kesamaan dijabarkan pada persamaan 5.

$$d(ab)c = \frac{(n_a+n_b)d_{ac}+(n_b+n_c)d_{bc}-n_c d_{ab}}{n_a+n_b+n_c} \quad (5)$$

Cluster Non Hierarki

Selain metode hierarki, *cluster* juga dapat diklasifikasikan berdasarkan metode *partitioning, density-based, grid-based,* dan *mode-based* [24]. Metode *partitioning* membentuk *cluster* dengan memecah obyek-obyek menjadi beberapa kelompok secara acak dengan menghitung *cost* dari *medoid* yang baru dengan lama [25]. Salah satu contoh metode *partitioning* adalah *cluster* menggunakan *k-means* dan *partitioning around medoids*. *K-means* merupakan metode yang mempartisi obyek ke dalam satu atau lebih kelompok. Metode ini dilakukan dengan menentukan jumlah *cluster* yang diinginkan dan selanjutnya obyek digabungkan sesuai keinginan jumlah *cluster* tersebut [26]. Metode *k-means* merupakan metode *cluster* sederhana dan sering digunakan karena mampu menangkap data yang cukup besar secara efektif dan efisien [27]. Perhitungan jarak antar setiap titik obyek dengan pusat *cluster* pada *k-means* dirumuskan dengan jarak *Euclidian* sesuai persamaan 8.

Partitioning Around Medoids (PAM) merupakan metode *cluster* dengan menggunakan nilai rata-rata obyek dalam suatu *cluster* sebagai titik referensinya dimana obyek yang paling terkonsentrasi dalam *cluster* disebut *medoid screened* [28]. Metode PAM memiliki kelebihan dalam mengatasi outlier yang berbeda dengan metode *k-means* dan tidak berdasarkan urutan obyek [29]. Algoritma yang digunakan dalam metode PAM adalah dengan menghitung jarak terdekat obyek *non-medoid* dengan *medoid* yang baru, sehingga didapatkan hasil selisih simpangan yang memiliki nilai lebih dari nol antar *medoid* [30]. Persamaan total simpangan [31] dalam *cluster* PAM dirumuskan sebagai persamaan 6.

$$Simpangan\ (S) = \text{cost}\ baru - cost\ lama \quad (6)$$

Validasi Cluster

Validasi *cluster* merupakan metode yang dilakukan dalam mengevaluasi model *clustering*

dengan tujuan menentukan jumlah *cluster* yang optimal [32]. Berdasarkan literatur, validasi *cluster* dapat dilakukan dengan validasi internal, validasi eksternal, analisis ukuran stabilitas dan validasi visual. Validasi internal merupakan cara validasi dengan menghitung indeks untuk mengukur kecocokan pengelompokan data berupa homogenitas *cluster* hingga struktur kedekatan data [33]. Sedangkan, validasi eksternal merupakan metode validasi dengan menggunakan informasi diluar model.

Selain itu, ukuran stabilitas menjadi bagian yang penting dalam menentukan validitas model *cluster* [34]. Ukuran stabilitas ini menandakan jika hasil *clustering* sebuah model diterapkan pada data dengan distribusi yang sama, maka akan menghasilkan *cluster* yang spesifik [35]. Pada penelitian kali ini digunakan metode validasi internal dan ukuran stabilitas. Metode kombinasi ini dilakukan untuk menentukan model *cluster* yang menghasilkan nilai validasi yang serupa serta partisi yang memiliki kemiripan, ketika *cluster* tersebut diterapkan kepada kumpulan data dengan distribusi yang sama.

TAHAPAN PENELITIAN

Tahapan yang dilakukan dalam karakterisasi kemiskinan di Indonesia tahun 2021 adalah sebagai berikut:

1. Menyusun subset data indikator kemiskinan

Data disusun dari enam indikator kemiskinan yang diambil sebagai data sekunder dari BPS. Data kemudian disusun dengan mengubah menjadi bentuk matriks *unlabeled data* untuk dapat dilakukan standardisasi data. Data yang digunakan dalam penelitian ini menggunakan tipe data numerik karena untuk *clustering* tidak direkomendasikan menggunakan tipe kategorik.

2. Melakukan normalisasi data menjadi data *scale*

Standardisasi data dilakukan dengan tujuan menyamakan skala pada setiap variabel. Hal ini dilakukan karena dalam perhitungan jarak atau *dissimilarity* sangat sensitif dengan perbedaan skala antara variabel pada data. Standardisasi data digunakan nilai *z-score* menggunakan persamaan 7.

$$z - score = \frac{x-\mu}{\sigma} \quad (7)$$

Keterangan:
x = nilai yang diamati
μ = rata-rata
σ = standar deviasi

3. Menentukan jarak dengan metode *Euclidean*

Dalam penyusunan *cluster* perlu memperhatikan jarak atau *dissimilarity.* Metode penghitungan yang digunakan pada penelitian ini adalah *Euclidean distance* [18]. Pemilihan metode ini didasarkan pada tujuan analisis *clustering* dengan menjadikan besaran nilai setiap variabel sebagai penjelas karakteristik yang membedakan tiap provinsi di Indonesia. *Euclidean distance* memiliki rumus, yaitu:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (8)$$

Keterangan:
d = jarak antar amatan
x_1 = titik latitude wilayah 1
x_2 = titik latitude wilayah 2
y_1 = titik longitude wilayah 1
y_2 = titik longitude wilayah 2

