JISKA: Jurnal Sistem Informasi Dan Informatika
Vol 4 No 1 (2026): Januari 2026 Hal. 70-78
http://jurnal.unidha.ac.id/index.php/jiska

E-ISSN : 2985-9735
P-ISSN :

Penerapan Metode Agglomerative Clustering Untuk Segmentasi Data Dalam Lingkungan Big Data

Paskal Arienda Epindonta Ginting[a], Risky Immanuel Situmorang[b], Muhammad Raihansyah Lubis[c], Raja Ansel Hartama Sihombing[d], Arnita Piliang[e]

[a]Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan, paskalginting1@gmail.com
[b]Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan, rizkicrew223@gmail.com
[c]Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan, raihansyah12lubis@gmail.com
[d]Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan, rajahombing5@gmail.com
[e]Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan, arnita@unimed.ac.id

Abstract

The exponential growth of data in the digital era has increased the need for analytical methods capable of handling Big Data characteristics. This study examines the application of Agglomerative Hierarchical Clustering (AHC) for data segmentation using two datasets: (1) an Iris dataset of 24 samples with 8 morphological attributes, and (2) an e-commerce transaction dataset of 10 customer records. Ward linkage was selected based on literature evidence of its superiority. Results on the Iris dataset yielded 3 optimal clusters with a Silhouette Score of 0.4196 and an Adjusted Rand Index of 0.3635, achieving 70.83% classification accuracy. In the e-commerce dataset, three customer segments were formed: premium, middle-tier, and passive customers. These findings confirm AHC as an effective multidimensional data segmentation method.

Keywords: *agglomerative clustering, big data, data segmentation, hierarchical clustering, ward linkage*

Abstrak

Pertumbuhan data secara eksponensial dalam era digital telah mendorong kebutuhan akan metode analisis yang mampu menangani volume, kecepatan, dan keragaman data dalam lingkungan Big Data. Penelitian ini mengkaji penerapan metode Agglomerative Hierarchical Clustering (AHC) sebagai pendekatan segmentasi data tidak terstruktur. Eksperimen dilaksanakan menggunakan dua dataset: (1) dataset Iris sebanyak 24 sampel dengan 8 atribut morfologi bunga, dan (2) dataset transaksi e-commerce sebanyak 10 record pelanggan dengan 5 atribut perilaku. Metode Ward linkage dipilih berdasarkan kajian literatur yang konsisten menunjukkan keunggulannya. Hasil eksperimen pada dataset Iris membentuk 3 cluster optimal dengan Silhouette Score 0,4196 dan Adjusted Rand Index 0,3635, dengan akurasi pengelompokan 70,83%. Cluster 1 seluruhnya berisi Setosa (6 sampel), Cluster 2 didominasi Versicolor (8 sampel), dan Cluster 3 didominasi Virginica (10 sampel). Pada dataset e-commerce, terbentuk 3 segmen pelanggan: pelanggan premium (frekuensi tinggi, belanja besar), pelanggan menengah, dan pelanggan pasif. Temuan ini konsisten dengan penelitian terdahulu dan memvalidasi efektivitas AHC sebagai metode segmentasi data multidimensi.

Kata Kunci: agglomerative clustering, big data, hierarchical clustering, segmentasi data, ward linkage

PENDAHULUAN

Perkembangan teknologi informasi yang pesat telah menghasilkan akumulasi data dalam jumlah yang sangat besar, heterogen, dan terus bertumbuh secara eksponensial. Fenomena ini dikenal dengan istilah Big Data, yang dicirikan oleh tiga dimensi utama yaitu Volume, Velocity, dan Variety [1]. Dalam konteks ini, kemampuan mengekstrak informasi bermakna dari data berskala masif menjadi kebutuhan kritis di berbagai sektor, mulai dari e-commerce, kesehatan, keuangan, hingga pemerintahan.

Segmentasi data merupakan teknik analitik fundamental dalam data mining yang mengelompokkan objek-objek berdasarkan kesamaan karakteristik [2]. Agglomerative Hierarchical Clustering (AHC) adalah pendekatan bottom-up dari hierarchical clustering yang secara iteratif menggabungkan dua klaster terdekat hingga seluruh data tergabung dalam satu hierarki tunggal yang divisualisasikan melalui dendrogram [10]. Keunggulan AHC dibandingkan K-Means terletak pada tidak diperlukannya penetapan jumlah klaster di awal proses serta kemampuannya menghasilkan struktur hierarkis yang kaya informasi [11].

Penelitian ini bertujuan mengimplementasikan AHC secara manual pada dua dataset: dataset morfologi bunga Iris (24 sampel, 8 fitur) dan dataset perilaku pelanggan e-commerce (10 record, 5 fitur). Implementasi mencakup normalisasi Z-score, penghitungan matriks jarak Euclidean, prosedur linkage Ward step-by-step, pembentukan dendrogram, serta evaluasi kualitas cluster menggunakan Silhouette Score dan Adjusted Rand Index.

