

KOMPARASI K-MEANS DAN AGGLOMERATIVE HIERARCHICAL CLUSTERING DALAM KLASTERISASI WILAYAH PRODUKSI PADI KARAWANG UNTUK DASAR PEMBERIAN SUBSIDI

Taufiqurrohman Yuarez^{1*}, Aries Suharso², Jajam Haerul Jaman³

^{1,2} Universitas Singaperbangsa Karawang; Jl. HS. Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat 41361 ³

Keywords:

Padi;
K-Means;
Agglomerative Hierarchical Clustering;
wilayah karawang.

Correspondent Email:

taufiqyuarez@gmail.com

Abstrak Kabupaten Karawang merupakan lumbung padi nasional di Jawa Barat, namun luas panen dan jumlah produksi antar kecamatan menunjukkan karakteristik yang beragam dan tidak merata. Kondisi ini mendorong perlunya pengelompokan wilayah agar distribusi bantuan pertanian dapat dilakukan secara lebih tepat sasaran. Penelitian ini bertujuan membandingkan kinerja algoritma K-Means dan Agglomerative Hierarchical Clustering (AHC) dalam klasterisasi wilayah produksi padi di Kabupaten Karawang berdasarkan variabel luas panen dan produksi periode 2023–2025. Data yang digunakan merupakan data sekunder Dinas Pertanian sebanyak 90 baris (30 kecamatan selama tiga tahun) yang dianalisis menggunakan tahapan Knowledge Discovery in Databases (KDD). Pada K-Means diperoleh jumlah klaster optimal $k=3$ dengan nilai SSE 28,83, sedangkan AHC menggunakan $k=2$ berdasarkan analisis dendrogram dengan metode Average Linkage dan nilai Cophenetic Correlation 0,7897. Evaluasi menggunakan Davies-Bouldin Index (DBI) dan Silhouette Score menunjukkan AHC lebih unggul dibanding K-Means. Hasil AHC mengelompokkan 79 data ke Klaster Rendah dan 11 data ke Klaster Tinggi. Hasil penelitian ini diharapkan menjadi acuan dalam penyaluran bantuan pertanian yang lebih tepat sasaran.



Copyright © 2026 The Author(s), Published by Jurnal Informatika dan Teknik Elektro Terapan. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Abstract. Karawang Regency is one of the national rice barns in West Java, yet harvest area and rice production across sub-districts show diverse and uneven characteristics. This study compares the performance of the K-Means and Agglomerative Hierarchical Clustering (AHC) algorithms in clustering rice production areas in Karawang Regency based on harvest area and total production variables for 2023–2025. The data consisted of 90 records from 30 sub-districts obtained from the Karawang Regency Agriculture Office. The analysis followed the Knowledge Discovery in Databases (KDD) methodology. For K-Means, the optimal number of clusters was $k = 3$ using the Elbow method with an SSE value of 28.83. For AHC, $k = 2$ was determined using dendrogram analysis with the Average Linkage method, producing a Cophenetic Correlation value of 0.7897. Evaluation using the Davies-Bouldin Index (DBI) and Silhouette Score showed that AHC outperformed K-Means. Based on AHC results, 79 data points were classified into the Low Cluster and 11 into the High Cluster. These findings are expected to support more targeted agricultural assistance policies..

1. PENDAHULUAN

Kabupaten Karawang merupakan lumbung padi nasional di Jawa Barat dengan rata-rata produksi mencapai $\pm 1.411.330$ ton per tahun, menjadikannya daerah penghasil padi terbesar kedua di Jawa Barat setelah Kabupaten Indramayu [1]. Namun, berdasarkan data BPS Jawa Barat periode 2023–2025, produksi padi Karawang berfluktuasi, yaitu 1.131.976,57 ton (2023), turun menjadi 1.041.530,95 ton (2024), dan naik kembali menjadi 1.194.712,00 ton (2025). Fluktuasi ini mencerminkan karakteristik luas panen dan produksi yang tidak merata antarwilayah kecamatan, akibat perbedaan kondisi pertanian, keterbatasan akses input, rendahnya adopsi teknologi, serta lemahnya akses informasi pasar [2]. Kondisi ini menyebabkan kebijakan distribusi subsidi yang seragam berpotensi tidak tepat sasaran, sehingga pengelompokan wilayah berdasarkan karakteristik produksi padi menjadi langkah penting agar alokasi bantuan dapat diarahkan secara lebih proporsional [3] [4].

Salah satu pendekatan yang dapat digunakan untuk mengelompokkan wilayah secara objektif adalah *clustering* pada *data mining*, yaitu metode pengelompokan data berdasarkan kemiripan antarobjek [5]. Penelitian terkait telah banyak menggunakan K-Means maupun Agglomerative Hierarchical Clustering (AHC). K-Means efisien dan mudah diinterpretasikan, namun sensitif terhadap nilai K awal dan *outlier* [6]. AHC menghasilkan *dendrogram* yang informatif tanpa perlu menentukan jumlah kluster di awal, tetapi proses penggabungannya bersifat *irreversible* dan sangat bergantung pada pemilihan jenis *linkage* [7]. Studi perbandingan menunjukkan bahwa keunggulan masing-masing algoritma sangat bergantung pada karakteristik data yang digunakan [8].

Meskipun penelitian klasterisasi data pertanian dengan K-Means dan AHC telah dilakukan di berbagai wilayah, belum terdapat kajian yang secara khusus membandingkan kedua algoritma tersebut pada data produksi padi tingkat kecamatan di Kabupaten Karawang. Belum diketahui algoritma mana yang paling optimal untuk mengklasifikasikan pola produksi padi dengan karakteristik geografis dan pertanian yang spesifik di wilayah tersebut, sehingga diperlukan studi komparatif yang dapat memberikan rekomendasi berbasis data secara ilmiah.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan membandingkan kinerja algoritma K-Means dan AHC dalam klasterisasi wilayah produksi padi di Kabupaten Karawang berdasarkan variabel luas panen dan jumlah produksi. Secara teoritis, penelitian ini diharapkan memperkaya kajian perbandingan metode *clustering* pada data produksi pertanian tingkat kecamatan. Secara praktis, hasilnya diharapkan menjadi acuan bagi pemerintah daerah dalam menyusun kebijakan distribusi subsidi dan program penyuluhan yang lebih tepat sasaran.