4. Evaluasi *cluster* menggunakan metode terbaik : Ward, K-Means, Diana, PAM

Evaluasi model yang dilakukan menggunakan ukuran validasi internal dan stabilitas. Ukuran internal yang digunakan berupa *connectivity, silhouette width* dan *Dunn index*. Dimana *connectivity* mengukur derajat koneksi dari *cluster* yang dilihat dari k-*nearest neighbor*. Ukuran *silhouette width* dan *dunn index* mengkombinasikan ukuran dari kepadatan dan separasi *cluster*. Dimana *silhouette width* adalah rata-rata dari masing-masing observasi yang diukur berdasarkan *degree of confidence* pada *cluster*. Sedangkan Dunn Index adalah rasio antara ukuran terkecil antar observasi yang tidak dalam satu *cluster* dengan ukuran terbesar jarak diantara *cluster* [36]. Ukuran stabilitas diukur berdasarkan average proportion of non-overlap (APN), average distance (AD), average distance between means (ADM), figure of merit (FOM). Nilai APN mengukur proporsi rata-rata jumlah observasi pada *cluster* yang saama antar data lengkap dan data yang dihilangkan satu variabel. Nilai AD mengukur jarak rata-rata provinsi di dalam *cluster* yang sama. ADM mengukur jarak rata-rata antar pusat *cluster*. Sedangkan, FOM mengukur variansi intra-*cluster* pada variabel yang dihapus. Nilai keempat kriteria tersebut ketika semakin kecil akan menghasilkan *cluster* yang semakin konsisten [37]. Adapun persamaan dari masing-masing kriteria stabilitas *cluster* ditunjukkan pada persamaan 9, 10, 11, dan 12.

$$APN = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\left(1 - \frac{n(C^{i,l} \cap C^{i,0})}{n(C^{i,0})}\right) \quad (9)$$

$$AD = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\frac{1}{n(C^{i,0})n(C^{i,l})}\left[\sum_{i \in C^{i,0}, j \in C^{i,l}} dist(i,j)\right] \quad (10)$$

$$ADM = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M} dist(\bar{x}_{Ci,l}, \bar{x}_{Ci,0}) \quad (11)$$

$$FOM = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\sum_{i \in Ck(l)} dist(\bar{x}_{i,l}, \bar{x}_{Ck(l)})} \quad (12)$$

Keterangan:
M = total observasi dalam kolom

N = total observasi dalam baris
i = jumlah observasi dalam *cluster* pada baris
l = jumlah observasi dalam *cluster* pada kolom
$C^{i,0}$ = *cluster* dengan observasi i menggunakan *clustering* asli
$C^{i,l}$ = *cluster* dengan observasi i menggunakan *clustering* pada data dengan kolom l dihapus
$\bar{x}_{Ci,l}$ = rata-rata observasi pada *cluster* yang mengandung observasi i pada data dengan kolom l dihapus
$\bar{x}_{Ci,0}$= rata-rata observasi pada *cluster* yang mengandung observasi i
$\bar{x}_{i,l}$ = nilai observasi ke-i pada kolom ke-l
$\bar{x}_{Ck(l)}$ = rata-rata dari cluster $Ck(l)$

5. Menyusun *cluster* dengan analisis *Principal Component Analysis* (PCA) pada masing-masing indikator kemiskinan.

Pada analisis *cluster*, jika jumlah fitur atau variabel yang digunakan cukup banyak, terdapat kemungkinan kasus *The Curse of Dimensionality*. Penggunaan banyak variabel pada model pencarian tetangga terdekat tidak dapat memberikan hasil yang terbaik. Hal ini juga akan memperburuk hasil analisis ketika *clustering* menggunakan matriks *Euclidean* yang disebabkan matriks jarak dengan dimensi yang tinggi tidak memiliki intuitif yang baik dibanding 2 atau 3 dimensi karena akan meratakan distribusi di beberapa titik [38].

Dalam mengatasi masalah ini, dilakukan metode penurunan dimensi yang digunakan dalam metode *clustering*. Hal ini akan menurunkan *noise* yang tidak penting dalam sebuah data. Reduksi data ini dapat dijadikan proses dalam *pre-processing* dengan tujuan tetap mempertahankan informasi penting dalam menganalisis sebuah data [39]. Maka dalam penelitian ini, akan dilakukan perbandingan antara model *clustering* tanpa proses *pre-processing* atau PCA dan dengan menggunakan PCA. Data dengan analisis PCA terlebih dahulu perlu dilakukan standardisasi data agar meminimalisir nilai ekstrem. Analisis PCA dilihat dengan nilai *eigen value* yang menjelaskan banyak variasi.

6. Membandingkan nilai *cluster* terbaik menggunakan PCA atau tanpa analisis PCA

Evaluasi *cluster* dengan PCA atau non-PCA dilakukan dengan analisis *average silhouette* yang digunakan untuk mempelajari jarak pemisahan antar cluster. Nilai ini dilihat dari Plot *silhouette* yang menampilkan ukuran kedekatan setiap titik dalam *cluster* dengan titik *cluster* tetangga. Nilai koefisien *silhouette* jika mendekati nilai 1 maka menunjukkan bahwa sampel memiliki jarak yang jauh dari *cluster* tetangga. Sedangkan, jika mendekati 0, sampel berada sangat dekat dengan *decision boundary* antar dua *cluster* yang bertetangga. Sehingga dalam evaluasi perbandingan model yang dilihat adalah nilai yang tertinggi.