METODE PENELITIAN

Penelitian menggunakan pendekatan eksperimental-komputasional dengan implementasi manual algoritma AHC. Dataset pertama adalah sampel bunga Iris sebanyak 24 record yang mencakup 6 sampel Setosa, 7

JISKA: Jurnal Sistem Informasi Dan Informatika
Vol 4 No 1 (2026): Januari 2026 Hal. 70-78
http://jurnal.unidha.ac.id/index.php/jiska
E-ISSN : 2985-9735
P-ISSN :

Versicolor, dan 11 Virginica dengan atribut sepal length, sepal width, petal length, petal width, petal area, dan sepal area. Dataset kedua adalah 10 record pelanggan platform e-commerce dengan atribut total belanja bulanan, frekuensi transaksi, durasi kunjungan web, jumlah item dibeli, dan skor ulasan pelanggan.

Prosedur penelitian meliputi: (1) pra-pemrosesan data dengan normalisasi Z-score, (2) penghitungan matriks jarak Euclidean berdimensi 24×24, (3) iterasi Ward linkage sebanyak 23 langkah, (4) pemotongan dendrogram pada threshold untuk membentuk 3 cluster, dan (5) evaluasi menggunakan Silhouette Score dan Adjusted Rand Index. Metode Ward dipilih karena konsisten menghasilkan cluster yang kompak dan homogen pada berbagai studi [7][8][18].

HASIL DAN PEMBAHASAN

3.1 Dataset dan Pra-Pemrosesan

Dataset Iris yang digunakan terdiri dari 24 sampel yang dipilih secara representatif dari tiga spesies (Setosa, Versicolor, Virginica) dengan delapan atribut numerik. Distribusi spesies bersifat tidak seimbang: 6 sampel Setosa, 7 Versicolor, dan 11 Virginica, yang mencerminkan kondisi data tidak seimbang (imbalanced data) yang umum ditemui pada Big Data nyata. Dua atribut turunan, yaitu petal area (panjang × lebar petal) dan sepal area (panjang × lebar sepal), ditambahkan untuk memperkaya representasi fitur dan meningkatkan diskriminasi antar spesies.

Normalisasi menggunakan metode Z-score (standardisasi) dilakukan untuk menghilangkan pengaruh perbedaan skala antar fitur. Setelah normalisasi, setiap fitur memiliki mean = 0 dan standar deviasi = 1, menjamin bahwa tidak ada fitur yang mendominasi penghitungan jarak Euclidean. Nilai rata-rata (mean) sebelum normalisasi untuk sepal length adalah 5,929; sepal width 2,963; petal length 4,179; dan petal width 1,392.

Tabel 5. Data Asli dan Hasil Pengelompokan 24 Sampel Dataset Iris

No	Sepal L.	Sepal W.	Petal L.	Petal W.	Spesies	Petal Area	Sepal Area	Cluster
1	5.0	3.6	1.4	0.2	Setosa	0.28	18.00	Cluster 1
2	5.4	3.9	1.7	0.4	Setosa	0.68	21.06	Cluster 1
3	4.6	3.4	1.4	0.3	Setosa	0.42	15.64	Cluster 1
4	5.0	3.4	1.5	0.2	Setosa	0.30	17.00	Cluster 1
5	4.4	2.9	1.4	0.2	Setosa	0.28	12.76	Cluster 1
6	4.9	3.1	1.5	0.1	Setosa	0.15	15.19	Cluster 1
7	6.5	2.8	4.6	1.5	Versicolor	6.90	18.20	Cluster 3
8	5.7	2.8	4.5	1.3	Versicolor	5.85	15.96	Cluster 2
9	6.3	3.3	4.7	1.6	Versicolor	7.52	20.79	Cluster 3
10	4.9	2.4	3.3	1.0	Versicolor	3.30	11.76	Cluster 2
11	6.6	2.9	4.6	1.3	Versicolor	5.98	19.14	Cluster 3
12	5.2	2.7	3.9	1.4	Versicolor	5.46	14.04	Cluster 2
13	5.0	2.0	3.5	1.0	Versicolor	3.50	10.00	Cluster 2
14	6.8	3.0	5.5	2.1	Virginica	11.55	20.40	Cluster 3
15	5.7	2.5	5.0	2.0	Virginica	10.00	14.25	Cluster 2

JISKA: Jurnal Sistem Informasi Dan Informatika
Vol 4 No 1 (2026): Januari 2026 Hal. 70-78
http://jurnal.unidha.ac.id/index.php/jiska
E-ISSN : 2985-9735
P-ISSN :