2. TINJAUAN PUSTAKA

2.1. Data Mining

Data Mining merupakan proses ekstraksi informasi yang bermanfaat dan pengetahuan baru dari sekumpulan data besar untuk mendukung pengambilan keputusan [9]. Dalam *data mining*, terdapat enam metode utama yang umum diimplementasikan, meliputi prediksi, deskripsi, klasifikasi, estimasi, asosiasi, dan *clustering* [10]. Pada penelitian ini, metode yang digunakan adalah *clustering* atau klasterisasi, yaitu proses pengelompokan data dimana data yang memiliki karakteristik serupa dimasukkan ke dalam satu kelompok kluster, sedangkan data yang berbeda ditempatkan pada kluster lain.

2.2. K-Means

K-Means *Clustering* merupakan salah satu metode pengelompokan data yang paling sering digunakan dalam analisis data. Algoritma ini bekerja dengan cara membagi data ke dalam jumlah K kelompok yang telah ditentukan sebelumnya, di mana K menunjukkan jumlah kluster yang diinginkan. Proses diawali dengan pemilihan K titik pusat (*centroid*) secara acak, kemudian setiap data dimasukkan ke dalam kluster dengan *centroid* terdekat. Posisi *centroid* diperbarui berdasarkan nilai rata-rata seluruh data dalam masing-masing kluster, dan tahapan ini diulang hingga tidak terjadi perubahan pada pengelompokan data [11]. Jarak antar data dengan *centroid* dihitung menggunakan rumus Euclidean Distance sebagaimana persamaan (1).

$$d(a_i, b_j) = \sqrt{\sum (a_i - b_j)^2} \quad (1)$$

Keterangan:

$d(a_i, b_j)$ = jarak antara data a_i dan centroid b_j

a_i = nilai data ke- i

b_j = nilai centroid klaster ke- j

2.3. Agglomerative Hierarchical Clustering (AHC)

Agglomerative Hierarchical Clustering (AHC) merupakan metode dalam *data mining* yang membentuk klaster secara bertahap dengan menggabungkan objek atau klaster yang memiliki tingkat kemiripan tertinggi hingga seluruh data tergabung dalam satu klaster besar, menggunakan *Euclidean Distance* sebagai ukuran jarak [12]. Metode AHC memiliki beberapa teknik penggabungan klaster, yaitu: (1) *Single Linkage*, berdasarkan jarak minimum antaranggota klaster; (2) *Complete Linkage*, berdasarkan jarak maksimum; (3) *Average Linkage*, berdasarkan rata-rata jarak seluruh pasangan data; dan (4) *Ward*, meminimalkan nilai *Error Sum of Squares* (ESS) untuk meningkatkan homogenitas klaster [13]. Hasil pengelompokan divisualisasikan dalam bentuk *dendrogram* yang menggambarkan hierarki penggabungan data [14].

$$D(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2)$$

Keterangan:

$D(x, y)$ = jarak Euclidean antara data x dan y

n = jumlah sampel data

x_i = nilai data ke- i pada variabel x

y_i = nilai data ke- i pada variabel y

2.4. Metode Elbow

Metode *Elbow* merupakan teknik evaluasi yang digunakan untuk menentukan jumlah klaster optimal dalam proses *clustering* dengan menganalisis nilai *Sum of Squared Errors* (SSE) pada berbagai jumlah klaster. Nilai SSE divisualisasikan dalam bentuk grafik, di mana titik “siku” (*elbow*) yang menandai penurunan nilai SSE mulai melambat digunakan sebagai acuan jumlah klaster optimal [15]. Pemilihan rentang nilai k perlu mempertimbangkan jumlah objek data, karena nilai k yang terlalu kecil menyebabkan *under-clustering* sedangkan terlalu besar menyebabkan *over-clustering* [16].

2.5. Cophenetic Correlation

Cophenetic Correlation digunakan untuk menentukan metode *linkage* terbaik pada *hierarchical clustering* dengan mengevaluasi kesesuaian antara jarak asli antardata dan jarak yang terbentuk pada *dendrogram*. Koefisien ini memiliki rentang nilai antara -1 hingga 1, di mana semakin tinggi nilainya, semakin baik metode *linkage* dalam merepresentasikan struktur data. Oleh karena itu, metode *linkage* dengan nilai *Cophenetic Correlation* tertinggi dipilih sebagai metode terbaik [17].

2.6. Standard Scaler

Standard Scaler merupakan metode *preprocessing* yang digunakan untuk standarisasi data dengan cara mengurangi nilai rata-rata (*mean*) pada setiap fitur dan membaginya dengan nilai standar deviasi, sehingga setiap fitur memiliki rata-rata nol dan variansi satu. Standarisasi ini bertujuan agar tidak ada fitur yang mendominasi akibat perbedaan skala data, sekaligus meningkatkan performa algoritma yang sensitif terhadap skala seperti K-Means [8].

2.7. Davies-Bouldin Index

Davies-Bouldin Index (DBI) merupakan metode evaluasi hasil *clustering* yang menilai tingkat kohesi dan separasi antara klaster. Kohesi mengacu pada kedekatan data terhadap pusat klasternya, sedangkan separasi menggambarkan jarak antarklaster. Semakin kecil nilai DBI, semakin baik kualitas pengelompokan yang dihasilkan [18].

2.8. Silhouette Coefficient

Silhouette Coefficient merupakan metode evaluasi *clustering* yang menggabungkan konsep kohesi (kedekatan data dalam satu klaster) dan separasi (jarak antarklaster). Nilai *Silhouette* dihitung pada setiap data untuk mengevaluasi kualitas masing-masing klaster maupun keseluruhan hasil *clustering* (Yolandari et al., 2025). Interpretasi nilainya adalah: 0,71-1,00 (Struktur Kuat), 0,51-0,70 (Struktur Baik), 0,26-0,50 (Struktur Lemah), dan $\leq 0,25$ (Tidak Terstruktur) [19].

