Journal of System & Technology
ISSN: 3110-0171
Vol. 2, No. 1, June 2026

KOMPARASI K-MEANS DAN AGGLOMERATIVE CLUSTERING PADA SEGMENTASI PELANGGAN RITEL

Muhammad Abidillah[1], Rehandra[1], Dian Ramadhani[1]*
[1]Dept.of Informatics Engineering, Faculty of Engineering, Universitas Riau
m.abidillah2005@gmail.com, rehandra2004@gmail.com, dianramadhani@lecturer.unri.ac.id
*Corresponding author: dianramadhani@lecturer.unri.ac.id

Abstract: Segmentasi pelanggan merupakan strategi krusial dalam industri ritel modern untuk mengoptimalkan pemasaran berbasis data. Namun, mayoritas studi komparatif terdahulu hanya mengevaluasi Agglomerative Clustering pada satu konfigurasi linkage default, sehingga sensitivitas performa algoritma terhadap pemilihan kriteria jarak belum banyak dikuantifikasi secara eksplisit. Penelitian ini bertujuan mengisi kesenjangan tersebut dengan membandingkan kinerja K-Means terhadap empat varian Agglomerative Clustering (Single, Complete, Average, dan Ward Linkage) dalam mengelompokkan perilaku transaksi pelanggan, dievaluasi secara kuantitatif menggunakan Silhouette Score dan Davies-Bouldin Index. Lingkungan penelitian dibangun menggunakan platform RapidMiner untuk memproses Synthetic Customer Dataset yang terdiri dari 500 observasi, dengan validasi metrik dihitung secara konsisten melalui scikit-learn. Tahap pra-pemrosesan meliputi seleksi fitur dan normalisasi Z-transformation terhadap lima atribut numerik: Age, TotalSpend, AverageOrderValue, TotalOrders, dan LoyaltyPoints. Hasil eksperimen menunjukkan bahwa K-Means (Silhouette=0,165; DBI=1,592) secara konsisten mengungguli seluruh varian Agglomerative Clustering yang diuji. Single Linkage mengalami chaining effect yang tercermin dari Silhouette Score negatif (-0,123), meskipun nilai DBI-nya tampak rendah (0,777) akibat artefak klaster singleton yang hampir tidak memiliki dispersi internal. Complete (Silhouette=0,082; DBI=1,931), Average (Silhouette=0,108; DBI=1,792), dan Ward Linkage (Silhouette=0,111; DBI=1,927) berhasil menghindari chaining effect namun tetap menghasilkan kualitas partisi yang lebih rendah dibandingkan K-Means. Temuan ini menegaskan bahwa keberhasilan Agglomerative Clustering pada data ritel padat sangat bergantung pada pemilihan kriteria linkage, dan kontribusi utama penelitian ini adalah kuantifikasi eksplisit atas sensitivitas tersebut, termasuk pengungkapan bahwa Davies-Bouldin Index dapat menyesatkan ketika diterapkan pada distribusi klaster yang sangat tidak seimbang.

Keywords—Agglomerative Hierarchical Clustering, Analisis Ritel, Data Mining, K-Means Clustering, RapidMiner

Article Information:
Received: 09-Jun-2026; Revised: 21-Jun-2026; Accepted: 26-Jun-2026; Published: 30-Jun-2026

This work is licensed under a **CC BY-SA**. Copyright ©2026 by Author. Published by Universitas Riau.

I. PENDAHULUAN

Ledakan data konsumen yang terus tumbuh pesat dalam industri ritel masa kini telah mendorong perubahan besar-besaran, mengubah data mentah menjadi aset berharga yang menentukan daya saing perusahaan. Ekosistem digital ritel tidak lagi mengandalkan intuisi manajerial semata, melainkan pada kemampuan sistem komputasi dalam mengolah jutaan catatan transaksi, interaksi antarmuka, dan riwayat keuangan konsumen. Tersembunyi di balik tumpukan *big data* tersebut, terdapat segmen-segmen perilaku konsumen yang mencerminkan berbagai macam persona ekonomi. Jika perusahaan mampu memahami dan membaca pola-pola ini, mereka akan memiliki keunggulan yang signifikan di pasar. Praktik segmentasi pelanggan yang presisi kini tidak lagi sekadar alat pendukung, melainkan telah menjadi fondasi utama dalam pengambilan keputusan strategis, mulai dari desain program loyalitas, personalisasi kampanye promosi yang sangat relevan, penetapan *tiered pricing*, hingga algoritma rekomendasi *cross-selling* yang dipersonalisasi [1].