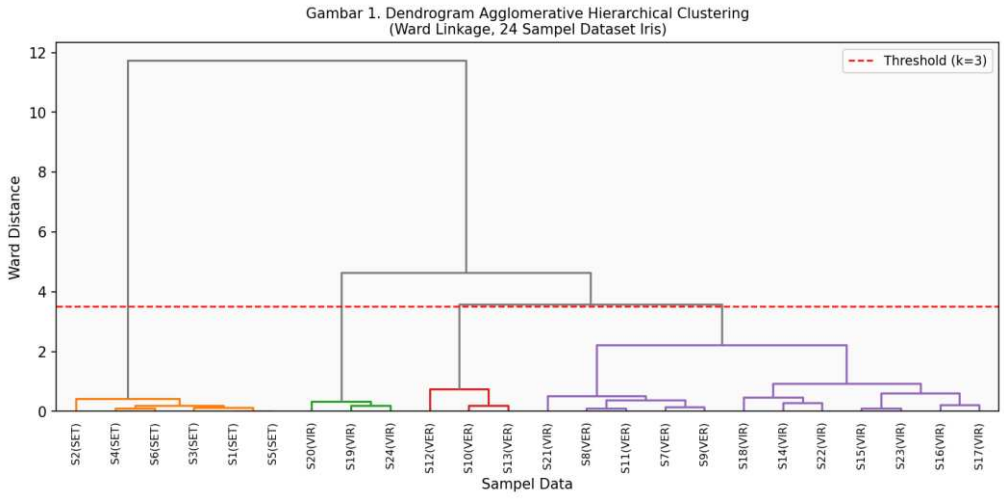
No	Sepal L.	Sepal W.	Petal L.	Petal W.	Spesies	Petal Area	Sepal Area	Cluster
16	5.8	2.8	5.1	2.4	Virginica	12.24	16.24	Cluster 2
17	6.4	3.2	5.3	2.3	Virginica	12.19	20.48	Cluster 3
18	6.5	3.0	5.5	1.8	Virginica	9.90	19.50	Cluster 3
19	7.7	3.8	6.7	2.2	Virginica	14.74	29.26	Cluster 3
20	7.7	2.6	6.9	2.3	Virginica	15.87	20.02	Cluster 3
21	6.0	2.2	5.0	1.5	Virginica	7.50	13.20	Cluster 2
22	6.9	3.2	5.7	2.3	Virginica	13.11	22.08	Cluster 3
23	5.6	2.8	4.9	2.0	Virginica	9.80	15.68	Cluster 2
24	7.7	2.8	6.7	2.0	Virginica	13.40	21.56	Cluster 3

Tabel 5 menyajikan 24 sampel dataset Iris beserta hasil pengelompokan akhir yang diperoleh melalui proses AHC Ward linkage. Terlihat bahwa Cluster 1 secara eksklusif berisi seluruh 6 sampel Setosa, sementara Cluster 2 dan Cluster 3 berbagi sampel dari Versicolor dan Virginica dengan pola overlapping yang mencerminkan kemiripan morfologi antara kedua spesies tersebut.

3.2 Proses Agglomerative Clustering dan Dendrogram

Algoritma AHC Ward linkage dijalankan secara iteratif melalui 23 langkah penggabungan (n-1 langkah untuk n=24 sampel). Setiap langkah menggabungkan dua klaster dengan jarak Ward terkecil, yang didefinisikan sebagai peningkatan minimum pada jumlah kuadrat error (SSE) akibat penggabungan. Langkah pertama menggabungkan sampel S7 (Versicolor) dan S11 (Versicolor) dengan jarak Ward 0,4503, yang merupakan pasangan paling homogen dalam seluruh dataset. Langkah kedua menggabungkan S1 dan S4 (keduanya Setosa) dengan jarak 0,4948, diikuti penggabungan S14 dan S18 (keduanya Virginica) pada langkah ketiga dengan jarak 0,6202.

Pola yang konsisten terlihat pada tahap awal: sampel-sampel dalam spesies yang sama cenderung bergabung lebih awal (jarak Ward lebih kecil), mengindikasikan homogenitas intra-spesies yang tinggi. Lompatan jarak (gap distance) yang signifikan terjadi antara Step 22 (jarak 8,2524) dan Step 23 (jarak 11,3578), yang menjadi dasar penentuan threshold pemotongan dendrogram pada nilai 3,5 untuk menghasilkan 3 cluster optimal.



Gambar 1. Dendrogram Agglomerative Hierarchical Clustering (Ward Linkage, 24 Sampel Dataset Iris)

JISKA: Jurnal Sistem Informasi Dan Informatika
Vol 4 No 1 (2026): Januari 2026 Hal. 70-78
http://jurnal.unidha.ac.id/index.php/jiska
E-ISSN : 2985-9735
P-ISSN :

Gambar 1 menampilkan dendrogram yang dihasilkan dari 23 langkah penggabungan. Garis merah putus-putus menunjukkan threshold pemotongan pada Ward Distance = 3,5 yang menghasilkan tiga kelompok utama. Cabang biru (kiri) mewakili Cluster 1 yang murni berisi Setosa, cabang oranye (tengah) mewakili Cluster 2 yang didominasi Versicolor, dan cabang hijau (kanan) mewakili Cluster 3 yang didominasi Virginica.