Tabel 1. Interpretasi Nilai Silhouette Coefficient

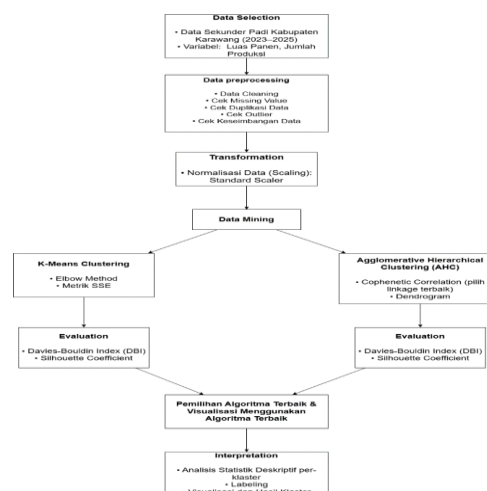
Rentang Nilai	Interpretasi
0,71 – 1,00	Struktur Kuat
0,51 – 0,70	Struktur Baik
0,26 – 0,50	Struktur Lemah
$\leq 0,25$	Tidak Terstruktur

2.9. Knowledge Discovery in Databases (KDD)

KDD (*Knowledge Discovery in Databases*) merupakan proses yang terstruktur dan sistematis untuk menemukan informasi bernilai dan pola-pola menarik dari kumpulan data berukuran besar [20]. Tahapan KDD yang digunakan dalam penelitian ini meliputi: (1) *Selection*, pemilihan data yang relevan; (2) *Preprocessing*, pembersihan data meliputi penanganan *missing value*, duplikasi, dan *outlier*; (3) *Transformation*, normalisasi data menggunakan *Standard Scaler*; (4) *Data Mining*, penerapan algoritma K-Means dan AHC; dan (5) *Evaluation*, evaluasi kualitas kluster menggunakan DBI dan *Silhouette Coefficient* [21].

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan *Knowledge Discovery in Databases* (KDD) yang meliputi tahapan seleksi data, prapemrosesan data, transformasi data, *data mining*, dan evaluasi pola. Gambaran alur penelitian secara umum dapat dilihat pada Gambar 1.



3.1. Data selection

Pada tahap data selection, digunakan data sekunder pertanian padi tingkat kecamatan di Kabupaten Karawang periode 2023–2025 yang diperoleh dari Dinas Pertanian Kabupaten Karawang dalam format Excel Rekapitulasi Kegiatan Serasi Padi (RKSP). Dari delapan atribut yang tersedia, dipilih dua variabel sebagai fitur clustering, yaitu Luas_Panen dan Jumlah_Produksi, karena keduanya paling merepresentasikan hasil pertanian padi setiap kecamatan. Atribut lainnya tidak digunakan karena hanya berfungsi sebagai identitas, memiliki korelasi tinggi dengan variabel terpilih, bersifat turunan, atau memiliki variasi data yang rendah. Hasil seleksi menghasilkan dataset sebanyak 90 data yang mencakup 30 kecamatan selama tiga tahun pengamatan.

3.2. Data Preprocessing

Tahap kedua adalah preprocessing data, yaitu proses pembersihan dan pemeriksaan dataset untuk memastikan kualitas data sebelum analisis. Tahapan yang dilakukan meliputi data cleaning, pemeriksaan *missing value*, pengecekan duplikasi data, identifikasi outlier menggunakan metode *Interquartile Range (IQR)*, serta pemeriksaan keseimbangan data untuk memastikan setiap kecamatan memiliki jumlah observasi yang konsisten selama periode pengamatan.

3.3. Data Transformation

Pada tahap transformation, data yang telah melalui preprocessing dinormalisasi menggunakan metode *Standard Scaler* sehingga setiap variabel memiliki rata-rata (mean) 0 dan standar deviasi 1. Proses ini bertujuan

menyamakan skala antarvariabel agar tidak terjadi dominasi variabel dengan nilai yang lebih besar dalam perhitungan jarak, sehingga data siap digunakan pada tahap clustering.

3.4. Data Mining

Pada tahap data mining, dilakukan proses clustering wilayah pertanian padi di Kabupaten Karawang berdasarkan variabel Luas_Panen dan Jumlah_Produksi menggunakan algoritma K-Means dan *Agglomerative Hierarchical Clustering* (AHC). Jumlah kluster optimal pada K-Means ditentukan menggunakan metode Elbow berdasarkan nilai SSE, sedangkan pada AHC dilakukan pemilihan metode linkage terbaik berdasarkan nilai *Cophenetic Correlation* dan hasil pengelompokan divisualisasikan melalui dendrogram. Selanjutnya, kinerja kedua algoritma dibandingkan untuk menentukan metode yang menghasilkan kluster paling optimal.

3.5. Evaluation

Pada tahap evaluation, hasil clustering dari algoritma K-Means dan *Agglomerative Hierarchical Clustering* (AHC) dievaluasi menggunakan metrik Davies-Bouldin Index (DBI) dan Silhouette Coefficient. Nilai DBI yang lebih rendah dan Silhouette Coefficient yang lebih tinggi menunjukkan kualitas kluster yang lebih baik. Berdasarkan hasil evaluasi tersebut, dipilih algoritma terbaik untuk menginterpretasikan pola persebaran wilayah produksi padi di Kabupaten Karawang.

3.6. Interpretation

Pada tahap interpretation, hasil klasterisasi dari algoritma terbaik dianalisis menggunakan statistik deskriptif untuk mengidentifikasi karakteristik setiap kluster dan memberikan label kategori wilayah. Selanjutnya, hasil klasterisasi divisualisasikan dalam bentuk peta distribusi wilayah produksi padi Kabupaten Karawang pada tahun 2023–2025 untuk menggambarkan pola persebaran spasial selama periode penelitian..

4. HASIL DAN PEMBAHASAN

Hasil penelitian ini berupa perbandingan algoritma K-Means dan *Agglomerative Hierarchical Clustering* (AHC) dalam mengelompokkan wilayah produksi padi di Kabupaten Karawang berdasarkan luas panen

dan jumlah produksi periode 2023–2025. Kinerja kedua algoritma dievaluasi menggunakan Silhouette Coefficient dan Davies-Bouldin Index (DBI) untuk menentukan metode yang paling optimal. Hasil klasterisasi dari algoritma terbaik kemudian dianalisis secara statistik deskriptif dan divisualisasikan dalam bentuk peta tematik (choropleth map) berbasis Sistem Informasi Geografis (SIG) untuk menggambarkan persebaran spasial wilayah produksi padi di Kabupaten Karawang.