Dalam melakukan segmentasi pasar tanpa campur tangan bias manusia, teknik *cluster analysis* (*unsupervised learning*) telah diakui sebagai pendekatan komputasi yang paling tepat. Paradigma ini membebaskan proses pengelompokan dari keterbatasan label prediktif yang kaku, membiarkan mesin menemukan pengelompokan alami berdasarkan kedekatan jarak antar data. Berangkat dari landasan historis inilah, penerapan arsitektur klasterisasi di era modern semakin mendominasi riset global, di mana keberhasilan segmentasi data berdimensi tinggi sangat bergantung pada kemampuan algoritma *clustering* dalam menyederhanakan kompleksitas operasional sekaligus menjaga transparansi metodologis bagi para pemangku kepentingan [2].

Dua algoritma yang paling banyak dibahas dalam literatur akademis maupun implementasi industri adalah K-Means Clustering dan Agglomerative Hierarchical Clustering. K-Means bekerja dengan cara membagi ruang data melalui partisi iteratif yang terus memperbarui posisi *centroid*, memaksa setiap titik data masuk ke dalam salah satu dari *k* kelompok terdekat hingga mencapai konvergensi [3].

Berbeda dengan sifat iteratif tersebut, Agglomerative Clustering tidak memerlukan penentuan jumlah kelompok

di awal. Algoritma ini bekerja secara *bottom-up*, membangun hierarki taksonomi berbentuk *dendrogram* dengan cara menggabungkan entitas yang paling dekat satu per satu, berdasarkan fungsi matematis *linkage* yang dipilih oleh analis [4].

Dalam lingkungan sistem *Customer Relationship Management* (CRM) ritel modern, data penjualan perusahaan dapat diolah menjadi pengetahuan berharga untuk mengoptimalkan strategi pemasaran. Penelitian oleh Ramadhani et al. [5] menegaskan bahwa teknik *data mining* sangat penting digunakan untuk menganalisis data pelanggan, memahami perilaku pembelian, serta preferensi produk guna menghasilkan landasan prediksi yang kuat.

Meskipun keandalan komputasional kedua pendekatan ini sudah cukup banyak dieksplorasi, perilaku dan respons masing-masing algoritma ketika berhadapan dengan kepadatan data spasial transaksi ritel masih membutuhkan kajian yang lebih spesifik. Hal ini disebabkan oleh karakteristik data transaksional yang unik dan dinamis, di mana salah satu metode *linkage* pada Agglomerative Clustering rentan terhadap distorsi spasial, sementara metode partisional cenderung menghasilkan pengelompokan yang lebih homogen [6].

Meskipun studi-studi tersebut [1]–[6] telah membandingkan K-Means dan Agglomerative Clustering secara umum, mayoritas penelitian hanya mengevaluasi Agglomerative Clustering pada satu konfigurasi *linkage* tunggal, umumnya tanpa menguji apakah hasil tersebut konsisten pada kriteria jarak lain. Akibatnya, kesimpulan salah satu algoritma sering bersifat spesifik terhadap konfigurasi yang diuji dan berisiko digeneralisasi secara berlebihan. Pada data transaksi ritel yang cenderung padat dan tidak terdistribusi secara terpisah, pemilihan kriteria *linkage* diketahui dapat mengubah struktur klaster secara drastis, mulai dari fenomena *chaining effect* pada Single Linkage hingga distribusi yang lebih seimbang pada Ward atau Complete Linkage. Kesenjangan inilah yang melatarbelakangi penelitian ini, alih-alih membandingkan kedua algoritma pada konfigurasi tunggal, penelitian ini secara sistematis mengevaluasi empat varian linkage Agglomerative Clustering terhadap K-Means menggunakan metrik validasi internal (Silhouette Score dan Davies-Bouldin Index), guna memberikan bukti kuantitatif mengenai sejauh mana pemilihan *linkage* memengaruhi keberhasilan segmentasi pada data ritel padat.