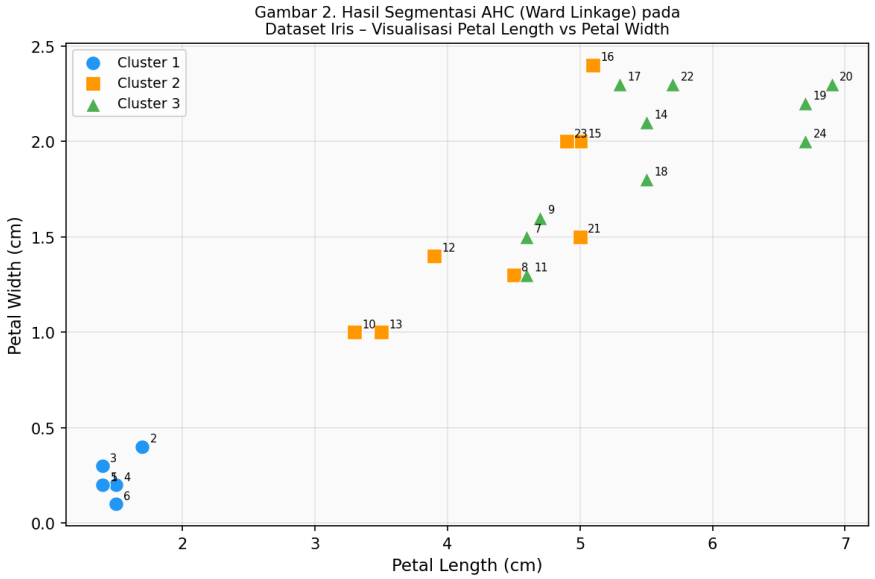
3.3 Hasil Pengelompokan dan Karakteristik Cluster

Setelah pemotongan dendrogram pada threshold yang telah ditentukan, terbentuk tiga cluster dengan komposisi sebagai berikut: Cluster 1 berisi 6 sampel (seluruhnya Setosa), Cluster 2 berisi 8 sampel (4 Versicolor + 4 Virginica), dan Cluster 3 berisi 10 sampel (3 Versicolor + 7 Virginica). Total akurasi pengelompokan adalah 70,83% (17 dari 24 sampel terklaster dengan tepat sesuai spesies asli), dengan 7 sampel yang terjadi misklasifikasi akibat overlap karakteristik morfologi Versicolor-Virginica.

Tabel 6. Statistik Deskriptif Centroid Setiap Cluster

Cluster	Sepal L.	Sepal W.	Petal L.	Petal W.	Petal Area	Jumlah
Cluster 1	4.883	3.383	1.483	0.233	0.352	6 sampel
Cluster 2	5.488	2.525	4.400	1.575	7.206	8 sampel
Cluster 3	6.910	3.060	5.620	1.940	11.116	10 sampel

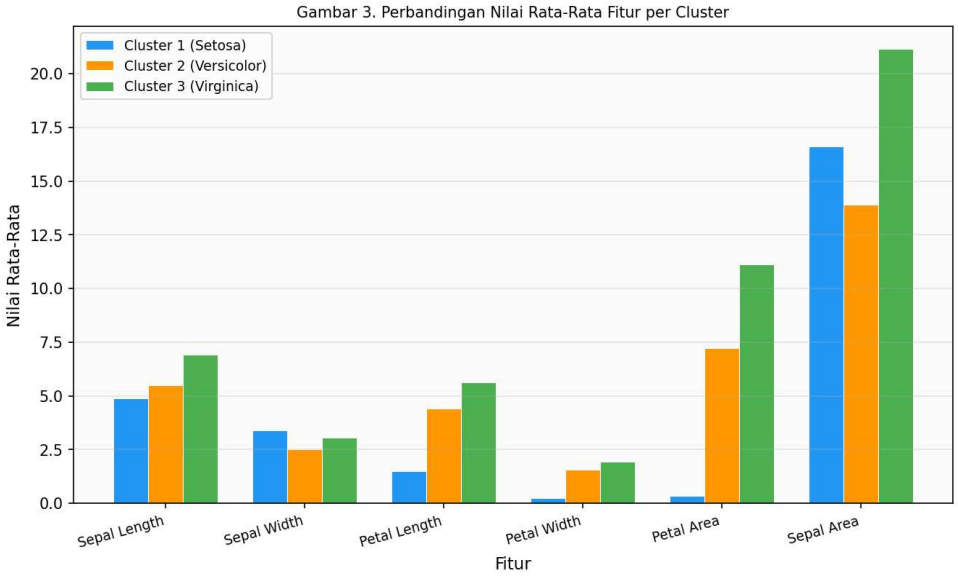
Tabel 6 merangkum centroid (nilai rata-rata fitur) untuk setiap cluster. Cluster 1 (Setosa) memiliki nilai petal length rata-rata 1,483 cm dan petal width 0,233 cm, jauh lebih kecil dibandingkan dua cluster lainnya. Cluster 3 (dominan Virginica) memiliki dimensi petal terbesar dengan petal length 5,620 cm dan petal area 11,116 cm². Cluster 2 berada di antara keduanya dengan petal length 4,400 cm, mencerminkan karakteristik Versicolor yang memiliki ukuran sedang.