Tabel 1. Daftar Variabel Penelitian

Variabel	Keterangan
X1	Persentase Penduduk Miskin (P0)
X2	Indeks Kedalaman Kemiskinan (P1)
X3	Indeks Keparahan Kemiskinan (P2)
X4	Tingkat Pengangguran Terbuka (TPT)
X5	Angka Melek Huruf (AMH)
X6	Rata-rata Lama Sekolah

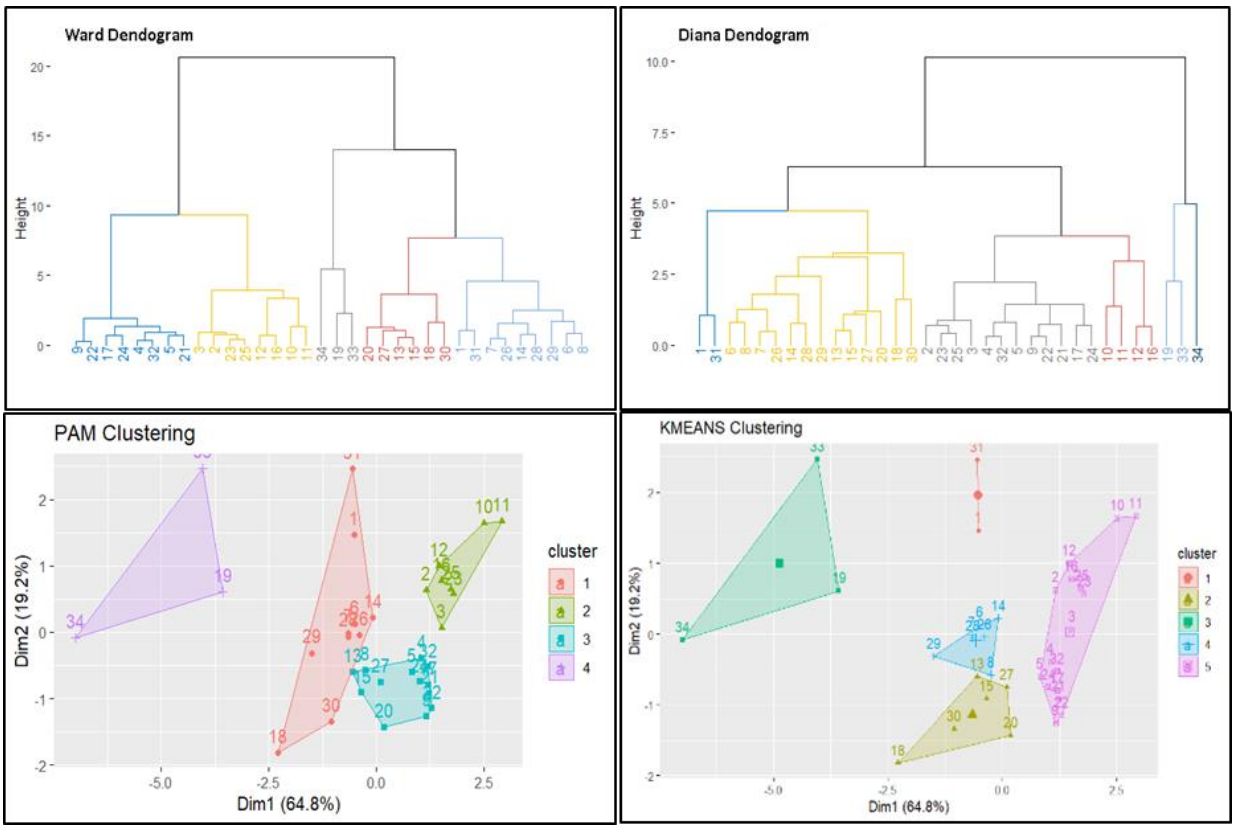
3. HASIL DAN PEMBAHASAN

EVALUASI MODEL TERBAIK

Dilihat pada Tabel 2, dengan menggunakan ukuran validasi berdasarkan ukuran internal *cluster* didapatkan bahwa seluruh ukuran menghasilkan jumlah yang beragam dari masing-masing metode. Pada metode ward, jumlah *cluster* yang digunakan adalah 3 atau 5. Untuk memudahkan analisis karakterisasi kemiskinan, digunakan jumlah pada *ward clustering* sejumlah lima *cluster*. Sedangkan pada metode K-Means sebanyak empat *cluster*, Diana sebanyak enam *cluster,* dan PAM sebanyak empat *cluster*. Pemilihan jumlah *cluster* yang lebih banyak disebabkan alasan penelitian untuk menemukan informasi lebih banyak lagi mengenai kategori kemiskinan yang ada di setiap provinsi di Indonesia.

Tabel 2. Pemilihan Model *Cluster* Terbaik Berdasarkan *Internal Measures*

Metode	***Internal Measures* (Jumlah *Cluster*)**		
	Connectivity	***Dunn***	**Silhouette**
Ward	0,781 (3)	0,899 (3)	0,832 (5)
K-Means	12,342 (3)	0,256 (4)	0,400 (3)
Diana	11,737 (3)	0,357 (6)	0,322 (6)
PAM	10,696 (3)	0,148 (3)	0,420 (4)



Gambar 2. Perbandingan *Cluster* Ward, Diana, PAM, dan K-Means

Tabel 3. Pemilihan Model *Cluster* Terbaik Berdasarkan *Stability Measures*

Metode	***Stability Measures* (Jumlah *Cluster*)**			
	APN	**AD**	**ADM**	**FOM**
Ward	0,094 (5)	1,682 (5)	0,643 (3)	0,739 (5)
K-Means	0,162 (5)	1,698 (5)	0,444 (5)	0,664 (5)
Diana	0,157 (3)	1,661 (6)	0,779 (3)	0,667 (6)
PAM	0,329 (4)	1,803 (4)	1,021 (3)	1,545 (4)

Dilihat pada Tabel 3, dengan menggunakan ukuran validasi berdasarkan ukuran stabilitas *cluster* didapatkan bahwa seluruh ukuran menghasilkan jumlah yang beragam dari masing-masing metode. Pada metode ward, jumlah *cluster* yang digunakan adalah 3 atau 5. Untuk memudahkan analisis karakterisasi kemiskinan, digunakan jumlah *cluster* pada *ward clustering* sejumlah lima *cluster*. Sedangkan pada metode *k-means* sebanyak empat *cluster*, Diana sebanyak enam *cluster,* dan PAM sebanyak empat *cluster*. Pemilihan jumlah *cluster* yang lebih banyak disebabkan alasan penelitian untuk menemukan informasi lebih banyak lagi mengenai kategori kemiskinan yang ada di setiap provinsi di Indonesia.