Keberhasilan dalam mengekstraksi informasi tersebut memerlukan pemahaman mendalam tentang integrasi statistik inferensial, pengenalan pola multivariat, dan kecerdasan analitik komputasional untuk mengidentifikasi pola tersembunyi dalam kumpulan basis data berskala besar [7].

Alasan teoretis mengenai pentingnya penanganan karakteristik data transaksional inilah yang mendasari mengapa pengujian variasi kriteria *linkage* pada Agglomerative dan evaluasi metrik jarak pada K-Means menjadi sangat penting dalam penelitian ini, guna memastikan segmen pasar yang terbentuk memiliki relevansi, stabilitas, dan manfaat nyata bagi pengembangan sistem rekomendasi pemasaran [8].

II. METODE PENELITIAN

A. Data Mining dan Cluster Analysis

Data mining merupakan proses penggalian pengetahuan yang tersembunyi dari kumpulan data berukuran besar. Larose [9] mendefinisikan *data mining* sebagai bidang ilmu yang memadukan metode statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi serta mengidentifikasi informasi berguna beserta pola-pola yang terkandung di dalam basis data berskala besar. Dalam konteks intelijen bisnis ritel, analisis klaster memegang peran sentral dalam mendukung paradigma *unsupervised machine learning*. Perbandingan operasional antara algoritma partisional dan algoritma hierarkis terbukti penting bagi intelijen bisnis ritel, di mana rekayasa fitur yang tepat berperan besar dalam meningkatkan performa algoritma *clustering* sekaligus menghasilkan segmentasi pelanggan yang lebih stabil dan akurat dalam mencerminkan perilaku pelanggan [10].

B. Data Collecting

Data yang digunakan dalam penelitian ini adalah *Synthetic Customer Dataset* yang dapat diakses secara publik melalui platform Kaggle. Dataset ini terdiri dari 500 baris observasi pelanggan dan 22 kolom atribut yang mencakup informasi demografis, perilaku belanja, dan preferensi pelanggan, meliputi *CustomerID*, *FirstName*, *LastName*, *Gender*, *Age*, *Email*, *PhoneNumber*, *Address*, *City*, *State*, *Country*, *CustomerTier*, *AccountCreationDate*, *LastPurchaseDate*, *TotalSpend*, *AverageOrderValue*, *TotalOrders*, *MostFrequentCategory*, *LoyaltyPoints*, *PreferredPaymentMethod*, *DiscountEligibility*, dan *PreferredContactMethod*. Dataset ini bersifat sintetis namun dirancang untuk merepresentasikan pola perilaku pelanggan ritel yang realistis, sehingga sesuai digunakan sebagai basis eksperimen segmentasi pelanggan.

C. Preprocessing Data

Tahap praproses bertujuan untuk menyiapkan data mentah agar siap digunakan dalam algoritma klaster, karena algoritma berbasis jarak Euclidean seperti K-Means dan Agglomerative mensyaratkan input berupa fitur numerik yang telah terstandarisasi.

Tahap pertama dalam praproses data adalah pengurangan dimensionalitas melalui seleksi fitur, di mana atribut-atribut yang tidak relevan secara analitis dieliminasi dari *dataset* guna menghindari bias komputasi yang tidak berkorelasi dengan perilaku konsumsi moneter.

Tahap berikutnya adalah standarisasi matriks dan normalisasi parameter untuk menyeragamkan distribusi setiap variabel data. Normalisasi ini sangat penting, mengingat data yang bersih, valid, dan terstandarisasi dengan baik merupakan syarat mutlak agar algoritma tidak mengalami bias dispersi maupun instabilitas pengelompokan akibat perbedaan satuan dan rentang nilai antar fitur [11]. Proses komputasi Z-Transformation

mengubah distribusi setiap titik data variabel x menggunakan formulasi 1.