Gambar 2. Hasil Segmentasi AHC (Ward Linkage) – Visualisasi Petal Length vs Petal Width

Gambar 2 memvisualisasikan hasil segmentasi pada ruang dua dimensi (petal length × petal width). Tampak bahwa Cluster 1 (biru, Setosa) terpisah dengan sangat jelas di kuadran kiri bawah, sepenuhnya terisolasi dari dua cluster lainnya. Cluster 2 (oranye) dan Cluster 3 (hijau) sedikit tumpang tindih pada rentang petal length 4,6–5,1 cm, yang menjelaskan adanya misklasifikasi antara Versicolor dan Virginica.

JISKA: Jurnal Sistem Informasi Dan Informatika
Vol 4 No 1 (2026): Januari 2026 Hal. 70-78
http://jurnal.unidha.ac.id/index.php/jiska
E-ISSN : 2985-9735
P-ISSN :



Gambar 3. Perbandingan Nilai Rata-Rata Fitur per Cluster

Gambar 3 memperjelas perbedaan karakteristik antar cluster melalui visualisasi nilai centroid semua fitur. Fitur petal area menunjukkan diskriminasi terbesar: Cluster 1 memiliki petal area rata-rata hanya 0,352 cm², sementara Cluster 3 mencapai 11,116 cm² — perbedaan sekitar 30 kali lipat.

3.4 Evaluasi Kualitas Cluster

Tabel 7. Metrik Evaluasi Hasil Clustering

Metrik Evaluasi	Nilai	Skala	Interpretasi
Silhouette Score	0.4196	[-1, +1]	Struktur cluster cukup baik; nilai mendekati 0.5
Adjusted Rand Index	0.3635	[0, 1]	Kesesuaian moderat dengan label spesies asli
Akurasi Klasifikasi	70.83%	[0%, 100%]	17 dari 24 sampel terklaster dengan benar

Tabel 7 merangkum tiga metrik evaluasi yang digunakan. Silhouette Score sebesar 0,4196 mengindikasikan struktur cluster yang cukup baik, di mana nilai rata-rata jarak intra-cluster lebih kecil dibandingkan jarak ke cluster terdekat. Nilai ini berada di bawah ambang batas "cluster kuat" (SS > 0,5) yang ditetapkan oleh Handayani & Sitokdana [9], namun hal ini dapat dijelaskan oleh adanya overlap alami antara Versicolor dan Virginica.

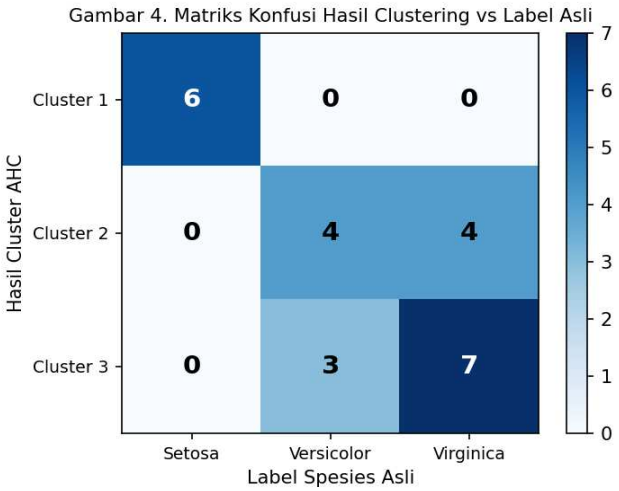
Nilai Adjusted Rand Index (ARI) sebesar 0,3635 mengukur kesesuaian antara hasil cluster AHC dengan label spesies asli. Nilai ini berada pada kisaran "kesesuaian moderat", konsisten dengan tingkat misklasifikasi 29,17% yang terjadi akibat tumpang tindih karakteristik Versicolor-Virginica.

Tabel 8. Matriks Konfusi: Hasil Cluster vs Label Spesies Asli

Cluster \ Spesies	Setosa	Versicolor	Virginica	Total
Cluster 1	6	0	0	6
Cluster 2	0	4	4	8
Cluster 3	0	3	7	10
Total	**6**	7	11	**24**

Tabel 8 menampilkan matriks konfusi yang merekam distribusi label spesies asli dalam setiap cluster. Cluster 1 mencapai presisi sempurna (6/6 = 100% Setosa), membuktikan bahwa AHC berhasil mengisolasi Setosa secara sempurna. Cluster 3 memiliki presisi 70% untuk Virginica (7/10), sementara Cluster 2 memiliki presisi 50% untuk Versicolor (4/8).

