JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i3.7251

PERBANDINGAN ALGORITMA K-MEANS DAN AGGLOMERATIVE HIERARCHICAL CLUSTERING UNTUK PENGELOMPOKAN DAERAH PENGHASIL PADI DI INDONESIA

Toto Abdulpatah[1*], Betha Nurina Sari[2], Susilawati[3]

[1,2,3]Universitas Singaperbangsa Karawang; Jl. HS. Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat 41361

Keywords:
Padi;
K-Means;
Agglomerative;
Silhouette Coeficient;
Davies Bouldin Index.

Corespondent Email:
totoabdulpatah45@gmail.com

Abstrak. Padi merupakan komoditas pangan utama di Indonesia yang menjadi bahan makanan pokok mayoritas masyarakat dan berperan penting dalam perekonomian nasional. Namun, tingginya volume impor beras menunjukkan bahwa produksi padi dalam negeri belum mencukupi kebutuhan pangan nasional. Oleh karena itu, diperlukan analisis karakteristik daerah penghasil padi berdasarkan luas panen,hasil produksi, dan produktivitas. Hal ini penting untuk mengidentifikasi daerah - daerah yang memiliki kontribusi tinggi maupun rendah terhadap produksi padi nasional. Penelitian ini bertujuan untuk membandingkan kinerja dua algoritma *clustering*, yaitu *K-Means* dan *Agglomerative Hierarchical Clustering* (AHC) untuk pengelompokan daerah penghasil padi di Indonesia. Berdasarkan hasil evaluasi, metode AHC (*Average Linkage*) menunjukkan kinerja yang lebih optimal dibandingkan *K-Means*, dengan nilai *Silhouette Coefficient* sebesar 0,723 dan *Davies-Bouldin Index* sebesar 0,229. Adapun *K-Means* menghasilkan nilai *Silhouette Coefficient* sebesar 0,696 dan *Davies-Bouldin Index* sebesar 0,404.

Abstract. *Rice is a primary food commodity in Indonesia, serving as the staple food for the majority of the population and playing a significant role in the national economy. However, the high volume of rice imports indicates that domestic rice production has not yet met national food demands. Therefore, an analysis of production characteristics based on harvested area, production yield, and productivity is necessary. This is important to identify regions that contribute either highly or minimally to national rice production. This study aims to compare the performance of two clustering algorithms, namely K-Means and Agglomerative Hierarchical Clustering (AHC), in grouping rice-producing regions in Indonesia. Based on the evaluation results, the AHC (Average Linkage) method demonstrated better performance compared to K-Means, with a Silhouette Coefficient value of 0.723 and a Davies-Bouldin Index of 0.229. Meanwhile, K-Means produced a Silhouette Coefficient of 0.696 and a Davies-Bouldin Index of 0.404.*

1. PENDAHULUAN

Padi (*Oryza sativa L.*) merupakan komoditas pangan nasional terbesar di Indonesia. Padi memiliki peran penting sebagai sumber karbohidrat, yang diolah menjadi beras dan kemudian menjadi nasi, sebagai makanan pokok yang dikonsumsi oleh sebagian besar masyarakat Indonesia [1]. Selain sebagai sumber pangan, budidaya padi juga menjadi mata pencaharian utama bagi banyak petani dan berkontribusi secara signifikan terhadap perekonomian nasional. Potensi ini didukung oleh faktor geografis dan iklim tropis seperti tanah yang subur serta curah hujan yang cukup, yang menjadikan sebagian besar wilayah Indonesia cocok untuk budidaya padi [2].

Meskipun Indonesia memiliki potensi besar dalam sektor pertanian padi dengan luas lahan yang memadai, dalam praktiknya Indonesia masih harus melakukan impor beras untuk menutupi kekurangan pasokan di dalam negeri. Menurut Badan Pusat Statistik Pada tahun 2023 total impor beras Indonesia mencapai 3,062 juta ton, dan kembali meningkat pada tahun 2024 mencapai 4,519 juta ton.

Berdasarkan data tersebut, diperlukan analisis lebih lanjut terhadap karakteristik produksi di masing-masing daerah di Indonesia. Hal ini penting untuk mengidentifikasi daerah daerah yang memiliki kontribusi tinggi maupun rendah terhadap produksi padi nasional. Dengan menganalisis data produksi, luas panen, dan produktivitas di tingkat daerah, dapat dilakukan pengelompokan daerah penghasil padi berdasarkan kemiripan karakteristiknya.

Salah satu pendekatan yang relevan dalam analisis data adalah teknik *data mining*, yang berfungsi untuk mengekstraksi informasi tersembunyi dari sekumpulan data guna memperoleh pengetahuan baru [3]. Dalam konteks ini, klasterisasi (*clustering*) merupakan metode yang umum digunakan untuk mengelompokkan data berdasarkan kemiripan karakteristik, di mana objek dalam satu klaster memiliki kesamaan lebih tinggi dibandingkan dengan objek di klaster lainnya [4]. Dalam praktiknya, terdapat berbagai algoritma *clustering*. Namun, dua metode yang sering digunakan dalam analisis data, khususnya pada studi komparatif, adalah *K-Means* dan *Agglomerative Hierarchical Clustering* (AHC).

Beberapa penelitian sebelumnya telah membandingkan kinerja algoritma *K-Means* dan *Agglomerative Hierarchical Clustering* (AHC) dalam berbagai kasus. Penelitian oleh Ramadhan dan Astuti (2024) mengenai segmentasi penjualan *online* menunjukkan bahwa algoritma AHC memiliki kualitas klaster yang lebih baik dengan nilai *silhouette score* sebesar 0,6363, dibandingkan *K-Means* yang hanya mencapai 0,5087 [5]. Sebaliknya, penelitian yang dilakukan oleh Samosir dan rekannya (2024) dalam konteks pengelompokan pengangguran menunjukkan bahwa *K-Means* lebih unggul, dengan nilai *silhouette score* sebesar 56,50%, dibandingkan AHC yang hanya memperoleh 43,69% [6].