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

μ adalah mean (rata-rata geometrik) dari fitur di dalam populasi, dan σ adalah standar deviasi dari distribusi fitur tersebut. Operasi aljabar linear ini mentransformasikan semua atribut menjadi unit dimensi yang tak berskala (scale-less) dengan nilai ekspektasi (rata-rata) tepat berada di 0 dan ukuran persebaran (varians) bernilai 1.

D. Algoritma K-Means Clustering

Algoritma K-Means yang diperkenalkan oleh MacQueen dan disempurnakan oleh Lloyd merupakan salah satu algoritma klaster paling populer berkat kesederhanaan dan efisiensinya. Model K-Means ini terbukti sangat efektif dalam membedakan pelanggan berdasarkan frekuensi dan nilai moneter transaksi mereka secara proporsional [12]. Prinsip utama K-Means adalah meminimalkan *Within-Cluster Sum of Squares* (WCSS) atau *inertia*, yaitu total jarak kuadrat setiap titik data terhadap centroid klasternya, yang dirumuskan sebagai berikut (persamaan 2)

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (2)$$

k adalah jumlah klaster, C_i adalah himpunan titik data dalam klaster ke-i, x adalah titik data, dan μ_i adalah centroid klaster ke-i.

Algoritma K-Means bekerja melalui empat langkah iteratif. Pertama, sebanyak $k$centroid diinisialisasi menggunakan metode K-Means++ untuk meminimalkan risiko konvergensi ke solusi lokal yang suboptimal. Kedua, setiap titik data ditetapkan ke klaster dengan centroid terdekat berdasarkan jarak Euclidean yang dirumuskan sebagai berikut (persamaan 3).

$$d(x, \mu_i) = \sqrt{\sum_{j=1}^{n} (x_j - \mu_{ij})^2} \quad (3)$$

$x_j$adalah nilai fitur ke-j dari titik data $x$dan μ_{ij}adalah nilai fitur ke-j dari centroid klaster ke-i. Ketiga, centroid setiap klaster diperbarui menjadi rata-rata dari seluruh titik data yang menjadi anggotanya, sebagaimana dirumuskan berikut (persamaan 4).

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \quad (4)$$

Keempat, langkah kedua dan ketiga diulang hingga penugasan klaster tidak berubah atau perubahan WCSS di bawah ambang batas tertentu. Kompleksitas komputasi K-Means adalah $O(n \cdot k \cdot i)$di mana $n$adalah jumlah data dan $i$adalah jumlah iterasi.

E. Algoritma Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering merupakan pendekatan *bottom-up* dalam membangun hierarki klaster. Algoritma ini dimulai dengan memperlakukan setiap titik data sebagai klaster tersendiri, kemudian secara bertahap menggabungkan dua klaster yang paling berdekatan hingga tersisa satu klaster tunggal atau jumlah klaster yang diinginkan tercapai. Proses penggabungan ini divisualisasikan melalui struktur pohon yang disebut *dendrogram*.

Kriteria penggabungan atau *linkage* menentukan cara pengukuran jarak antar klaster. Pemilihan fungsi *linkage* ini sangat krusial, mengingat struktur klaster yang dihasilkan sangat sensitif terhadap parameter kriteria jarak dan dapat dievaluasi secara dinamis menggunakan platform analitik visual untuk mengamati visualisasi *dendrogram* yang terbentuk [13]. *Single Linkage* menggunakan jarak minimum antara titik-titik dalam dua klaster, dirumuskan pada persamaan 5 sebagai berikut.

$$d_{single}(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \quad (5)$$

Complete Linkage menggunakan jarak maksimum antara titik-titik dalam dua klaster, dirumuskan pada persamaan 6.

$$d_{complete}(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y) \quad (6)$$

Average Linkage menggunakan rata-rata jarak seluruh pasangan titik lintas klaster, dirumuskan pada persamaan 7.

$$d_{average}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y) \quad (7)$$

Ward Linkage meminimalkan peningkatan total varians intra-klaster saat dua klaster digabungkan, dirumuskan pada persamaan 8.

$$d_{ward}(C_i, C_j) = \frac{|C_i||C_j|}{|C_i| + |C_j|} \|\mu_i - \mu_j\| \quad (8)$$

$\mu_i$dan $\mu_j$masing-masing adalah centroid klaster $C_i$dan C_j. Kompleksitas komputasi Agglomerative Clustering adalah $O(n^2 \log n)$, yang lebih berat dibandingkan K-Means namun memberikan fleksibilitas dalam menentukan jumlah klaster setelah proses penggabungan selesai.