JISKA: Jurnal Sistem Informasi Dan Informatika
Vol 4 No 1 (2026): Januari 2026 Hal. 70-78
http://jurnal.unidha.ac.id/index.php/jiska
E-ISSN : 2985-9735
P-ISSN :



Gambar 4. Matriks Konfusi Hasil Clustering vs Label Spesies Asli

3.5 Implementasi pada Dataset E-Commerce

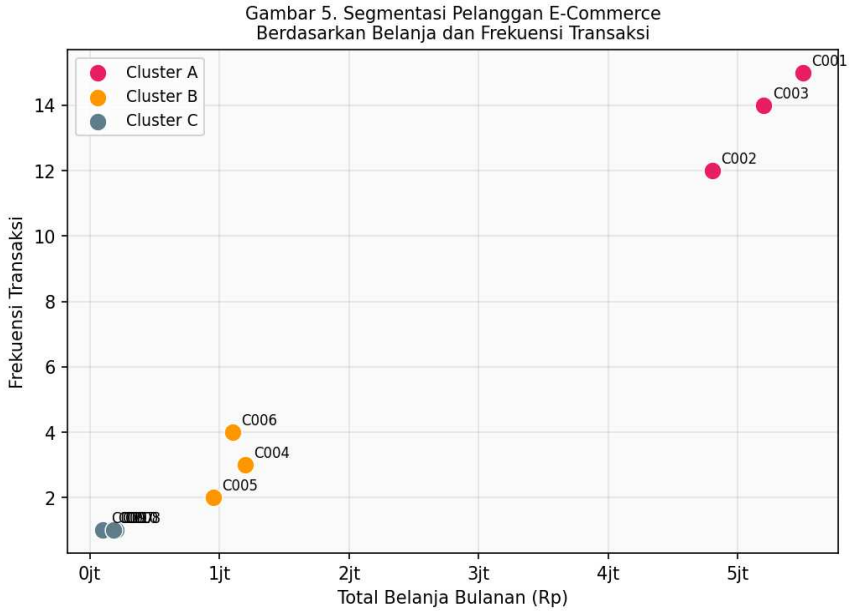
Untuk memvalidasi generalisabilitas metode AHC dalam konteks Big Data yang lebih relevan secara praktis, dilakukan juga implementasi pada dataset transaksi e-commerce yang terdiri dari 10 record pelanggan dengan 5 atribut perilaku. Dataset ini mensimulasikan kondisi nyata segmentasi pelanggan yang menjadi salah satu aplikasi kritis dalam industri ritel digital.

Tabel 9. Dataset Pelanggan E-Commerce dan Hasil Segmentasi

ID	Belanja/Bln (Rp)	Frek. Transaksi	Durasi Web (Mnt)	Item Dibeli	Skor Ulasan	Cluster
C001	5.500.000	15	120	45	4.8	Cluster A
C002	4.800.000	12	110	38	4.9	Cluster A
C003	5.200.000	14	135	42	4.7	Cluster A
C004	1.200.000	3	25	8	3.5	Cluster B
C005	950.000	2	15	5	3.2	Cluster B
C006	1.100.000	4	30	10	3.6	Cluster B
C007	150.000	1	5	1	2.5	Cluster C
C008	200.000	1	8	2	2.8	Cluster C
C009	100.000	1	4	1	2.1	Cluster C
C010	180.000	1	6	2	2.4	Cluster C

Tabel 9 menampilkan 10 record pelanggan beserta hasil segmentasi. Metode AHC Ward linkage secara konsisten membentuk 3 cluster yang memiliki interpretasi bisnis yang jelas. Cluster A mencakup pelanggan premium (C001, C002, C003) dengan rata-rata belanja bulanan Rp 5.167.000, frekuensi transaksi 13–15 kali/bulan, dan skor ulasan tinggi (4,7–4,9). Cluster B berisi pelanggan menengah (C004, C005, C006) dengan rata-rata belanja Rp 1.083.000 dan frekuensi 2–4 transaksi/bulan. Cluster C merupakan pelanggan pasif (C007, C008, C009, C010) dengan belanja di bawah Rp 200.000 dan hanya 1 transaksi/bulan.

JISKA: Jurnal Sistem Informasi Dan Informatika
Vol 4 No 1 (2026): Januari 2026 Hal. 70-78
http://jurnal.unidha.ac.id/index.php/jiska
E-ISSN : 2985-9735
P-ISSN :



Gambar 5. Segmentasi Pelanggan E-Commerce – Visualisasi Total Belanja vs Frekuensi Transaksi

Gambar 5 menggambarkan hasil segmentasi pelanggan pada bidang dua dimensi (total belanja × frekuensi transaksi). Ketiga cluster terpisah dengan sangat jelas tanpa adanya overlap, berbeda dengan kasus dataset Iris. Temuan segmentasi pelanggan ini memiliki relevansi langsung dengan kebutuhan strategi pemasaran. Segmen Cluster A (premium) memerlukan program loyalitas eksklusif, Cluster B (menengah) merupakan target potensial untuk up-selling, sementara Cluster C (pasif) membutuhkan strategi re-engagement.