Berdasarkan latar belakang tersebut, penelitian ini akan membandingkan kinerja algoritma *K-Means* dan *Agglomerative Hierarchical Clustering* (AHC) untuk mengelompokkan daerah penghasil padi di Indonesia.

2. TINJAUAN PUSTAKA

2.1. Knowledge Discovery in Databases

KDD (*Knowledge Discovery in Databases*) adalah proses eksplorasi, analisis, dan pemrosesan data dalam jumlah besar untuk memperoleh informasi yang bernilai atau pengetahuan yang bermanfaat [7]. Tahapan dalam KDD (*Knowledge Discovery in Databases*) meliputi beberapa langkah utama, yaitu : Seleksi Data, Prapemrosesan Data, Transformasi Data, *Data Mining* dan Evaluasi Pola [8].

2.2. Standard Scaler

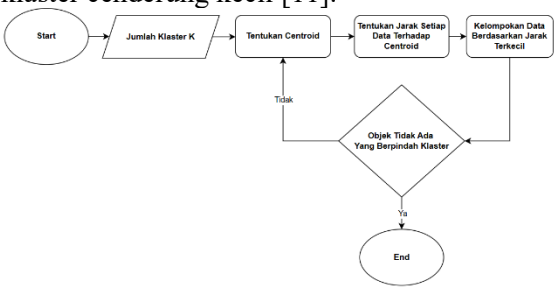
Standard Scaler merupakan salah satu teknik *preprocessing* yang digunakan untuk melakukan standarisasi pada data. Metode ini bekerja dengan mengurangi nilai rata-rata (*mean*) dari setiap fitur dan membaginya dengan standar deviasinya, sehingga setiap fitur memiliki distribusi dengan rata-rata nol dan variansi satu. Proses ini penting untuk memastikan bahwa tidak ada fitur yang mendominasi proses analisis atau pemodelan karena perbedaan skala. Standarisasi ini juga membantu meningkatkan kinerja algoritma pembelajaran mesin yang sensitif terhadap skala data, seperti *K-Means* dan algoritma berbasis jarak lainnya [9].

2.3. Reduksi Dimensi

Tingginya dimensi dan kompleksitas data dapat memberikan dampak negatif terhadap akurasi proses klasifikasi. Oleh karena itu, diperlukan tahapan reduksi dimensi dan kompleksitas data guna meminimalkan potensi kesalahan dalam proses klasifikasi. Salah satu metode yang dapat digunakan untuk mereduksi dimensi data adalah algoritma *Principal Component Analysis* (PCA). PCA bekerja dengan membentuk sejumlah dimensi baru yang disusun berdasarkan besar varian yang dikandung oleh masing-masing dimensi. Proses ini menghasilkan *principal components* yang diperoleh melalui dekomposisi nilai *eigen* (*eigenvalue*) dan vektor *eigen* (*eigenvector*) dari matriks kovarians data [10].

2.4. K-Means

Algoritma *K-Means* adalah metode pengelompokan data *non-hierarki* yang membagi data ke dalam satu atau lebih klaster. Objek dengan karakteristik yang serupa akan dikelompokkan ke dalam klaster yang sama, sedangkan objek dengan karakteristik yang berbeda akan ditempatkan dalam klaster lain. Dengan demikian, variasi data dalam satu klaster cenderung kecil [11].

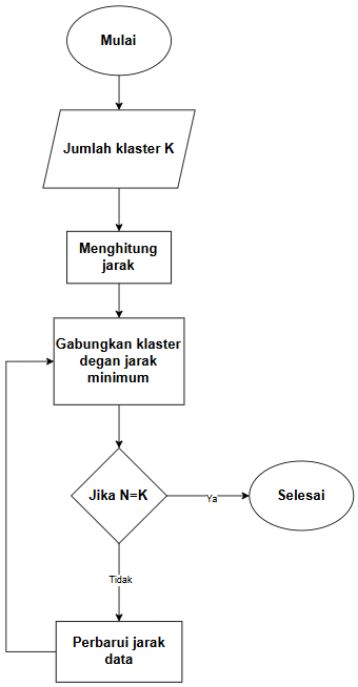


Gambar 2. 1 Diagram alur K-Means

Gambar 2. 1 menyajikan diagram alur algoritma *K-Means* yang menggambarkan tahapan-tahapan proses pengelompokan data, mulai dari inisialisasi centroid hingga pembentukan klaster akhir.

2.5. Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering (AHC) adalah teknik dalam analisis eksplorasi data yang digunakan untuk mengelompokkan data berdasarkan kesamaan karakteristiknya, sehingga membentuk kelompok atau klaster. Metode ini menyusun data dalam bentuk hierarki, yang kemudian direpresentasikan dalam struktur pohon, memungkinkan analisis hubungan antar kelompok dengan lebih jelas [12].



Gambar 2. 2 Diagram alur Agglomerative

Gambar 2. 2 menyajikan diagram alur dari algoritma *Agglomerative Hierarchical Clustering* yang menggambarkan proses pengelompokan data secara *bottom-up*, dimulai dari setiap data sebagai klaster tunggal hingga terbentuknya hierarki klaster berdasarkan kemiripan antar objek.

Agglomerative Hierarchical Clustering terdiri dari empat metode utama, yaitu :

2.5.1. Single Linkage

Single Linkage adalah teknik dalam analisis klaster hierarki yang mengelompokkan objek berdasarkan pasangan dengan jarak terdekat terlebih dahulu.

2.5.2. Complete Linkage

Complete Linkage adalah teknik dalam analisis klaster hierarki yang mengelompokkan objek berdasarkan jarak terjauh atau kemiripan terkecil antar objek.