PERBANDINGAN HASIL *CLUSTER*

Hasil dendogram dari kedua metode pada Gambar 2 yaitu Ward dan Diana menunjukkan dendogram yang *compact*. Hal ini ditunjukkan dengan pembagian partisi yang cukup jelas sehingga pemotongan dendogram untuk membagi *cluster* lebih mudah. Metode *Ward* dan *Diana* sama-sama membentuk jumlah *cluster* yang optimal sebanyak lima *cluster*. Dari grafik *cluster* ward terdapat dua *cluster* besar yang dipecah kembali menjadi *cluster* yang lebih kecil, kemiripan antar provinsi berdasarkan tingkat kemiskinan terlihat. Sedangkan pada *cluster Diana* pada pembagian dua *cluster* diawal terdapat satu pohon yang membagi hanya menjadi 3 provinsi yaitu NTT, Papua Barat dan Papua. Sedangkan jika dilihat dari Gambar 2 hasil *clustering* dari metode *k-means* dan PAM menunjukkan jumlah *cluster* optimal yang berbeda, Hal ini disebabkan karena pendekatan metode PAM yang merupakan model *robust* dari *k-means* dan lebih tidak sensitif terhadap outlier. Metode PAM membagi titik-titik di sekitar medoid, sedangkan *k-means* menggunakan titik buatan. Sehingga pada *cluster k-means* terlihat untuk Aceh dan Maluku

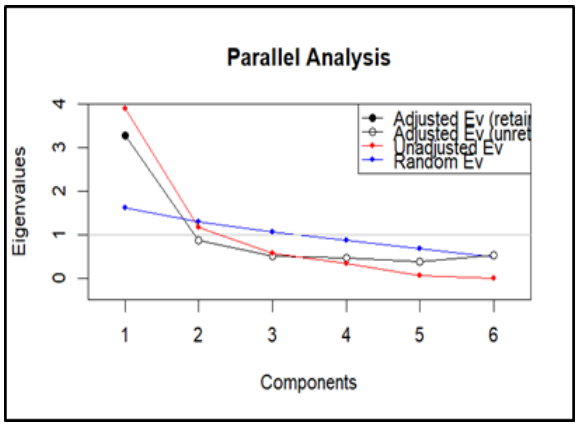
terpisah menjadi satu *cluster* yang berbeda. Dari gambar dapat dilihat *cluster* yang memiliki pembagian terbaik adalah *cluster ward* dan PAM.

Analisis PCA Pada *Clustering*

Tabel 4. Analisis Komponen Utama pada Variabel Kemiskinan

Important Components	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
Standard Deviasi	1,971	1,075	0,759	0,579	0,217
Eigenvalue	3,885	1,155	0,575	0,335	0,046
Proporsi Varians	0,647	0,193	0,096	0,056	0,008
Proporsi Kumulatif	0,647	0,840	0,936	0,992	0,999

Nilai varians kumulatif dari dua komponen utama adalah 0,84. Tetapi, nilai ini perlu dijelaskan dengan menggunakan *Kaiser Criterion* dan Grafik *Parallel Analysis* untuk membuktikan apakah lebih banyak komponen utama atau tidak [40]. Jika dilihat dari kriteria yang ada, *eigenvalues* yang memiliki nilai lebih dari satu adalah dua faktor. Sehingga berdasarkan hal itu, dua kriteria komponen utama dapat dipilih untuk menjelaskan karakteristik kemiskinan di Indonesia. Berdasarkan analisis kontribusi variabel terhadap dimensi pertama, variabel yang berkontribusi banyak terhadap komponen utama pertama adalah variabel ukuran kemiskinan P0, P1, dan P2 yang mengukur persentase kemiskinan di Indonesia. Sedangkan pada kontribusi variabel dimensi kedua, variabel yang berkontribusi adalah variabel indikator kemiskinan yaitu TPT dan RLS. Uniknya, AMH memiliki kontribusi terhadap dimensi kedua dengan nilai dibawah 20%. Hal ini sesuai fakta BPS yang menyebutkan bahwa AMH tidak relevan lagi dalam mengukur perhitungan IPM karena perubahan penyebaran data AMH di masing-masing provinsi di Indonesia.



Gambar 3. Grafik *Parallel Analysis*

Berdasarkan grafik *parallel analysis* pada Gambar 3 ditunjukkan bahwa cukup hanya satu faktor yang harus dipertahankan. Hanya faktor pertama yang memenuhi perysaratan *eigenvalues* yang diperoleh lebih tinggi daripada *eigenvalues* dari data yang acak. Grafik tersebut menunjukkan hasil yang berbeda dengan *Kaiser Criterion* yang menyarankan untuk menggunakan dua faktor. Namun, karena nilai varians yang dapat dijelaskan oleh dua faktor mencapai nilai 84%, penggunaan dua faktor dipilih untuk mendapatkan lebih banyak informasi mengenai karakterisasi kemiskinan di Indonesia.

Evaluasi Model Dengan PCA

Berdasarkan Tabel 5 diatas, secara keseluruhan evaluasi model *clustering* menggunakan PCA memiliki nilai *average silhouette* yang lebih tinggi daripada solusi model tanpa PCA. Artinya, model *clustering* PCA memiliki jarak yang jauh dari *cluster* tetangganya. Hal ini menunjukkan bahwa masalah *curse of dimensionality* pada *clustering* dapat diatasi dengan menggunakan PCA. Hasil analisis PCA juga memberikan performa pada model *clustering* yang lebih baik dalam semua model yang digunakan yaitu: *K-means, Agnes, Diana,* dan PAM.