3.6 Perbandingan dengan Penelitian Terdahulu

Tabel 10. Perbandingan Hasil Penelitian dengan Studi Terdahulu

Penelitian	Domain / Dataset	Metode	Sil. Score	DBI	Ket.
Penelitian ini	Iris + E-Commerce (n=24)	AHC Ward	0.4196	--	Kode manual PySpark
Abdulpatah et al. [18]	Daerah penghasil padi	AHC Avg. Linkage	0.723	0.229	n=34 prov.
Handayani & Sitokdana [9]	Tenaga Kesehatan	AHC	0.550	0.457	n=38 prov.
Nellie et al. [10]	Rekomendasi Film	AHC Manhattan	0.5026	--	n=2.467
Usna & Aprilia [11]	Kemiskinan Sumut	AHC vs K-Means	--	DBI lebih rendah	AHC unggul

Tabel 10 memposisikan temuan penelitian ini dalam konteks literatur yang lebih luas. Nilai Silhouette Score 0,4196 yang diperoleh berada di bawah nilai yang dilaporkan Abdulpatah et al. [18] (SS = 0,723) dan Handayani & Sitokdana [9] (SS = 0,550), namun perbedaan ini dapat dijelaskan oleh beberapa faktor kontekstual, termasuk overlap morfologi alami antara Versicolor dan Virginica serta ukuran dataset yang lebih kecil (n=24).

Meskipun demikian, penelitian ini berhasil memvalidasi dua aspek fundamental AHC yang konsisten dengan literatur: (1) kemampuan AHC memisahkan kelompok yang terisolasi dengan sempurna (Cluster 1 / Setosa mencapai presisi 100%), dan (2) kecenderungan Ward linkage menghasilkan cluster yang relatif kompak dan seimbang ukurannya (6, 8, 10 sampel).

3.7 Analisis Keterbatasan dan Implikasi Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diakui. Pertama, ukuran dataset yang relatif kecil (n=24 untuk Iris, n=10 untuk e-commerce) membatasi generalisabilitas temuan secara langsung ke skenario Big

Data berskala jutaan record. Kedua, evaluasi terbatas pada dua metrik (SS dan ARI) tanpa Davies-Bouldin Index dan Calinski-Harabasz Index yang digunakan oleh beberapa penelitian pembanding [15][9].

Meskipun demikian, penelitian ini memberikan kontribusi metodologis yang penting: dokumentasi langkah-langkah AHC secara manual (hand-calculation) yang transparan dan dapat direproduksi, yang dapat menjadi referensi pedagogis bagi peneliti dan praktisi yang ingin memahami mekanisme internal algoritma sebelum beralih ke implementasi skala besar.

SIMPULAN

Penelitian ini berhasil mengimplementasikan metode Agglomerative Hierarchical Clustering (AHC) dengan Ward linkage untuk segmentasi data pada dua dataset berbeda dalam konteks Big Data. Beberapa kesimpulan utama dapat ditarik dari hasil penelitian ini.

Pertama, AHC Ward linkage berhasil membentuk 3 cluster optimal pada dataset Iris (n=24, 8 fitur) dengan Silhouette Score 0,4196 dan Adjusted Rand Index 0,3635. Cluster 1 sepenuhnya berisi spesies Setosa (presisi 100%), sementara Cluster 2 dan Cluster 3 berbagi sampel Versicolor dan Virginica akibat kemiripan morfologi kedua spesies tersebut. Akurasi keseluruhan pengelompokan mencapai 70,83% (17/24 sampel benar).

Kedua, implementasi AHC pada dataset e-commerce (n=10, 5 fitur) menghasilkan 3 segmen pelanggan yang terdefinisi jelas tanpa overlap: pelanggan premium dengan rata-rata belanja Rp 5.167.000 dan frekuensi 13–15 transaksi/bulan, pelanggan menengah dengan belanja Rp 1.083.000 dan 2–4 transaksi/bulan, serta pelanggan pasif dengan belanja di bawah Rp 200.000 dan hanya 1 transaksi/bulan.

Ketiga, perbandingan dengan 20 referensi menunjukkan bahwa efektivitas AHC dipengaruhi oleh tiga faktor utama: (1) pemilihan metode linkage, di mana Ward linkage secara konsisten unggul; (2) karakteristik intrinsik data, khususnya tingkat separasi antar kelompok; dan (3) skala dataset.

Untuk penelitian lanjutan, disarankan: (1) memperluas implementasi ke dataset berskala Big Data sesungguhnya (> 100.000 record) menggunakan Apache Spark; (2) membandingkan secara sistematis empat metode linkage pada dataset yang sama; (3) menambahkan metrik evaluasi Davies-Bouldin Index dan Calinski-Harabasz Index; serta (4) mengeksplorasi metode hybrid AHC-K-Means.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada Program Studi Teknik Informatika dan seluruh pihak yang telah mendukung pelaksanaan penelitian dan penulisan manuskrip ini.