4.1. Data selection

Data yang digunakan merupakan data sekunder pertanian padi Kabupaten Karawang periode 2023–2025 yang diperoleh dari Dinas Pertanian Kabupaten Karawang dalam bentuk file Excel Rekapitulasi Kegiatan Serasi Padi (RKSP). Data dari ketiga tahun tersebut direkap dan digabungkan menjadi satu dataset yang terdiri atas 90 data dari 30 kecamatan dengan delapan atribut, yaitu No, Kecamatan, Luas_Tanam, Luas_Puso, Luas_Panen, Jumlah_Produksi, Provitas, dan Tahun.

No	Kecamatan	Luas_Tanam	Luas_Puso	Luas_Panen	Jumlah_Produksi	Provitas	Tahun
0 1	PANGKALAN	4241.0	0.0	4432.0	27156.50	61.27	2023
1 2	TEGALWARU	4207.0	0.0	4174.0	21219.10	50.84	2023
2 3	CIAMPEL	1356.0	0.0	1558.0	8981.45	57.65	2023
3 4	TELUKJAMBE TIMUR	1398.0	117.0	1888.0	7770.49	41.16	2023
4 5	TELUKJAMBE BARAT	4249.0	48.0	6063.0	36567.58	60.31	2023

Gambar 4. 1 Tampilan awal data

Selanjutnya dilakukan seleksi atribut untuk menentukan variabel yang digunakan dalam proses clustering. Dari delapan atribut yang tersedia, dipilih dua variabel utama, yaitu Luas_Panen dan Jumlah_Produksi, karena keduanya paling merepresentasikan karakteristik produksi padi pada setiap kecamatan. Atribut lainnya tidak digunakan karena hanya berfungsi sebagai identitas, memiliki korelasi tinggi dengan variabel terpilih, bersifat turunan, atau memiliki variasi data yang rendah. Hasil seleksi menghasilkan dataset dengan 90 data dan lima atribut, yaitu Tahun, Kecamatan, Luas_Panen, dan Jumlah_Produksi, yang selanjutnya digunakan pada tahap analisis berikutnya.

	Tahun	Kecamatan	Luas_Panen	Jumlah_Produksi
0	2023	PANGKALAN	4432.00	27156.50
1	2023	TEGALWARU	4174.00	21219.10
2	2023	CIAMPEL	1558.00	8981.45
3	2023	TELUKJAMBE TIMUR	1888.00	7770.49
4	2023	TELUKJAMBE BARAT	6063.00	36567.58
...	...	...	...	...
85	2025	CILEBAR	7021.00	48175.90
86	2025	CIBUAYA	9810.00	66406.05
87	2025	TIRTAJAYA	13460.00	87615.80

Gambar 4. 2 Data setelah seleksi

4.2. Data Preprocessing

Tahap preprocessing dilakukan untuk memastikan kualitas dataset sebelum proses clustering. Hasil pemeriksaan menunjukkan bahwa tidak terdapat missing value pada seluruh variabel, sehingga data dapat langsung digunakan pada tahap berikutnya. Hasil pemeriksaan disajikan pada Gambar 4.3.

Tahun	0
Kecamatan	0
Luas_Panen	0
Jumlah_Produksi	0
dtype:	int64

Gambar 4. 3 Cek missing value

Hasil pemeriksaan menunjukkan bahwa dataset tidak mengandung data duplikat, sehingga seluruh 90 data dapat digunakan pada tahap berikutnya Gambar 4.4.

```
df_cl.duplicated().sum()
np.int64(0)
```

Gambar 4. 4 cek diplikasi data

Deteksi outlier menggunakan metode Interquartile Range (IQR) menunjukkan adanya beberapa data pencilan pada variabel Luas_Panen dan Jumlah_Produksi, yang berasal dari kecamatan dengan kapasitas produksi relatif tinggi. Hasil deteksi outlier disajikan pada Gambar 4.5.

```
Outlier di Luas_Panen:
20    17353.0
49    17959.0
Name: Luas_Panen, dtype: float64
-----
Outlier di Jumlah_Produksi:
49    131317.6
Name: Jumlah_Produksi, dtype: float64
-----
```

Gambar 4. 5 Cek Outlier

Pemeriksaan keseimbangan data dilakukan untuk memastikan setiap kecamatan memiliki jumlah observasi yang konsisten selama periode penelitian. Hasil pemeriksaan menunjukkan bahwa seluruh 30 kecamatan masing-masing memiliki tiga data pengamatan pada periode 2023–2025, sehingga dataset bersifat seimbang dan siap digunakan pada tahap berikutnya. Hasil pemeriksaan ditunjukkan pada Gambar 4.6.

```
kecamatan_counts = df_cl['Kecamatan'].value_counts()
print("Kecamatan Counts:\n", kecamatan_counts)
```

Kecamatan Counts:

Kecamatan	3
PANGKALAN	3
TEGALWARU	3
CIAMPEL	3
TELUKJAMBE TIMUR	3
TELUKJAMBE BARAT	3
KLARI	3
CIKARPEK	3
PURWASARI	3
TIRTAMULYA	3
JATISARI	3
BAHYUSARI	3
KOTABANU	3
CILAMAYA WETAN	3
CILAMAYA KULON	3
LEMAHABANG	3
TELAGASARI	3
HADALAYA	3
KARAWANG TIMUR	3
KARAWANG BARAT	3
RAMAHERTA	3
TEMPURAN	3
KUTAWALUYA	3
RENGASDENGKLOK	3
JAYAKERTA	3
PEDES	3
CILEBAR	3
CIBUAYA	3
TIRTAJAYA	3
BATUJAYA	3
PAKISJAYA	3

Name: count, dtype: int64

Gambar 4. 6 Cek duplikasi Data

4.3. Data Transformation

Tahap selanjutnya adalah transformasi data melalui standarisasi menggunakan metode StandardScaler untuk menyamakan skala antarvariabel sebelum proses clustering. Proses ini dilakukan karena variabel Luas_Panen dan Jumlah_Produksi memiliki rentang nilai yang berbeda, sehingga standarisasi diperlukan agar tidak terjadi dominasi variabel tertentu dalam perhitungan jarak. Implementasi proses standarisasi ditunjukkan pada Gambar 4.7.

```
df_norm = df_eda.copy()

# Standarisasi data
features = ['Luas_Panen', 'Jumlah_Produksi']

X = StandardScaler()
X = X.fit_transform(df_norm[features])
```

Gambar 4. 7 Proses Standarisasi

Hasil standarisasi menunjukkan bahwa kedua variabel telah berada pada skala yang seragam dengan nilai yang terdistribusi di sekitar nol, sehingga siap digunakan pada tahap data mining. Sementara itu, atribut Tahun dan Kecamatan tetap dipertahankan sebagai

identitas data dan tidak ikut distandarisasi. Hasil standarisasi ditunjukkan pada Gambar 4.10.