Tabel 5. Evaluasi Model

Metode	**Nilai *Average illhouette* tanpa PCA**	**Nilai *Average illhouette* dengan PCA**
Ward	0,32	0,48
K-Means	0,30	0,37
Diana	0,32	0,45
PAM	0,28	0,42

Karakterisasi Model Terbaik

Dari Tabel 6 dapat dilihat bahwa jumlah provinsi yang tersebar dari masing-masing *cluster* adalah sebanyak 2, 9, 9, 12, dan 2. Setiap *cluster* terkelompok secara tersebar dari berbagai provinsi khususnya pada *cluster* satu yang tidak berdekatan. Tetapi, terdapat *cluster* yang juga berdekatan seperti *Cluster* 5 hal ini menunjukkan tingkat kemiskinan di beberapa daerah tertentu yang berdekatan memiliki karakteristik yang hampir sama. Analisis lebih lanjut mengenai nilai rata-rata variabel karakteristik kemiskinan terlampir pada Tabel 7.

Berdasarkan tabel rata-rata variabel karakteristik pada masing-masing *cluster*, ditunjukkan bahwa *Cluster* 5 adalah *cluster* dengan

tingkat kemiskinan yang bernilai sangat tinggi. Hal ini disebabkan karena nilai variabel Indeks Kedalaman Kemiskinan, Indeks Keparahan Kemiskinan dan Persentase Penduduk Miskin adalah yang paling tinggi daripada *cluster* lainnya. Di sisi lain, berdasarkan ketiga variabel tersebut, *Cluster* 3 adalah *cluster* dengan tingkat kemiskinan yang sangat rendah. Jika dilihat dari provinsi yang tercakup dalam *Cluster* 3 adalah didominasi provinsi yang berada di pulau Kalimantan dan sekitar pulau Nusa Tenggara termasuk Bali sementara *cluster* 5 terdiri atas provinsi Papua dan Papua Barat.

Tabel 6. Anggota *Cluster* Provinsi

Cluster	Anggota	Jumlah
1	Aceh, Maluku	2
2	Sumatera Utara, Sumatera Barat, Kep. Riau, DKI Jakarta, Jawa Barat, Banten, Kalimantan Tengah, Kalimantan Timur, Sulawesi Utara	9
3	Riau, Jambi, Kep. Bangka Belitung, Bali, NTT, Kalimantan Barat, Kalimantan Selatan, Kalimantan Utara, Maluku Utara	9
4	Sumatera Selatan, Bengkulu, Lampung, Jawa Tengah, DI Yogyakarta, Jawa Timur, NTB, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat	12
5	Papua Barat, Papua	2

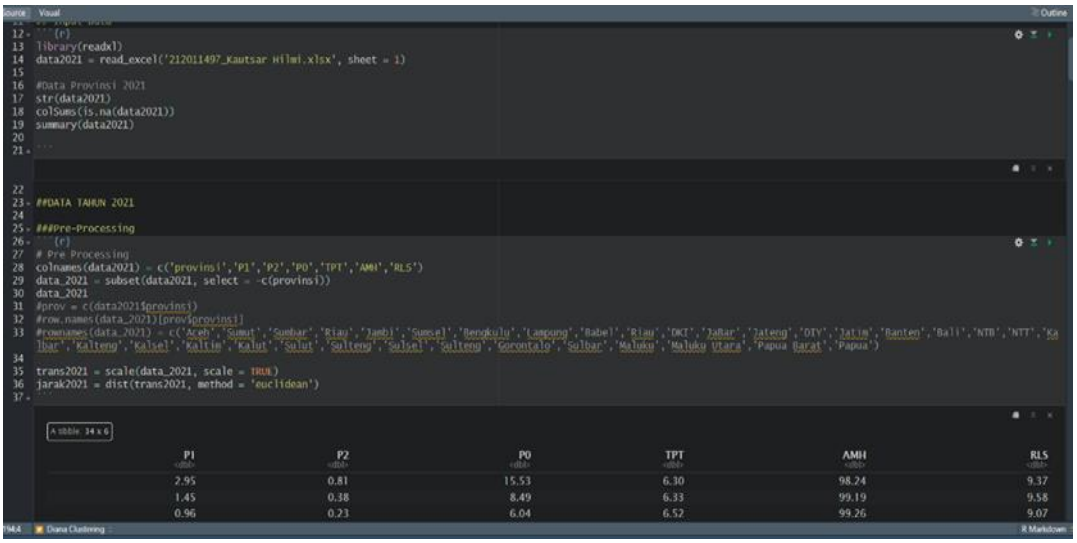
Cluster 1 mempunyai karakteristik dengan persentase Rata-rata Lama Sekolah yang paling tinggi dibandingkan dengan *cluster* lainnya. Pada *cluster* 1 juga memiliki karakteristik angka Tingkat Pengangguran Terbuka dan Angka Melek Huruf yang lebih tinggi daripada beberapa *cluster* lainnya. Hal ini menunjukkan bahwa pada provinsi yang tercakup dalam *cluster* 1 memiliki kualitas Pendidikan yang baik dengan dilihat dari AMH dan RLS. Pendidikan yang berkualitas pada provinsi tersebut dapat mendukung penurunan tingkat kemiskinan. Tetapi, persentase pengangguran pada *cluster* ini perlu ditinjau lebih lanjut karena dengan angka yang cukup tinggi daripada *cluster* lainnya, dapat memicu peningkatan tingkat kemiskinan.