2.5.3. Average Linkage

Average Linkage adalah teknik dalam analisis klaster hierarki yang mengelompokkan objek berdasarkan rata-rata jarak antara semua pasangan objek dalam klaster.

2.5.4. Ward Linkage

Ward Linkage adalah teknik analisis klaster hierarki yang didasarkan pada kehilangan informasi akibat penggabungan objek ke dalam suatu klaster.

2.6. Metode Elbow

Metode *elbow* merupakan pendekatan yang umum digunakan dalam menentukan jumlah klaster yang optimal pada proses *clustering*. Tujuan utama dari metode ini adalah untuk memperoleh jumlah klaster yang paling sesuai guna menghasilkan pengelompokan data yang efektif. Proses kerja metode *elbow* melibatkan perhitungan tingkat kohesi dan pemisahan antar klaster. Kohesi mengacu pada tingkat kedekatan data dalam suatu klaster, sedangkan pemisahan menunjukkan tingkat perbedaan antar klaster. Kedua indikator tersebut dapat diukur menggunakan nilai *Sum of Squares Error* (SSE), yang mencerminkan total kesalahan kuadrat dari data terhadap pusat klasternya [13].

2.7. Koefisien korelasi Chopenetic

Koefisien korelasi *cophenetic* merupakan ukuran statistik yang menilai tingkat kesesuaian antara matriks jarak asli dengan matriks jarak *cophenetic* yang diperoleh dari dendrogram. Ukuran ini dimanfaatkan untuk memilih metode klasterisasi hierarki yang paling sesuai, seperti *single linkage*, *average linkage*, *complete linkage* dan *ward linkage*. Nilai koefisien ini berada dalam rentang -1 hingga 1. Semakin tinggi nilai koefisien korelasi *cophenetic*, semakin baik *dendrogram* tersebut mencerminkan hubungan asli antar objek, yang berarti hasil pengelompokan dinilai lebih akurat [14].

2.8. Dendrogram

Dendrogram merupakan diagram berbentuk pohon yang menggambarkan hubungan hierarkis antar objek dalam suatu data. Diagram ini umumnya digunakan sebagai hasil visual dari proses klasterisasi hierarkis, dengan tujuan utama untuk membantu menentukan cara terbaik dalam mengelompokkan objek berdasarkan kemiripannya [15].

2.9. Silhouette Coeficient

Silhouette Coefficient digunakan untuk menilai tingkat keseragaman anggota dalam suatu klaster serta perbedaan antara klaster yang berbeda. Nilai koefisien ini berkisar antara -1 hingga 1, di mana semakin mendekati 1 menunjukkan bahwa pembentukan klaster semakin optimal, sedangkan jika mendekati -1, maka kualitas klaster yang terbentuk semakin buruk [16].

Tabel 2. 1 Interpretasi nilai *silhouette coeficient*

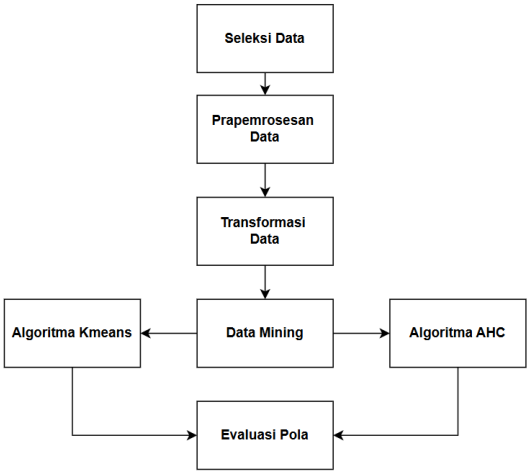
Rentang Nilai	Interpretasi
0,71 – 1,00	Struktur kuat
0,51 – 0,70	Struktur baik
0,26 – 0,50	Struktur lemah
≤ 0,25	Tidak terstruktur

2.10. Davies Bouldin Index

Davies Bouldin Index digunakan untuk mengukur tingkat kemiripan antar klaster dengan membandingkan jarak antara klaster yang berbeda terhadap ukuran klaster itu sendiri. *Davies-Bouldin Index* digunakan untuk menilai perbandingan antara jarak di dalam klaster dan jarak antar klaster, di mana semakin rendah nilainya, semakin baik kualitas pengelompokan yang dihasilkan [17].

3. METODE PENELITIAN

Metode penelitian menjelaskan alur penelitian secara umum yang dilakukan peneliti agar mampu menghasilkan penelitian yang sistematis guna mendapatkan hasil akhir berupa pengelompokan daerah penghasil padi di Indonesia. Adapun tahapan pada penelitian ini dapat digambarkan pada **Gambar 3. 1**.



Gambar 3. 1 Alur Metodologi Penelitian

3.1. Seleksi Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang diambil dari situs Badan Pusat Statistik (BPS). Data tersebut meliputi 38 provinsi dan terdiri atas empat atribut. dimana peneliti hanya mengambil 2 dataset yaitu dataset padi 2023 dan 2024.

Atribut yang digunakan dalam penelitian ini dapat ditunjukkan pada **Tabel 3. 1**.

Tabel 3. 1 Deskripsi variabel penelitian

Atribut	Satuan	Keterangan
Provinsi	Non satuan	Wilayah administratif
Luas Panen	Ha	Hektar
Produksi	Ton GKG	Ton Gabah Kering Giling
Produktivitas	Ku/Ha	Kuintal per hektar

3.2. Prapemrosesan Data

Tahap kedua adalah prapemrosesan data. Pada tahap ini dilakukan pembersihan data (*data cleaning*) untuk menghapus informasi yang tidak relevan, kesalahan penginputan, data duplikat dan menangani nilai kosong (*missing value*) yang dapat mempengaruhi hasil analisis.