	Tahun	Kecamatan	Luas_Panen	Jumlah_Produksi
0	2023	PANGKALAN	-0.654788	-0.701038
1	2023	TEGALWARU	-0.717950	-0.928888
2	2023	CIAMPEL	-1.358379	-1.398512
3	2023	TELUKJAMBE TIMUR	-1.277591	-1.444983
4	2023	TELUKJAMBE BARAT	-0.255499	-0.339885
...	...	...	...	...
85	2025	CILEBAR	-0.020969	0.105589
86	2025	CIBUAYA	0.661813	0.805177
87	2025	TIRTAJAYA	1.555378	1.619108
88	2025	BATUJAYA	0.545566	0.729988
89	2025	PAKISJAYA	-0.109346	-0.004007

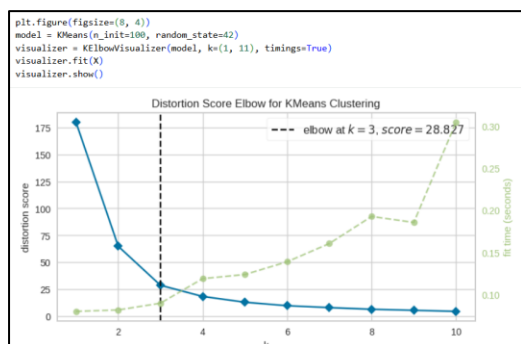
90 rows x 4 columns

Gambar 4. 8 Hasil Standarisasi Standard scaler

4.4. Data Mining

4.4.1. Algoritma K-Means

Tahap awal implementasi K-Means adalah menentukan jumlah kluster optimal menggunakan metode Elbow berdasarkan nilai Sum of Squared Errors (SSE). Grafik Elbow pada Gambar 4.11 menunjukkan bahwa penurunan nilai SSE mulai melandai pada $k = 3$, yang mengindikasikan bahwa jumlah kluster optimal adalah tiga. Nilai SSE untuk setiap jumlah kluster disajikan pada Tabel 4.3.



Gambar 4. 9 Grafik Elbow Method

Tabel 4. 1 Nilai Inertia Per-Kluster

Jumlah Kluster (K)	Nilai SSE
1	180,00
2	64,96
3	28,83
4	18,47
5	13,11
6	9,88
7	8,12
8	6,60

Jumlah Kluster (K)	Nilai SSE
9	5,63
10	4,66

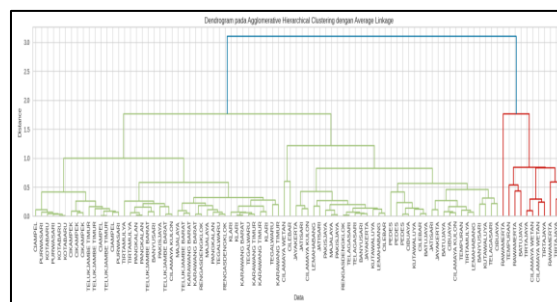
Pemilihan $k = 3$ tidak hanya didasarkan pada hasil metode Elbow, tetapi juga mempertimbangkan kemudahan interpretasi hasil. Dengan tiga kluster, wilayah produksi padi di Kabupaten Karawang dapat dikelompokkan menjadi kategori produksi tinggi, sedang, dan rendah sehingga lebih mudah digunakan sebagai dasar pengambilan keputusan. Oleh karena itu, nilai $k = 3$ digunakan dalam proses clustering menggunakan algoritma K-Means.

4.4.2. Algoritma Agglomerative Hierarchical clustering

Implementasi Agglomerative Hierarchical Clustering (AHC) diawali dengan pemilihan metode linkage terbaik berdasarkan nilai Cophenetic Correlation. Empat metode linkage yang dievaluasi adalah single, complete, average, dan ward. Hasil perbandingan pada Tabel 4.2 menunjukkan bahwa Average Linkage memiliki nilai Cophenetic Correlation tertinggi sebesar 0,789683, sehingga dipilih sebagai metode linkage yang paling mampu merepresentasikan struktur hubungan antar data.

Tabel 4. 2 Hasil Seleksi Linkage

Linkage	Cophenetic Correlation
Average	0.789683
Complete	0.778897
Single	0.722322
Ward	0.636958



Gambar 4. 10 Visualisasi Dendrogram Average

Selanjutnya, jumlah klaster ditentukan melalui analisis dendrogram pada Gambar 4.10. Berdasarkan struktur hierarki yang terbentuk, pemotongan dendrogram pada threshold distance sekitar 3,0 menghasilkan dua klaster. Oleh karena itu, model AHC dibangun menggunakan konfigurasi Average Linkage dengan jumlah klaster sebanyak dua.

4.5. Evaluation

hasil evaluasi kedua algoritma disajikan pada Tabel 4.3.

Tabel 4. 3 Perbandingan Hasil Evaluasi

Metode Evaluasi	Nilai K-Means	Nilai AHC
Davies-Bouldin Index	0,5403	0,4870
Silhouette Score	0,5673	0,5677

Berdasarkan nilai Davies-Bouldin Index (DBI) dan Silhouette Score, Agglomerative Hierarchical Clustering (AHC) menunjukkan performa yang lebih baik dibanding K-Means. AHC memperoleh nilai DBI yang lebih rendah sebesar 0,4870 dan nilai Silhouette Score yang sedikit lebih tinggi sebesar 0,5677, yang mengindikasikan bahwa klaster yang terbentuk lebih padat, kohesif, dan memiliki pemisahan antar kelompok yang lebih baik. Berdasarkan interpretasi Silhouette Score, hasil tersebut termasuk dalam kategori struktur klaster yang baik. Oleh karena itu, AHC ditetapkan sebagai model terbaik dan digunakan pada tahap interpretasi selanjutnya.