Cluster 2 memiliki karakteristik dengan TPT dan AMH yang paling tinggi dibandingkan dengan *cluster* lainnya. Selain itu, nilai RLS juga tertinggi kedua daripada *cluster* lainnya. Indikasi dari karakteristik ini menunjukkan bahwa kualitas Pendidikan pada *cluster* 2 memiliki kriteria yang cukup sehingga *cluster* 2 tergolong dalam tingkat kemiskinan yang rendah. Tetapi masalah pengangguran akan menjadi tantangan dengan angka yang cukup tinggi. Diperlukan kebijakan pemerintah untuk dapat memberikan pelatihan kerja dan memperluas lapangan pekerjaan kepada masyarakat sehingga dapat meningkatkan tenaga kerja yang terampil dan mampu bersaing dengan lainnya.

Cluster 3 memiliki karakteristik dengan tingkat kemiskinan yang sangat rendah yang dibuktikan dengan rata-rata *cluster* dengan variabel P1, P2, dan P0 yang sangat rendah daripada *cluster* lainnya. Pada provinsi yang termasuk dalam *cluster* ini menunjukkan bahwa rata-rata masyarakat telah mendapatkan pendidikan yang cukup dan mampu dalam membaca serta memiliki

Tabel 7. Rata-rata Variabel Karakteristik *Cluster*

Cluster	P1	P2	P0	TPT	AMH	RLS	Keterangan
1	3,220	0,935	15,915	6,615	98,830	9,7	Tinggi
2	1,085	0,264	6,631	7,993	99,046	9,625	Rendah
3	0,828	0,178	5,874	4,835	97,876	8,764	Sangat Rendah
4	2,045	0,514	11,771	4,456	94,912	8,284	Sedang
5	5,646	1,890	23,213	4,313	90,216	7,380	Sangat Tinggi



Gambar 4. Analisis *Cluster* menggunakan R-Studio

tingkat pengangguran yang rendah. Kebijakan yang perlu dilakukan pemerintah dalam mengawal provinsi pada *cluster* ini adalah dengan mempertahankan stabilitas perekonomian dan indikator sosial kependudukan untuk meningkatkan angka pendidikan, pendapatan dan persentase tenaga kerja yang ada.

Cluster 4 memiliki karakteristik berupa angka pengangguran, AMH, dan RLS yang cukup rendah sehingga tingkat kemiskinan di *cluster* ini tergolong Sedang. Indikasi dari karakteristik ini adalah kualitas pendidikan di *cluster* ini sudah cukup baik. Walaupun memiliki tingkat pengangguran yang rendah, perlu menjadi perhatian penting untuk masyarakat yang termasuk memiliki pendapatan di bawah garis kemiskinan yang persentasenya cukup banyak dengan memperhatikan kualitas Pendidikan maupun pekerjaan yang ada pada *cluster* ini. Provinsi yang tercakup dalam *cluster* ini adalah provinsi yang memiliki angka pendapatan masyarakat yang cenderung dibawah rata-rata sehingga tingkat kemiskinan dapat dibilang cukup tinggi atau sedang.

Cluster 5 memiliki karaketeristik dengan tingkat kemiskinan yang sangat tinggi. *Cluster* ini memiliki AMH dan RLS yang nilainya paling rendah dibandingkan *cluster* lainnya. Indikasi dari fenomena ini adalah tingginya tingkat kemiskinan yang disebabkan kualitas Pendidikan yang sangat rendah. Perlu dilakukan pemfokusan kebijakan pemerintah dengan meningkatkan sarana dan prasrana Pendidikan pada provinsi yang tercakup dalam *cluster* ini. Dengan adanya peningkatan tersebut akan memperbaiki SDM yang dapat lebih berkualitas dan meningkatkan perekonomian sehingga tingkat kemiskinan akan berkurang. Provinsi Papua Barat dan Papua perlu menjadi fokus pembangunan dalam kebijakan pemerintah selanjutnya.

Analisis *cluster* dalam penelitian ini dilakukan menggunakan R-Studio yang tercantum pada Gambar 4 mengenai proses input dan *read* data dalam proses analisis data. Penelitian ini juga memperkuat pemanfaatan penggunaan metode machine learning, baik clustering maupun classification, bermanfaat dalam menganalisis data kemiskinan [41] dan pembangunan manusia [42].

.

4. KESIMPULAN

Dari keseluruhan metode *cluster* yang digunakan, masing-masing memiliki evaluasi jumlah *cluster* optimal yang berbeda. Metode *cluster* terbaik dalam karakterisasi kemiskinan di Indonesia tahun 2021 adalah *cluster ward linkage* yang merupakan *hierarchial* clustering. Penggunaan analisis PCA dalam model *clustering* memberikan hasil yang terbaik daripada *clustering* tanpa menggunakan PCA. Hal ini menunjukkan pendekatan PCA dapat mengatasi reduksi dimensi dan *noise* data sehingga meningkatkan perfoma *cluster* dalam karakterisasi kemiskinan di provinsi Indonesia. Metode *cluster ward* dengan pendekatan PCA dapat digunakan dalam memodelkan karakterisasi yang optimal dalam menentukan kebijakan kesejahteraan masyarakat yang tepat sasaran.

Berdasarkan hasil *clustering,* pada tahun 2021 Provinsi Papua masih tergolong provinsi dengan tingkat kemiskinan yang sangat tinggi. Hal ini ditunjukkan dengan kualitas Pendidikan yang rendah (AMH dan RLS rendah). Pemerintah dapat mengambil langkah prioritas pada pembangunan sarana Pendidikan di provinsi Papua untuk penurunan tingkat kemiskinan. Provinsi seperti Jakarta, Jawa Barat dan Riau merupakan *cluster*

dengan tingkat kemiskinan rendah, tetapi angka Tingkat Partisipasi Pengangguran (TPT) yang tinggi, sehingga pelaksanaan kebijakan berupa pelatihan kerja perlu dilakukan untuk menunjang produktivitas tenaga kerja di seluruh provinsi di Indonesia untuk memiliki daya saing. Sedangkan untuk penelitian selanjutnya dapat ditambahkan beberapa variabel sebagai indikator kemiskinan dan pengganti variabel Angka Melek huruf yang juga memberikan karakteristik terhadap kemiskinan di Indonesia.