3.3. Transformasi Data

Tahap ketiga adalah Transformasi Data. Pada tahap ini, data yang telah melalui proses prapemrosesan akan diubah ke dalam format yang sesuai untuk digunakan pada tahap selanjutnya, yaitu proses *clustering* dalam data mining. Transformasi data bertujuan untuk menyelaraskan struktur data, baik dari segi format maupun skala, agar algoritma dapat bekerja secara optimal. Dalam penelitian ini, transformasi dilakukan melalui proses normalisasi menggunakan metode *Standard Scaler* untuk menyamakan skala antar atribut. Selain itu, dilakukan pula reduksi dimensi menggunakan algoritma *Principal Component Analysis* (PCA) guna menyederhanakan kompleksitas data serta meningkatkan efisiensi dan visualisasi pada tahap *clustering*.

3.4. Data Mining

Pada tahap ini, algoritma atau metode analisis diterapkan untuk menemukan pola-pola penting dari data yang telah diproses sebelumnya. Penelitian ini menggunakan dua algoritma, yaitu *K-Means* dan *Agglomerative Hierarchical Clustering* (AHC), untuk melakukan pengelompokan terhadap daerah penghasil padi di Indonesia.

3.5. Evaluasi Pola

Tahap terakhir dalam penelitian ini adalah evaluasi, yang bertujuan untuk menilai kualitas hasil *clustering* terhadap data yang telah melalui seluruh tahapan sebelumnya. Evaluasi dilakukan untuk memastikan bahwa pengelompokan yang dihasilkan benar-benar mencerminkan struktur alami data. Pada penelitian ini, digunakan dua metrik evaluasi, yaitu *Silhouette Coefficient* dan *Davies-Bouldin Index. Silhouette Coefficient* mengukur seberapa mirip suatu objek dengan klasternya sendiri dibandingkan dengan klaster lain, sedangkan *Davies-Bouldin Index* mengevaluasi rasio antara jarak dalam klaster dan antar klaster, di mana nilai yang lebih rendah menunjukkan hasil *clustering* yang lebih baik.

4. HASIL DAN PEMBAHASAN

Hasil dari penelitian ini berupa perbandingan implementasi dua algoritma *clustering*, yaitu *K-Means* dan *Agglomerative Hierarchical Clustering* (AHC) untuk mengelompokkan daerah penghasil padi di Indonesia. Perbandingan kedua algoritma tersebut dievaluasi menggunakan dua metrik pengukuran kualitas klaster, yaitu *Silhouette Coefficient* dan *Davies-Bouldin Index*. Hasil klasterisasi kemudian divisualisasikan pada peta.

4.1. Data Seleksi

Tahap pertama dalam penelitian ini adalah melakukan seleksi data pada dataset padi di Indonesia. Dataset yang digunakan diperoleh dari situs resmi Badan Pusat Statistik (BPS), yang menyediakan informasi mengenai Luas Panen, Produksi, dan Produktivitas Padi menurut provinsi di Indonesia. Adapun data yang digunakan adalah data padi tahun 2023 dan 2024 pada **Gambar 4. 1** dan **Gambar 4. 2**.

Provinsi	Luas Panen 2023	Produksi 2023	Produktivitas 2023
Aceh	254287.38	1404234.82	55.22
Sumatera Utara	406109.49	2087474.15	51.40
Sumatera Barat	300564.77	1482468.79	49.32
Papua Tengah	2094.20	9273.22	44.28
Papua Pegunungan	14.14	61.63	43.59

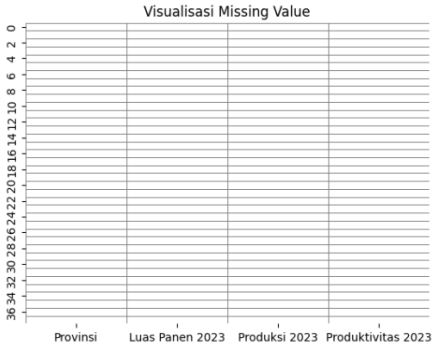
Gambar 4. 1 Dataset Padi Indonesia 2023

Provinsi	Luas Panen 2024	Produksi 2024	Produktivitas 2024
Aceh	301196.35	1659966.28	55.11
Sumatera Utara	419463.48	2204875.51	52.56
Sumatera Barat	295278.98	1356467.93	45.94
Papua Tengah	1436.12	6072.38	42.28
Papua Pegunungan	9.66	42.38	43.87

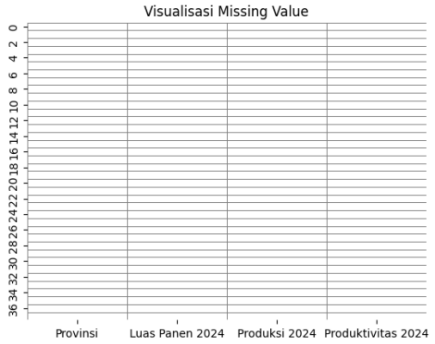
Gambar 4. 2 Dataset Padi Indonesia 2024

4.2. Prapemrosesan Data

Tahap kedua pada penelitian ini adalah proses pembersihan dan pemeriksaan data untuk meningkatkan kualitas data. Pemeriksaan pertama yang dilakukan adalah memastikan bahwa data tidak mengandung nilai kosong (*missing value*).