4.6. Interpretation

Tahap interpretasi dilakukan terhadap hasil klasterisasi AHC yang telah terpilih sebagai model terbaik. Prosesnya diawali dengan memisahkan data berdasarkan tahun pengamatan (2023, 2024, dan 2025), kemudian dilanjutkan dengan analisis statistik deskriptif per klaster, pemberian label, visualisasi klaster, dan pemetaan distribusi spasial wilayah produksi padi di Kabupaten Karawang.

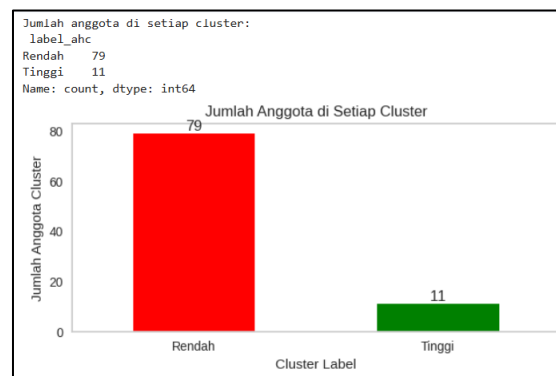
Rincian statistik deskriptif untuk setiap klaster disajikan pada Gambar 4.11.

Label_ahc	Luas_Panen			Jumlah_Produksi		
	min	Max	Mean	Min	Max	Mean
0	12309.00	17959.0	14820.272727	78442.92	131317.60	95502.867273
1	693.59	14003.0	6032.604937	4687.32	71902.45	38451.481519

Gambar 4. 11 Statistik Deskriptif per variabel

Berdasarkan gambar tersebut, Klaster 0 memiliki nilai rata-rata luas panen dan jumlah produksi yang lebih tinggi dibanding Klaster 1. Selain itu, rentang nilai pada Klaster 0 relatif lebih konsisten, yang menunjukkan karakteristik wilayah dengan produksi padi tinggi. Sebaliknya, Klaster 1 memiliki rentang nilai yang lebih bervariasi dengan rata-rata luas panen dan jumlah produksi yang lebih rendah, sehingga merepresentasikan wilayah dengan tingkat produksi yang lebih rendah.

Distribusi jumlah anggota pada setiap klaster ditunjukkan pada Gambar 4.17. Hasil tersebut menunjukkan bahwa Klaster Rendah mendominasi dengan 79 anggota, sedangkan Klaster Tinggi terdiri atas 11 anggota. Kondisi ini mengindikasikan bahwa sebagian besar wilayah di Kabupaten Karawang memiliki luas panen dan jumlah produksi yang relatif lebih rendah dibandingkan wilayah lainnya.



Selanjutnya, pengelompokan data per tahun disajikan pada tabel berikut.

Tabel 4. 4 Hasil Klaster 2023

Label	Nama Kecamatan
Rendah	Pangkalan, Tegalwaru, Ciampel, Telukjambe Timur, Telukjambe Barat, Klari, Cikampek, Purwasari,

	Tirtamulya, Jatisari, Banyusari, Kotabaru, Cilamaya Wetan, Cilamaya Kulon, Lemahabang, Telagasari, Majalaya, Karawang Timur, Karawang Barat, Kutawaluya, Rengasdengklok, Jayakarta, Pedes, Cilebar, Cibuaya, Pakisjaya
Tinggi	Rawamerta, Tempuran, Tirtajaya, Batujaya

Tabel 4. 5 Hasil Klaster 2024

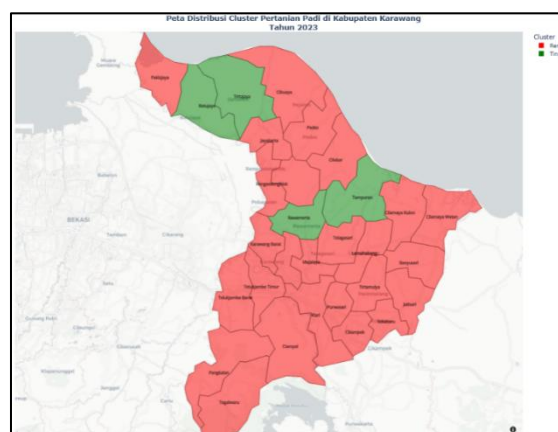
Label	Nama Kecamatan
Rendah	Pangkalan, Tegalwaru, Ciampel, Telukjambe Timur, Telukjambe Barat, Klari, Cikampek, Purwasari, Tirtamulya, Jatisari, Banyusari, Kotabaru, Cilamaya Kulon, Lemahabang, Telagasari, Majalaya, Karawang Timur, Karawang Barat, Tempuran, Kutawaluya, Rengasdengklok, Jayakarta, Pedes, Cilebar, Cibuaya, Batujaya, Pakisjaya
Tinggi	Cilamaya Wetan, Rawamerta, Tirtajaya

Tabel 4. 6 Hasil Klaster 2025

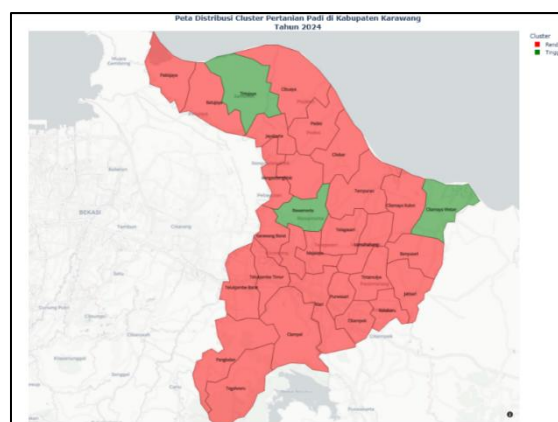
Label	Nama Kecamatan
Rendah	Pangkalan, Tegalwaru, Ciampel, Telukjambe Timur, Telukjambe Barat, Klari, Cikampek, Purwasari, Tirtamulya, Jatisari, Banyusari, Kotabaru, Cilamaya Kulon, Lemahabang, Telagasari, Majalaya, Karawang Timur, Karawang Barat, Kutawaluya, Rengasdengklok, Jayakarta, Pedes, Cilebar, Cibuaya, Batujaya, Pakisjaya

Tinggi	Cilamaya Wetan, Rawamerta, Tempuran, Tirtajaya
--------	--

Hasil klasterisasi kemudian divisualisasikan dalam bentuk *choropleth map* berbasis SIG menggunakan *library* Folium dan format GeoJSON untuk setiap tahun pengamatan. Warna merah mewakili Klaster Rendah dan warna hijau mewakili Klaster Tinggi.