F. Metrik Evaluasi

Evaluasi kualitas hasil klasterisasi secara objektif dilakukan menggunakan dua metrik evaluasi internal yang terstandar. Pertama, *Silhouette Score* mengukur seberapa mirip suatu titik data terhadap klasternya sendiri dibandingkan dengan klaster terdekat lainnya, dengan nilai berkisar antara -1 hingga 1, di mana nilai positif yang tinggi menunjukkan pemisahan antar kelompok yang baik. Kedua, *Davies-Bouldin Index* (DBI) mengukur rata-rata kemiripan antara setiap klaster dengan klaster yang paling menyerupainya, di mana nilai yang lebih kecil mengindikasikan kepadatan internal yang tinggi sekaligus keterpisahan antar klaster yang baik [14].

Kualitas pembagian klaster dievaluasi menggunakan persamaan 9.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (9)$$

$a(i)$adalah rata-rata jarak titik data i terhadap seluruh titik dalam klasternya sendiri, dan $b(i)$adalah rata-rata jarak titik data i terhadap seluruh titik dalam klaster terdekat yang bukan klasternya. Nilai *Silhouette Score* berkisar antara -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan kualitas pemisahan yang lebih baik.

Davies-Bouldin Index (DBI) mengukur rata-rata kemiripan antara setiap klaster dan klaster yang paling menyerupainya, dirumuskan dengan persamaan 10 sebagai berikut.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{s_i + s_j}{d(\mu_i, \mu_j)}\right) \quad (10)$$

$s_i$adalah rata-rata jarak titik-titik dalam klaster $i$terhadap centroidnya, dan $d(\mu_i, \mu_j)$adalah jarak antara centroid klaster $i$dan j. Nilai DBI yang lebih kecil mengindikasikan klaster yang lebih kompak dan terpisah dengan baik.

Untuk memastikan konsistensi metodologis antar seluruh konfigurasi yang dibandingkan, nilai Silhouette Score dan Davies-Bouldin Index pada Tabel 1 dihitung menggunakan pustaka scikit-learn pada Python dengan parameter k=5, Euclidean distance, Z-transformation yang identik dengan konfigurasi pada RapidMiner, sehingga kelima metode dievaluasi pada kondisi pra-pemrosesan dan jumlah klaster yang sama persis.

G. *RapidMiner*

RapidMiner adalah perangkat lunak yang berperan sebagai alat pembelajaran untuk keperluan ilmu data mining [15]. Dalam penelitian ini, RapidMiner digunakan karena menyediakan operator bawaan untuk kedua algoritma yang diuji, yakni *Clustering (k-Means)* dan *Agglomerative Clustering*, sehingga eksperimen perbandingan dapat dijalankan dalam *workflow* paralel menggunakan sumber data yang identik. Platform ini juga mendukung operator *Normalize* untuk standardisasi data, *Select Attributes* untuk seleksi fitur, serta visualisasi hasil berupa *Centroid Chart*, *Heat Map*, dan *Dendrogram* secara langsung dalam antarmuka grafis.

III. HASIL DAN PEMBAHASAN

Eksperimen klasterisasi dijalankan terhadap *Synthetic Customer Dataset* yang telah melalui tahap praproses, menghasilkan dataset bersih dengan 500 observasi dan 5 fitur numerik. Kedua algoritma dieksekusi secara paralel pada RapidMiner menggunakan *workflow* yang identik sebagaimana dijelaskan pada bagian sebelumnya.