Gambar 4. 3 *Visualisasi Pengecekan Missing value Dataset Padi 2023*



Gambar 4. 4 Visualisasi Pengecekan Missing Value Dataset Padi 2024

Pada **Gambar 4. 3** dan **Gambar 4. 4** dilakukan pengecekan terhadap kedua dataset menggunakan visualisasi *heatmap* untuk mengidentifikasi adanya nilai kosong atau anomali dalam data. *Heatmap* ini menampilkan representasi visual dari nilai-nilai dalam *dataframe*, di mana kotak berwarna putih menandakan data yang terisi, sedangkan kotak berwarna hitam biasanya menunjukkan nilai kosong. Berdasarkan hasil visualisasi, seluruh kotak pada *heatmap* tampak berwarna putih dan tidak ditemukan kotak berwarna hitam. Hal ini mengindikasikan bahwa tidak terdapat nilai kosong dalam dataset, sehingga data dinyatakan bersih dan siap digunakan untuk tahap analisis selanjutnya.

Pemeriksaan kedua dilakukan terhadap keberadaan data duplikat. Pada **Gambar 4. 5** Tidak terdapat baris duplikat dalam *dataframe*, sehingga data dinyatakan bersih dari redundansi.

```
# Cek jumlah baris yang duplikat
jumlah_duplikat = dataset.duplicated().sum()
print(f"Jumlah baris duplikat: {jumlah_duplikat}")
```

```
Jumlah baris duplikat: 0
```

Gambar 4. 5 Pengecekan Data Duplikat

Setelah seluruh data tersaring, kedua dataset yang terpisah kemudian digabungkan menjadi satu kesatuan dataset yang utuh dan siap digunakan dalam proses klasterisasi.

Provinsi	Luas Panen 2023	Produksi 2023	Produktivitas 2023	Luas Panen 2024	Produksi 2024	Produktivitas 2024
Aceh	254287.38	1404234.82	55.22	301196.35	1659966.28	55.11
Sumatera Utara	406109.49	2087474.15	51.40	419463.48	2204875.51	52.56
Sumatera Barat	300564.77	1482468.79	49.32	295278.98	1356467.93	45.94
Papua Tengah	2094.20	9273.22	44.28	1436.12	6072.38	42.28
Papua Pegunungan	14.14	61.63	43.59	9.66	42.38	43.87

Gambar 4. 6 Dataset Gabungan

Pada **Gambar 4. 6** akan dilakukan penggantian nama – nama atribut agar lebih sederhana dan minimalis. Berikut disajikan pada **Gambar 4. 7**.

Provinsi	LP23	PI23	PS23	LP24	PI24	PS24
Aceh	254287.38	1404234.82	55.22	301196.35	1659966.28	55.11
Sumatera Utara	406109.49	2087474.15	51.40	419463.48	2204875.51	52.56
Sumatera Barat	300564.77	1482468.79	49.32	295278.98	1356467.93	45.94
Papua Tengah	2094.20	9273.22	44.28	1436.12	6072.38	42.28
Papua Pegunungan	14.14	61.63	43.59	9.66	42.38	43.87

Gambar 4. 7 Penggantian Nama Atribut

4.3. Transformasi Data

Tahap ketiga dalam penelitian ini adalah transformasi data, yaitu proses mengubah format, struktur, atau nilai data agar sesuai dengan kebutuhan analisis lebih lanjut. Transformasi ini sangat penting untuk memastikan bahwa data berada dalam kondisi yang optimal sebelum masuk ke tahap *data mining*.

Hasil dari proses normalisasi ditunjukkan pada **Gambar 4. 8**, yang memperlihatkan dataset setelah dilakukan transformasi menggunakan *standard scaler* terhadap atribut-atribut numerik.

Provinsi	LP23	PI23	PS23	LP24	PI24	PS24
Aceh	-0.032520	-0.006480	1.180257	0.087108	0.107964	1.184815
Sumatera Utara	0.308111	0.264866	0.695765	0.366868	0.332960	0.879375
Sumatera Barat	0.071309	0.024590	0.431958	0.073110	-0.017352	0.086431
Papua Tengah	-0.598344	-0.560485	-0.207268	-0.621973	-0.574938	-0.351965
Papua Pegunungan	-0.603011	-0.564143	-0.294781	-0.625347	-0.577428	-0.161514

Gambar 4. 8 Normalisasi Dataset

Setelah proses normalisasi dilakukan terhadap dataset, tahap selanjutnya adalah reduksi dimensi menggunakan metode

Principal Component Analysis (PCA). Reduksi dimensi ini bertujuan untuk menyederhanakan data dengan mengubah enam atribut awal menjadi tiga komponen utama (*principal components*). Berikut disajikan hasil dari Reduksi Dimensi pada **Gambar 4. 9**.

Provinsi	PC1	PC2	PC3
Aceh	0.866304	1.432813	0.036130
Sumatera Utara	1.091822	0.671486	0.140421
Sumatera Barat	0.238790	0.291288	-0.236689
Papua Tengah	-1.224786	0.217151	-0.086099
Papua Pegunungan	-1.194263	0.279297	0.111607

Gambar 4. 9 Reduksi Dimensi Data yang sudah di normalisasi

4.4. Data Mining

Tahap ke empat dalam penelitian ini adalah *data mining*, yaitu proses *clustering* data untuk mengelompokkan data ke dalam beberapa klaster yang memiliki kesamaan, sehingga masing-masing kelompok merepresentasikan karakteristik yang serupa di antara anggotanya. Pada penelitian ini akan ada dua penerapan algoritma yaitu algoritma *K-Means* dan algoritma *Agglomerative Hierarchical Clustering*.