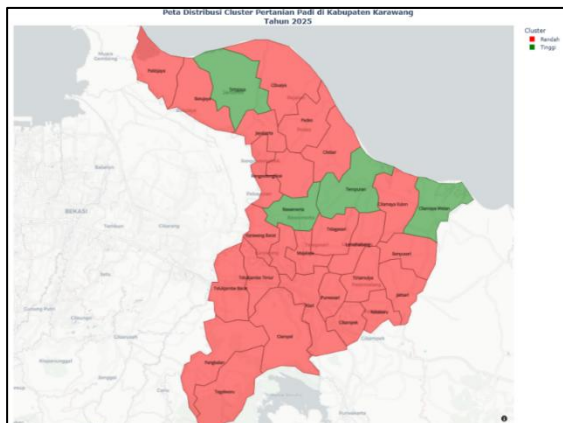


Berdasarkan Gambar 4.18, Pada tahun 2023, Klaster Tinggi (hijau) mencakup Rawamerta, Tempuran, Tirtajaya, dan Batujaya yang berada di bagian utara dan tengah Kabupaten Karawang. Klaster Rendah (merah) mendominasi sebagian besar wilayah tengah dan selatan. Pola ini menunjukkan bahwa wilayah dengan luas panen dan jumlah produksi yang lebih tinggi cenderung terkonsentrasi di bagian utara Kabupaten Karawang.



Berdasarkan Gambar 4.19, Pada tahun 2024, terjadi perubahan pada komposisi Klaster Tinggi. Tempuran dan Batujaya turun ke

Klaster Rendah, sementara Cilamaya Wetan masuk ke Klaster Tinggi untuk pertama kalinya bersama Rawamerta dan Tirtajaya. Klaster Rendah tetap mendominasi wilayah tengah dan selatan Kabupaten Karawang.



Berdasarkan Gambar 4.20, Pada tahun 2025, Klaster Tinggi kembali meluas dengan masuknya Tempuran, sehingga mencakup Cilamaya Wetan, Rawamerta, Tempuran, dan Tirtajaya. Sementara Klaster Rendah tetap mendominasi wilayah tengah dan selatan seperti tahun-tahun sebelumnya.

Secara keseluruhan, peta selama periode 2023–2025 menunjukkan bahwa wilayah dengan luas panen dan jumlah produksi yang lebih tinggi cenderung terkonsentrasi di bagian utara Kabupaten Karawang, sementara wilayah dengan nilai rendah lebih banyak tersebar di bagian tengah dan selatan. Informasi ini diharapkan bisa menjadi acuan bagi pemerintah daerah dalam menentukan prioritas pengembangan dan distribusi bantuan pertanian padi yang lebih tepat sasaran.

5. KESIMPULAN

- a. Klasterisasi wilayah produksi padi di Kabupaten Karawang berhasil dilakukan menggunakan algoritma K-Means dan AHC berdasarkan variabel Luas Panen dan Jumlah Produksi periode 2023 - 2025. Pada K-Means, jumlah klaster optimal ditetapkan pada $k = 3$ melalui metode Elbow dengan nilai SSE 28,83. Pada AHC, jumlah klaster ditetapkan pada $k = 2$ berdasarkan hasil analisis dendrogram dengan metode Average Linkage yang

menghasilkan nilai Cophenetic Correlation tertinggi sebesar 0,7897.

- b. Dari hasil evaluasi menggunakan Davies-Bouldin Index dan Silhouette Score, AHC terbukti lebih baik dibanding K-Means dengan nilai DBI 0,4870 dan Silhouette Score 0,5677, sedangkan K-Means memperoleh nilai DBI 0,5403 dan Silhouette Score 0,5673. Hal ini terjadi karena K-Means menentukan centroid secara acak di awal sehingga kurang mampu menangkap data yang variasinya besar seperti Kecamatan Rawamerta dan Tempuran, sementara AHC membangun klaster secara bertahap berdasarkan kedekatan jarak asli antar data sehingga pengelompokan yang terbentuk lebih mencerminkan kondisi nyata. Atas dasar itu, AHC dipilih sebagai model terbaik dan digunakan pada tahap interpretasi yang mencakup analisis statistik deskriptif per klaster, pemberian label, visualisasi scatter plot, serta pemetaan distribusi spasial wilayah produksi padi di Kabupaten Karawang. Hasilnya menunjukkan 79 dari 90 data masuk ke Klaster Rendah, sementara hanya 11 data yang masuk Klaster Tinggi, dengan wilayah berluas panen dan jumlah produksi tinggi umumnya terkonsentrasi di bagian utara Karawang seperti Rawamerta, Cilamaya Wetan, Tirtajaya, dan Tempuran.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Dr. Aries Suharso, S.Si., M.Kom. dan Dr. Jajam Haerul Jaman, S.E., M.Kom., selaku dosen Program Studi Informatika Universitas Singaperbangsa Karawang, atas bimbingan, arahan, dan dukungan yang telah diberikan selama proses penelitian dan penyusunan artikel ini.

DAFTAR PUSTAKA

- [1] E. Tahir, T. Setiady, and M. Fadhilah, "Implementasi Asuransi Usaha Tani Padi Sebagai Perlindungan Hukum Bagi Petani Di Kabupaten Karawang Akibat Bencana Banjir," *J. Yustia*, vol. 11, no. 2021, pp. 139–159, 2025.
- [2] A. Ghaffar, A. Zaki, and Y. Y. Setyawan, "Determinasi Produktivitas Padi Nasional Dengan Analisis Data Panel Provinsi Di Indonesia 2020 – 2024," *J. Kumpul. Artik. Agribisnis*, vol. 05, no. 01, pp. 139–158, 2025.