DAFTAR PUSTAKA

[1] N. Nurwati, "Kemiskinan : Model Pengukuran, Permasalahan dan Alternatif Kebijakan," *J. Kependud. Padjadjaran*, vol. 10, no. 1, 2008.

[2] Nasir, M. Saichudin, and Maulizar, "Analisis Faktor-faktor yang Mempengaruhi Kemiskinan Rumah Tangga di Kabupaten Purworejo," *J. Eksek.*, vol. 5, no. 4, 2008.

[3] J. Sumanta, "Fenomena Lingkaran Kemiskinan: Analisis Ekonometrika Regional," *J. Kebijak. Ekon.*, vol. 1, 2015.

[4] World Bank, *World Development Report 2000/2001: Attacking Poverty*. 2000.

[5] C. Clayman, S. Srinivasan, and R. Sangwan, "K-means Clustering and Principal Components Analysis of Microarray Data of L1000 Landmark Genes," *Procedia Comput. Sci.*, vol. 168, pp. 97–104, Jan. 2020, doi: 10.1016/j.procs.2020.02.265.

[6] E. Xhafaj and Nurja, "The Principal Components Analysis and Cluster Analysis as Tools for the Estimation of Poverty, an Albanian Case Study," *Int. J. Sci. Res.*, pp. 2319–7064, Jan. 2013.

[7] N. Febianto and N. Palasara, "Analisa Clustering K-Means Pada Data Informasi Kemiskinan Di Jawa Barat Tahun 2018," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 8, pp. 130–140, Aug. 2019, doi: 10.32736/sisfokom.v8i2.653.

[8] N. Afira and A. W. Wijayanto, "Analisis Cluster dengan Metode Partitioning dan Hierarki pada Data Informasi Kemiskinan Provinsi di Indonesia Tahun 2019," *Komputika J. Sist. Komput.*, vol. 10, no. 2, pp. 101–109, 2021, doi: 10.34010/komputika.v10i2.4317.

[9] R. P. Michelle, R. Aurelius, and C. Robielos, "Applying Clustering Algorithm on Poverty Analysis in a Community in the Philippines," *Proc. Int. Conf. Ind. Eng. Oper. Manag. Monterrey, Mex. Novemb. 3-5*, pp. 1511–1521, 2021, [Online]. Available: http://ieomsociety.org/proceedings/2021monterrey/251.pdf

[10] E. Ji, M. Kim, and S. J. Oh, "A Study on the Welfare State Model in Korea Keimyung University , Daegu , Republic of Korea," vol. 7, no. 3, pp. 33–38, 2020.

[11] N. Thamrin and A. W. Wijayanto, "Comparison of Soft and Hard Clustering: A Case Study on Welfare Level in Cities on Java Island: Analisis cluster dengan menggunakan hard clustering dan soft clustering untuk pengelompokkan tingkat kesejahteraan kabupaten/kota di pulau Jawa," *Indones. J. Stat. Its Appl.*, vol. 5, pp. 141–160, Mar. 2021, doi: 10.29244/ijsa.v5i1p141-160.

[12] A. 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, p. 66, Sep. 2021, doi: 10.24843/JIK.2021.v14.i02.p01.

[13] W. M. Fauziyah and A. I. Achmad, "Penerapan Analisis Cluster Hybrid untuk Pengelompokan Kabupaten/Kota di Provinsi Jawa Barat Berdasarkan Indikator Kemiskinan Tahun 2022," *Bandung Conf. Ser. Stat.*, vol. 3, no. 2, pp. 566–574, 2023, doi: 10.29313/bcss.v3i2.8610.

[14] I. N. L. Fitriana and M. O. Mabruri, "Cluster Analysis Of Covid-19 Impact On Poverty In Indonesia Using Self-Organizing Map Algorithm," *J. Apl. Stat. Komputasi Stat.*, vol. 14, no. 1, pp. 85–94, 2022, doi: 10.34123/jurnalasks.v14i1.389.

[15] Soemartini and E. Supartini, "Analisis K-Means Cluster Untuk Pengelompokan Kabupaten / Kota Di Jawabarat Berdasarkan Indikator Masyarakat," *Konf. Nas. Penelit. Mat. dan Pembelajarannya II (KNPMP II)*, no. Knpmp Ii, pp. 144–154, 2017.

[16] C. C. Astuti and V. Rezania, "Cluster Analysis for Grouping Districts in Sidoarjo Regency Based on Education Indicators," *KnE Soc. Sci.*, pp. 311–317, 2021, doi: 10.18502/kss.v7i10.11233.

[17] Ş. Yilmaz and S. Sener, "Analysis of The Countries According to The Prosperity Level with Data Mining," *Alphanumeric J.*, vol. 10, no. 2, pp. 85–104, 2022, doi: 10.17093/alphanumeric.1002461.

[18] P.-N. Tan, M. Steinbach, V. Kumar, T. Pang-Ning, M. Steinbach, and V. Kumar, "Introduction to data mining: Instructur's," *Libr. Congr.*, p. 769, 2006.

[19] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis.: Pearson Prentice Hall*. 2007.

[20] Narwati, "Pengelompokan Mahasiswa Menggunakan Algoritma K-Means," *J. Din. Inform.*, vol. 2, no. 2, pp. 1–7, 2010.

[21] M. Goreti, Y. N. Nasution, and S. Wahyuningsih, "Perbandingan Hasil Analisis Cluster dengan Menggunakan Metode Single Linkage dan Metode C-Means," *EKSPONENSIAL; Vol 7 No 1*, Nov. 2017, [Online]. Available: http://jurnal.fmipa.unmul.ac.id/index.php/exponensial/article/view/15

[22] B. Everitt, *Cluster analysis*, vol. 14, no. 1. 1980. doi: 10.1007/BF00154794.