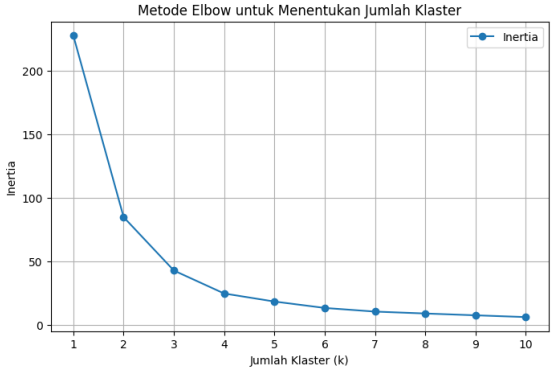
4.4.1. Algoritma K-Means

Pada algoritma *K-Means*, langkah awal yang dilakukan adalah menentukan jumlah klaster (k) yang optimal. Salah satu metode yang digunakan untuk tujuan ini adalah Metode *Elbow*, yang dilakukan dengan menghitung nilai *inertia* atau *Sum of Squared Errors* (SSE) untuk berbagai jumlah klaster (k), biasanya dalam rentang tertentu. Berikut ini adalah nilai *inertia* untuk masing-masing jumlah klaster (k) yang dapat dilihat pada **Tabel 4. 1**.

Tabel 4. 1 Nilai *Inertia* tiap cluster

Jumlah Klaster	SSE
2	84.80
3	42.71
4	24.63
5	18.30
6	13.25
7	10.38
8	8.88
9	7.45
10	6.06

Selanjutnya, nilai - nilai *inertia* tersebut divisualisasikan dalam bentuk grafik untuk melihat titik siku (*elbow point*), yaitu titik di mana penurunan nilai inertia mulai melambat secara signifikan. Titik ini dianggap sebagai jumlah klaster yang paling optimal karena setelah titik tersebut, penambahan jumlah klaster tidak memberikan penurunan *inertia* yang signifikan.



Gambar 4. 10 Grafik Metode Elbow

Pada **Gambar 4. 10** menunjukan bahwa sumbu horizontal (X) ditampilkan jumlah klaster (k) yang diuji, sedangkan sumbu vertikal (Y) menunjukkan nilai *inertia* yang diperoleh untuk masing-masing nilai k. Terlihat bahwa nilai inertia menurun tajam dari k = 1 hingga sekitar k = 2, dan setelah itu penurunannya mulai melambat. Titik tekukan atau *elbow* pada grafik ini terletak pada k = 2, yang mengindikasikan bahwa jumlah klaster yang optimal adalah 2.

Maka setelah didapatkan jumlah klaster optimalnya, selanjutnya yaitu proses *modelling clustering k-means*. Berikut program *k-means* pada **Gambar 4. 11.**

```
from sklearn.cluster import KMeans

kmean = KMeans(n_clusters=2, random_state=42)
labels_kmean = kmean.fit_predict(x_pca)

labels_kmean
```

```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int32)
```

Gambar 4. 11 Program *Clustering K-Means*

4.4.1. Algoritma Agglomerative

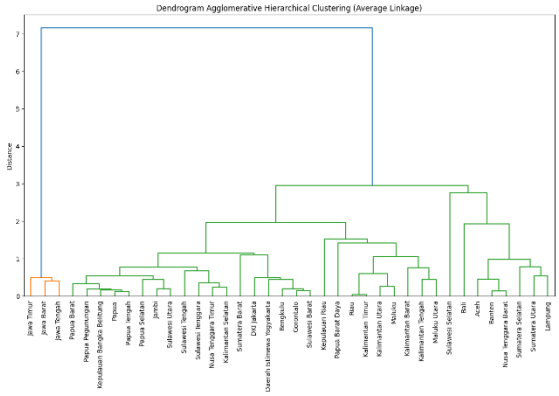
Langkah awal yang dilakukan adalah menentukan metode optimal dari algoritma agglomerative untuk digunakan dalam pembentukan klaster yang efektif. Ada empat metode yang digunakan yaitu : *single linkage*, *complete linkage*, *average linkage* dan *ward linkage*.

Untuk menentukan metode terbaik di antara keempatnya, dilakukan pengujian menggunakan koefisien korelasi *cophenetic*. Koefisien ini digunakan untuk mengukur sejauh mana dendrogram yang dihasilkan mampu merepresentasikan jarak asli antar objek dalam data. Nilai koefisien korelasi *cophenetic* berada dalam rentang -1 hingga 1, dengan nilai yang semakin mendekati 1 menunjukkan bahwa struktur klaster yang terbentuk semakin sesuai dengan struktur data aslinya. Oleh karena itu, metode *linkage* yang menghasilkan nilai koefisien korelasi *cophenetic* tertinggi dipilih sebagai metode paling optimal untuk proses *clustering* selanjutnya.

Tabel 4. 2 Koefisien korelasi *cophenetic*

Metode	Koefisien korelasi *cophenetic*
Single Linkage	0,918
Complete Linkage	0,818
Average Linkage	0,922
Ward Linkage	0,893

Pada **Tabel 4. 2** ditunjukkan bahwa metode yang paling optimal untuk algoritma *Agglomerative* dalam penelitian ini adalah metode *Average Linkage*, dengan nilai koefisien korelasi *cophenetic* sebesar 0,922. Nilai ini menunjukkan bahwa struktur dendrogram yang dihasilkan oleh metode tersebut mampu merepresentasikan jarak asli antar objek dengan sangat baik.



Gambar 4. 12 Dendrogram *Average Linkage*

Dendrogram metode *Average Linkage* pada **Gambar 4. 12** menunjukkan bahwa data terbagi menjadi dua kelompok utama, yaitu kelompok yang ditandai dengan garis kuning dan kelompok dengan garis hijau, yang keduanya dihubungkan oleh garis biru. Hal ini mengindikasikan bahwa jumlah klaster optimal untuk algoritma *Agglomerative Hierarchical Clustering* (AHC) dengan metode *Average Linkage* adalah sebanyak dua klaster.