- [3] D. Sahara, Abubakar, and L. Nur'azkiya, "Strategi Pengembangan Agribisnis Komoditas Padi Di Kabupaten Karawang Pasca Pandemi Covid-19," *J. Pemikir. Masy. Ilm. Berwawasan Agribisnis.*, vol. 8, no. 1, pp. 168–181, 2022.
- [4] M. A. Khumaini, "Pemkab Karawang Perluas lahan Untuk Kejar Target Produksi Gabah," *Antara News Megapolitan*, Apr. 2026. [Online]. Available: <https://megapolitan.antaranews.com/berita/417017/kabupaten-karawang-terus-kejar-produksi-padi-tahun-2025>
- [5] A. Adriansyah, D. Fitri, Fauzi, Farhan, Juliana, and Khafi, "Penerapan Decision Tree Dalam Data Mining Menggunakan Bahasa Pemrograman Python," *JRIIN J. Ris. Inform. dan Inov.*, vol. 3, no. 7, pp. 2083–2087, 2025.
- [6] N. L. Wibowo, M. A. Soeleman, and A. Z. Fanani, "Antlion Optimizer Algorithm Modification for Initial Centroid," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 158, pp. 1–6, 2023.
- [7] N. A. Maori and H. Mulyo, "Metode Single Linkage Pada Agglomerative Hierarchical Clustering Dalam Penentuan Tingkat Kepuasan Mahasiswa Terhadap Layanan Akademik," *J. SIMETRIS*, vol. 15, no. 2, pp. 1–10, 2024.
- [8] T. Abdulfatah, B. N. Sari, and Susilawati, "Algoritma K-Means Dan Agglomerative Hierarchical Clustering Untuk Pengelompokan Daerah Penghasil Padi Di Indonesia," *JITET (Jurnal Inform. dan Tek. Elektro Ter.)*, vol. 13, no. 3, 2025.
- [9] D. Mayangsari and T. Wahyudi, "Klasifikasi Seleksi Penerimaan Karyawan Baru Menggunakan Algoritma C4.5," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. January, pp. 26–34, 2025.
- [10] F. M. Hana, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5," 2020.
- [11] N. Hendrastuty, "Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa," *J. Ilim. Inform. Dan Ilmu Komput.*, vol. 3, pp. 46–56, 2024.
- [12] R. T. Aldisa, "Data Mining Penentuan Jurusan Siswa Menggunakan Metode Agglomerative Hierarchical Clustering (AHC)," *J. MEDIA Inform. BUDIDARMA*, vol. 7, no. April, pp. 873–880, 2023, doi: 10.30865/mib.v7i2.6092.
- [13] Lisa and T. Handhayani, "Perbandingan Bisecting K-Means Dan Agglomerative Hierarchical Clustering Untuk Pengelompokan Daerah Penghasil Kacang Hijau," *J. Mnemon.*, vol. 8, no. 2, pp. 175–182, 2025.
- [14] B. Hartono, V. Lusiana, and I. H. Al Amin, "Perbandingan Proses Klasterisasi Data Menggunakan K-Means Clustering dan Agglomerative Hierarchical Clustering," *JURIKOM (Jurnal Ris. Komputer)*, vol. 12, no. 4, pp. 628–635, 2025, doi: 10.30865/jurikom.v12i4.8766.
- [15] M. Sholeh and K. Aeni, "Perbandingan Evaluasi Metode Davies Bouldin , Elbow DAN Silhouette Pada Model Clustering Dengan Menggunakan Algoritma K-Means," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.)*, vol. 8, no. 1, 2023.
- [16] B. Hartono and V. Lusiana, "Analisis Metode Elbow SSE , Silhouette Score , dan Jaccard Stability dalam Pemilihan Jumlah Kluster Data yang Optimal TIN: Terapan Informatika Nusantara," *TIN Terap. Inform. Nusant.*, vol. 6, no. 8, pp. 1521–1532, 2026, doi: 10.47065/tin.v6i8.9271.
- [17] S. D. Raihannabil, "Penerapan Cophenetic Correlation pada Pemilihan Metode Pembentukan Dendrogram untuk Mengelompokkan Alat Kontrasepsi Peserta KB Aktif dengan Pendekatan Bottom-Up (Studi pada Level Kabupaten / Kota di Provinsi NTT Tahun 2023)," *JSTAR (Jurnal Stat. Ter.)*, vol. 4, no. 2, pp. 1–16, 2024.
- [18] A. Amrullah, I. Purnamasari, B. N. Sari, and A. Voutama, "Analisis Cluster Faktor Penunjang Pendidikan Menggunakan Algoritma K-Means (Studi Kasus : Kabupaten Karawang)," *JIRE (Jurnal Inform. Rekayasa Elektron.)*, vol. 5, no. 2, pp. 244–252, 2022.
- [19] N. A. Yolandari, L. E. Butarbutar, G. C. H. Rajagukguk, M. F. Zulfi, Arnita, and F. Ramadhani, "Analisis Perbandingan K-Means Dan DBSCAN Dalam Pengelompokan Data Travel Review Ratings Menggunakan Evaluasi Silhouette Index Dan Davies-Bouldin Index," *JITET (Jurnal Inform. dan Tek. Elektro Ter.)*, vol. 13, no. 3, 2025.
- [20] D. F. Angraeni, N. Rahaningsih, R. D. Dana, and C. L. Rohmat, "Pemanfaatan Algoritma K-Means Dalam Analisis Data Penjualan Toko Buyung Upik JS Di Lazada," *JITET (Jurnal Inform. dan Tek. Elektro Ter.)*, vol. 13, no. 2, pp. 1297–1306, 2025.
- [21] M. Subur, Martanto, and U. Hayati, "Clustering Penduduk Miskin Untuk Penerima Bantuan Sosial Menggunakan Algoritma K-Means (Studi Kasus: Desa Grege)," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 8, no. 1, pp. 789–795, 2024.