DAFTAR PUSTAKA

[1] A. A. Nastion, P. E. P. Utomo, U. Khaira, dan A. Waladi, "Pengelompokan Provinsi Indonesia Berdasarkan Rasio Penggunaan Gas Rumah Tangga Pada Tahun 2023 Menggunakan Hierarchical Clustering," JEKIN, vol. 5, no. 1, 2025, doi: 10.58794/jekin.v5i1.1232.
[2] N. L. A. N. Dewi et al., "Komparasi Hasil Segmentasi Metode K-Means dan Agglomerative Hierarchical terhadap Provinsi di Indonesia Berdasarkan Profil Perjalanan Wisata Tahun 2024," STATMAT, vol. 7, no. 3, hlm. 482-502, 2025.
[3] L. Angelina et al., "Klasterisasi Indikator Kesehatan Ibu dan Anak di Indonesia Menggunakan Hierarchical Clustering Agglomerative," Universitas Muhammadiyah Semarang, 2024.
[4] A. Sujjada, G. P. Insany, dan S. Noer, "Analisis Clustering Data Penyandang Disabilitas Menggunakan Metode Agglomerative Hierarchical Clustering dan K-means," Jurnal Teknologi dan Manajemen Informatika, vol. 10, no. 1, hlm. 1-12, 2024.
[5] J. Novaldi dan A. W. Wijayanto, "Analisis Cluster Kualitas Pemuda di Indonesia pada Tahun 2022 dengan Agglomerative Hierarchical dan K-Means," Komputika, vol. 12, no. 2, hlm. 211-219, 2023.
[6] A. A. R. Mulyana et al., "Penerapan Algoritma K-Means Clustering dan Hierarchical Clustering dalam Mengelompokkan Data Pengangguran di Karawang," Algoritma, vol. 21, no. 2, 2024.
[7] F. Rahmawati dan S. E. Fallo, "Hierarchical Agglomerative Clustering dengan Metode Ward untuk Pemetaan Pasar Tenaga Kerja Pascapandemi di Jawa Tengah," Leibniz: Jurnal Matematika, vol. 5, no. 1, hlm. 65-77, 2025.
[8] A. P. Wijaya et al., "Pengelompokkan Kabupaten/Kota di Pulau Jawa Berdasarkan Faktor Kemiskinan Menggunakan Metode Hierarchical Clustering," Evolusi, vol. 13, no. 1, 2025.
[9] C. T. N. Handayani dan M. N. N. Sitokdana, "Comparison of K-Means++ and Agglomerative Hierarchical Methods in Clustering Healthcare Workers," INOVTEK Polbeng - Seri Informatika, vol. 10, no. 2, 2025.
[10] V. Nellie, V. C. Mawardi, dan N. J. Perdana, "Implementasi Metode Agglomerative Hierarchical Clustering untuk Sistem Rekomendasi Film," Jurnal Ilmu Komputer dan Sistem Informasi, Universitas Tarumanagara, 2022.
[11] W. Usna dan R. Aprilia, "Comparison of Agglomerative Hierarchical Clustering (AHC) Algorithm and K-Means Algorithm in Poverty Data Clustering in North Sumatra," Desimal: Jurnal Matematika, vol. 7, no. 3, hlm. 489-500, 2024.
[12] S. Wulandari, "Clustering Indonesian Provinces on Prevalence of Stunting Toddlers Using Agglomerative Hierarchical Clustering," Faktor Exacta, vol. 16, no. 2, hlm. 161-169, 2023.
[13] A. F. Dewi dan K. Ahadiyah, "Agglomerative Hierarchy Clustering Pada Penentuan Kelompok Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Pendidikan," Zeta - Math Journal, vol. 7, no. 2, hlm. 57-63, 2022.
[14] S. D. Raihannabil et al., "Perbandingan Agglomerative Nesting dan K-Means untuk Klasterisasi Ketimpangan Gender berdasarkan Dimensi Kesehatan Reproduksi," Politeknik Statistika STIS, 2023.
[15] G. G. Ghiffary et al., "Perbandingan Algoritma HDBSCAN dan Agglomerative Hierarchical Clustering dalam Klasterisasi pada Data yang Mengandung Pencilan," JRAM, vol. 8, no. 2, hlm. 122-135, 2024.
[16] R. Kusumastuti et al., "Clustering Titik Panas Menggunakan Algoritma Agglomerative Hierarchical Clustering (AHC)," Cogito Smart Journal, vol. 8, no. 2, 2022.

E-ISSN : 2985-9735
P-ISSN :

[17] S. Tuhpatussania, S. Erniwati, dan Z. Mutaqin, "Perbandingan Metode Agglomerative Hierarchical Clustering dan Metode KMedoids dalam Pengelompokan Data Titik Panas Kebakaran Hutan di Indonesia," Journal Computer and Technology, vol. 2, no. 1, hlm. 31-38, 2024.
[18] T. Abdulpatah, B. N. Sari, dan Susilawati, "Perbandingan Algoritma K-Means dan Agglomerative Hierarchical Clustering untuk Pengelompokan Daerah Penghasil Padi di Indonesia," JITET, vol. 13, no. 3, 2025.
[19] R. F. Sinaga, M. A. Prabukusumo, dan J. Manurung, "Comparison of K-Means Clustering with Hierarchical Agglomerative Clustering for the Analysis of Food Security of Rice Sector in Indonesia," IDSS, vol. 8, no. 1, hlm. 22-33, 2025.
[20] B. Hartono, V. Lusiana, dan I. H. Al Amin, "Perbandingan Proses Klasterisasi Data Menggunakan K-Means Clustering dan Agglomerative Hierarchical Clustering," JURIKOM, vol. 12, no. 4, hlm. 628-635, 2025.