Setelah menentukan jumlah klaster yang optimal dengan parameter. Langkah selanjutnya adalah melakukan proses *clustering* menggunakan algoritma *Agglomerative* dengan metode *Average Linkage* pada **Gambar 4. 13**.

```
from sklearn.cluster import AgglomerativeClustering
from scipy.cluster.hierarchy import linkage

agglo_average = AgglomerativeClustering(n_clusters=2, linkage='average')
labels_agglo_average = agglo_average.fit_predict(x_pca)

labels_agglo_average
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

Gambar 4. 13 Program *Clustering* AHC *Average Linkage*

4.5. Evaluasi Pola

Tahap kelima dalam penelitian ini adalah evaluasi pola, yang bertujuan menilai kualitas hasil *clustering* secara objektif. Evaluasi dilakukan untuk memastikan bahwa pengelompokan yang dihasilkan merepresentasikan struktur data dengan baik. Penilaian dilakukan menggunakan dua metrik, yaitu *Silhouette Coefficient dan Davies-Bouldin Index* (DBI). Berikut disajikan hasil evaluasi *clustering* pada **Tabel 4. 3**.

Tabel 4. 3 Evaluasi Metrik Clustering

Algoritma	*Silhouette Coeficient*	*Davies Bouldin Index*
K-Means	0,696	0,404
Agglomerative Average Linkage	0,723	0,229

Tabel 4. 3 menunjukkan bahwa algoritma yang paling optimal adalah algoritma *Agglomerative* dengan metode *Average Linkage*. Hal ini didasarkan pada hasil evaluasi menggunakan dua metrik, yaitu *Silhouette Coefficient* dan *Davies-Bouldin Index* (DBI), di mana keduanya menunjukkan nilai yang lebih baik dibandingkan algoritma *K-Means*.

Algoritma *K-Means* menghasilkan nilai *Silhouette Coefficient* sebesar 0,696 yang termasuk dalam kategori struktur klaster yang baik. Sementara itu, algoritma *Agglomerative Average Linkage* menghasilkan nilai *Silhouette*

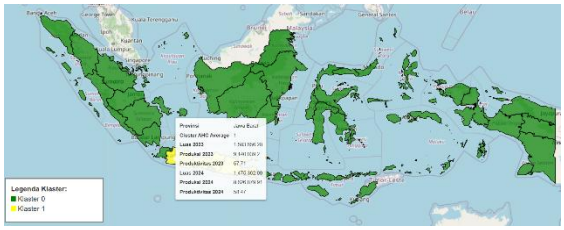
Coefficient sebesar 0,723, yang termasuk dalam kategori struktur klaster yang kuat.

Untuk metrik *Davies-Bouldin Index*, *K-Means* memperoleh nilai sebesar 0,404, sedangkan *Agglomerative Average Linkage* memperoleh nilai yang lebih rendah, yaitu sebesar 0,229. Karena pada *Davies-Bouldin Index* semakin rendah nilainya maka semakin baik kualitas pengelompokan yang dihasilkan, maka dapat disimpulkan bahwa *Agglomerative Average Linkage* memberikan hasil pengelompokan yang lebih optimal dibandingkan *K-Means*.

Selanjutnya, untuk memperkuat interpretasi hasil klasterisasi, dilakukan visualisasi dalam bentuk peta sebaran klaster. Visualisasi ini bertujuan untuk memberikan gambaran spasial mengenai distribusi provinsi dalam setiap klaster, sehingga memudahkan dalam memahami pola-pola geografis yang terbentuk dari hasil *clustering*. Berikut disajikan hasil visualisasi peta klasterisasi dengan algoritma *agglomerative average linkage* pada **Gambar 4. 14**, **Gambar 4. 15** dan **Gambar 4. 16**.

Gambar 4. 14 Peta hasil *clustering* AHC *average linkage*

Gambar 4. 15 Provinsi Kalimantan Timur masuk ke klaster 0



Gambar 4. 16 Provinsi Jawa Barat masuk ke klaster 1

Berdasarkan visualisasi tersebut, dapat dilihat bahwa hasil pengelompokan (*clustering*) terhadap 38 provinsi di Indonesia dilakukan berdasarkan karakteristik data komoditas padi yang dianalisis, mencakup tiga variabel utama yaitu luas panen, hasil produksi, dan produktivitas pada tahun 2023 dan 2024. Proses klasterisasi dilakukan menggunakan algoritma *Agglomerative Hierarchical Clustering* dengan metode *average linkage*, yang menghasilkan dua kelompok klaster.

Klaster 1 terdiri atas tiga provinsi, yaitu Jawa Barat, Jawa Tengah, dan Jawa Timur, yang menunjukkan karakteristik yang relatif serupa dan berbeda secara signifikan dibandingkan provinsi lainnya. Sementara itu, sebanyak 35 provinsi lainnya tergolong ke dalam Klaster 0, yang tersebar di berbagai wilayah di luar ketiga provinsi tersebut.

5. KESIMPULAN

a. Hasil pengelompokan menggunakan algoritma *K-Means* menunjukan bahwa data berhasil dibagi kedalam dua klaster. Klaster 0 terdiri dari 4 provinsi dan klaster 1 terdiri dari 34 provinsi.
b. Hasil pengelompokan menggunakan algoritma *agglomerative* dengan metode *average linkage* menunjukan bahwa data berhasil dibagi kedalam dua klaster. Klaster 0 terdiri dari 35 provinsi dan klaster 1 terdiri dari 3 provinsi.
c. Hasil perbandingan algoritma *clustering* antara *K-Means* dan *Agglomerative* berdasarkan evaluasi menggunakan metrik *Silhouette Coefficient* dan *Davies Bouldin Index* menunjukkan bahwa algoritma *Agglomerative* dengan metode *Average Linkage* memberikan hasil yang lebih baik, dengan nilai *Silhouette Coefficient* sebesar 0,723 dan *Davies Bouldin Index* sebesar 0,229. Nilai tersebut lebih unggul dibandingkan *K-Means* yang menghasilkan *Silhouette Coefficient* sebesar 0,696 dan *Davies-Bouldin Index* sebesar 0,404.