[23] R. Ramadani and A. Salma, "Metode Average Linkage Dan Ward Dalam Pengelom Pokan Kesejahteraan Sumatera Barat Tahun 2021," *J. Math. UNP*, vol. 7, no. 3, pp. 11–24, 2022.

[24] G. Abdillah, F. A. Putra, and F. Renaldi, "Penerapan Data Mining Pemakaian Air Pelanggan untuk Menentukan Klasifikasi Potensi Pemakaian Air Pelanggan Baru di PDAM Tirta Raharja Menggunakan Algoritma K-means," *Semin. Nas. Teknol. Inf. dan Komun. 2016*, pp. 18–19, 2016.

[25] S. Agarwal, *Data Mining: Data Mining Concepts and Techniques*. 2013. doi: 10.1109/ICMIRA.2013.45.

[26] R. H. B. Bangun, "Analisis Klaster Non Heirarki Dalam Pengelompokan Kabupaten/Kota di Sumatera Utara Berdasarkan Faktor Produksi Padi," *J. Agribisnis Sumatera Utara*, vol. 4, no. 1, pp. 54–61, 2016.

[27] M. Sammour and Z. ali othman, "An Agglomerative Hierarchical Clustering with Various Distance Measurements for Ground Level Ozone Clustering in Putrajaya, Malaysia," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, p. 1127, Dec. 2016, doi: 10.18517/ijaseit.6.6.1482.

[28] R. C. Balabantaray, C. Sarma, and M. Jha, "Document Clustering using K-Means and K-Medoids," 2015, [Online]. Available: http://arxiv.org/abs/1502.07938

[29] E. Atmaja, "Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta," *Int. J. Appl. Sci. Smart Technol.*, vol. 1, pp. 33–44, Jun. 2019, doi: 10.24071/ijasst.v1i1.1859.

[30] S. Defiyanti, M. Jajuli, and N. Rohmawati, "Optimalisasi K-MEDOID dalam Pengklasteran Mahasiswa Pelamar Beasiswa dengan CUBIC CLUSTERING CRITERION," *J. Teknol. dan Sist. Inf.*, vol. 3, p. 211, May 2017, doi: 10.25077/TEKNOSI.v3i1.2017.211-218.

[31] N. Qona'ah, A. R. Devi, and I. M. G. M. Dana, "Laboratory Clustering using K-Means, K-Medoids, and Model-Based Clustering," *Indones. J. Appl. Stat.*, vol. 3, no. 1, p. 64, 2020, doi: 10.13057/ijas.v3i1.40823.

[32] T. Susilowati, D. Sugiarto, and I. Mardianto, "Uji Validasi Algoritme Self-Organizing Map (SOM) dan K-Means untuk Pengelompokan Pegawai," *Tek. Inform. Fak. Teknol. Ind.*, vol. 4, no. 6, pp. 1171–1178, 2021.

[33] S. Akhanli and C. Hennig, "Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes," *Stat. Comput.*, vol. 30, Sep. 2020, doi: 10.1007/s11222-020-09958-2.

[34] T. Ullmann, C. Hennig, and A. L. Boulesteix, "Validation of cluster analysis results on validation data: A systematic framework," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 12, no. 3, pp. 1–19, 2022, doi: 10.1002/widm.1444.

[35] C. Hennig, "Cluster-wise assessment of cluster stability," *Comput. Stat. Data Anal.*, vol. 52, no. 1, pp. 258–271, 2007, doi: https://doi.org/10.1016/j.csda.2006.11.025.

[36] G. Brock, V. Pihur, S. Datta, and S. Datta, "ClValid: An R package for cluster validation," *J. Stat. Softw.*, vol. 25, no. 4, pp. 1–22, 2008, doi: 10.18637/jss.v025.i04.

[37] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set," *J. Stat. Softw.*, vol. 61, no. 6 SE-Articles, pp. 1–36, Nov. 2014, doi: 10.18637/jss.v061.i06.

[38] C. Aggarwal, A. Hinneburg, and D. Keim, "On the Surprising Behavior of Distance Metric in High-Dimensional Space," *First publ. Database theory, ICDT 200, 8th Int. Conf. London, UK, January 4 - 6, 2001 / Jan Van den Bussche ... (eds.). Berlin Springer, 2001, pp. 420-434 (=Lecture notes Comput. Sci. ; 1973)*, Feb. 2002.

[39] A. Ben-Hur and I. Guyon, "Detecting stable clusters using principal component analysis.," *Methods Mol. Biol.*, vol. 224, pp. 159–182, 2003, doi: 10.1385/1-59259-364-X:159.

[40] J. Horn, "A rationale and test for the number of factors in factor analysis," *Psychometrika*, vol. 30, no. 2, pp. 179–185, 1965, [Online]. Available:

https://econpapers.repec.org/RePEc:spr:psycho:v:30:y:1965:i:2:p:179-185

[41] Q. Iman and A. W. Wijayanto, "Klasifikasi Rumah Tangga Penerima Beras Miskin (Raskin)/Beras Sejahtera (Rastra) di Provinsi Jawa Barat Tahun 2017 dengan Metode Random Forest dan Support Vector Machine," *J. Sist. dan Teknol. Inf.*, vol. 9, no. 2, p. 178, 2021, doi: 10.26418/justin.v9i2.44137.

[42] E. Luthfi and A. W. Wijayanto, "Analisis perbandingan metode hirearchical, k-means, dan k-medoids clustering dalam pengelompokkan indeks pembangunan manusia Indonesia," *Inovasi*, vol. 17, no. 4, pp. 761–773, 2021, doi: 10.30872/jinv.v17i4.10106.