UCAPAN TERIMA KASIH

Penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada Ibu Betha Nurina Sari, M.Kom. dan Ibu Susilawati, M.Si., selaku Dosen Prodi Informatika Universitas Singaperbangsa Karawang, atas bimbingan, arahan, serta dukungan yang telah diberikan

selama proses penyusunan penelitian ini. Kontribusi dan masukan yang diberikan sangat berarti dalam menyempurnakan hasil penelitian yang tertuang dalam jurnal ini.

DAFTAR PUSTAKA

[1] S. Wijayanto and M. Y. Fathoni, "Pengelompokkan produktivitas tanaman padi di Jawa Tengah menggunakan metode clustering K-Means," *JUPITER: Jurnal Penelitian Ilmu dan Teknologi Komputer*, vol. 13, no. 2, pp. 212–219, 2021.

[2] H. Azari, D. Hartanti, and A. A. Sari, "Pengelompokan produksi padi dan beras Provinsi Jawa Timur dengan metode Agglomerative Hierarchical Clustering," *Infotek: Jurnal Informatika dan Teknologi*, vol. 7, no. 2, pp. 379–389, 2024.

[3] A. Aziz, *Penerapan Algoritma K-Means dan Fuzzy C-Means untuk Pengelompokan Kabupaten Kota Berdasarkan Produksi Padi di Provinsi Jawa Barat*, Doctoral dissertation, Universitas Buana Perjuangan Karawang, 2021.

[4] H. Oktavianto, B. S. Bakti, and A. F. Cobantoro, "Perbandingan kinerja Agglomerative Clustering pada data stunting untuk segmentasi wilayah," *BINA INSANI ICT Journal*, vol. 10, no. 2, pp. 145–154, 2023.

[5] G. Ramadhan and Y. Astuti, "Perbandingan kinerja algoritma K-means dan Agglomerative Clustering untuk segmentasi penjualan online pada customer retail," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 9, no. 1, pp. 92–96, 2024.

[6] E. F. Situmorang, Y. A. Sihombing, and E. L. D. Samosir, "Perbandingan K-Means dengan Hierarchical Clustering untuk pengelompokkan tingkat pengangguran di Sumatera Utara," *TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi*, vol. 4, no. 2 (SEMNASTIK), pp. 155–159, 2024.

[7] A. Supriyadi, A. Triayudi, and I. D. Sholihati, "Perbandingan algoritma K-Means dengan K-Medoids pada pengelompokan armada kendaraan truk berdasarkan produktivitas," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 6, no. 2, pp. 229–240, 2021.

[8] K. K. Munawar and A. I. Purnamasari, "Implementasi algoritma K-Means Clustering pada klasterisasi kasus HIV di Jawa Barat," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 2, pp. 1092–1099, 2023.

[9] G. Sumantri, M. D. Novianto, and P. P. Prihastuti, "Implementasi Fuzzy C-Means dalam Pengelompokan Provinsi di Indonesia untuk Pemerataan Kualitas Pendidikan," in *Prosiding Seminar Pendidikan Matematika dan Matematika*, vol. 8, Mar. 2023.

[10] D. Hediyati and I. M. Suartana, "Penerapan Principal Component Analysis (PCA) untuk reduksi dimensi pada proses clustering data produksi pertanian di Kabupaten Bojonegoro," *JIEET (Journal of Information Engineering and Educational Technology)*, vol. 5, no. 2, pp. 49–54, 2021.

[11] C. Nisa and W. Yustanti, "Studi perbandingan algoritma klastering dalam pengelompokan persediaan produk (studi kasus: Subdirektorat Perencanaan Sarana Prasarana dan Logistik PTN X)," *Journal of Emerging Information System and Business Intelligence (JEISBI)*, vol. 2, no. 3, pp. 14–20, 2021.

[12] G. G. Ghiffary, K. Alifviansyah, A. Fitrianto, E. Erfiani, and L. R. D. Jumansyah, "Perbandingan algoritma HDBSCAN dan Agglomerative Hierarchical Clustering dalam klasterisasi pada data yang mengandung pencilan," *Jurnal Riset dan Aplikasi Matematika (JRAM)*, vol. 8, no. 2, pp. 122–135, 2024.

[13] P. Vania and B. N. Sari, "Perbandingan Metode Elbow dan Silhouette untuk Penentuan Jumlah Klaster yang Optimal pada Clustering Produksi Padi menggunakan Algoritma K-Means," *Jurnal Ilmiah Wahana Pendidikan*, vol. 9, no. 21, pp. 547–558, 2023.

[14] S. D. Raihannabil, H. M. A. Ilyas, H. N. Shafira, M. A. Riani, N. N. Hastin, and T. K. H. Siregar, "Perbandingan Agglomerative Nesting dan K-Means untuk klasterisasi ketimpangan gender berdasarkan dimensi kesehatan reproduksi," in *Seminar Nasional Official Statistics*, vol. 2024, no. 1, pp. 459–470, Nov. 2024.

[15] A. M. Sroyer, S. A. Mandowen, and F. Reba, "Analisis cluster penyakit malaria Provinsi Papua menggunakan metode Single Linkage dan K-Means," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 7, no. 3, pp. 147–154, 2022.

[16] A. Septianingsih, "Pemetaan kabupaten kota di Provinsi Jawa Timur berdasarkan tingkat kasus penyakit menggunakan pendekatan Agglomeratif Hierarchical Clustering," *Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 3, no. 2, pp. 367–386, 2022.

[17] A. M. Sikana and A. W. Wijayanto, "Analisis perbandingan pengelompokan Indeks Pembangunan Manusia Indonesia tahun 2019 dengan metode Partitioning dan Hierarchical Clustering," *Jurnal Ilmu Komputer*, vol. 14, no. 2, pp. 66–78, 2021.