A. *Hasil K-Means Clustering*

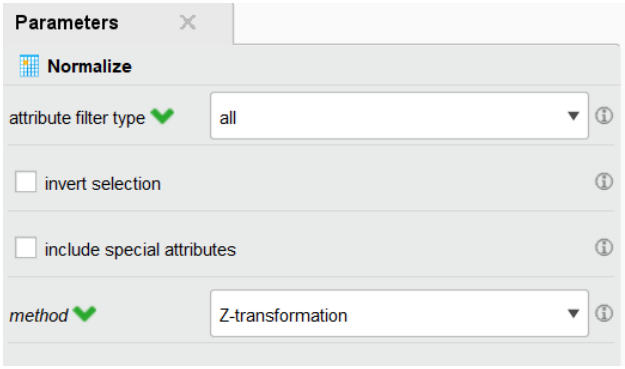
Dari 22 atribut awal, tahap seleksi fitur mengeliminasi 17 atribut yang tidak relevan secara analitis, mencakup atribut identifikasi pribadi (*CustomerID*, *FirstName*, *LastName*, *Email*, *PhoneNumber*, *Address*), atribut lokasi (*City*, *State*, *Country*), atribut kategoris yang tidak dikuantifikasi (*Gender*, *CustomerTier*, *MostFrequentCategory*, *PreferredPaymentMethod*, *DiscountEligibility*, *PreferredContactMethod*), serta atribut temporal mentah (*AccountCreationDate*, *LastPurchaseDate*). Dataset yang tersisa memuat lima fitur numerik yang relevan untuk klasterisasi, yaitu *Age*, *TotalSpend*, *AverageOrderValue*, *TotalOrders*, dan *LoyaltyPoints*.

```
df_clean = df.drop(columns=kolom_drop)

print("Shape setelah drop:", df_clean.shape)
print("\nKolom tersisa:")
print(df_clean.columns.tolist())
```

```
Shape setelah drop: (500, 5)

Kolom tersisa:
['Age', 'TotalSpend', 'AverageOrderValue', 'TotalOrders', 'LoyaltyPoints']
```

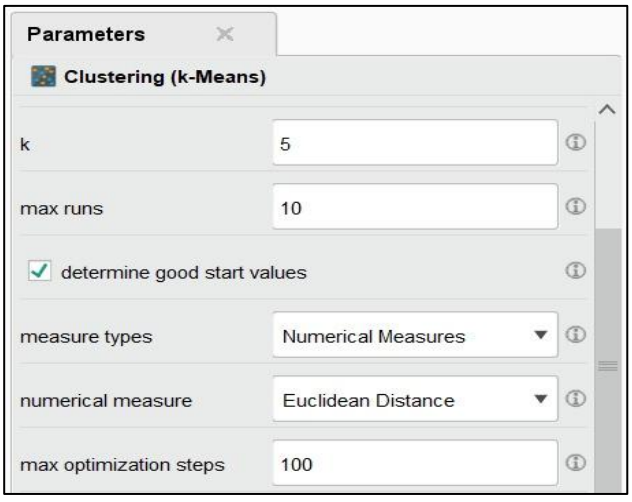
Gambar 1. Penghapusan Atribut

Sebelum normalisasi, rentang nilai antar kelima fitur sangat timpang. *Age* berkisar antara 18 hingga 80, sementara *LoyaltyPoints* berkisar antara 7 hingga 9.992, sehingga tanpa normalisasi fitur dengan rentang besar akan mendominasi perhitungan jarak Euclidean dan menghasilkan klaster yang bias. Normalisasi Z-transformasi kemudian diterapkan melalui operator *Normalize* pada RapidMiner, mentransformasi seluruh fitur sehingga memiliki rata-rata 0 dan standar deviasi 1.



Gambar 2. Konfigurasi parameter operator normalize

B. *Hasil K-Means Clustering*

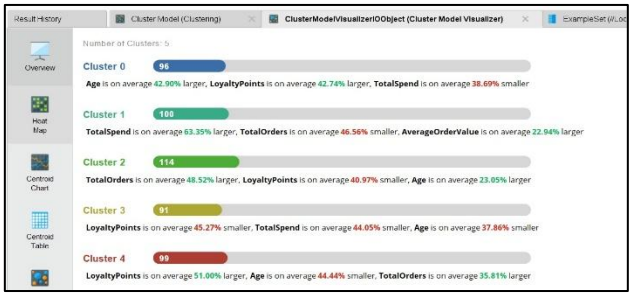


Gambar 3. Konfigurasi parameter operator K-Means

Algoritma K-Means dijalankan dengan parameter k=5, *max runs* 10, dan *measure type Euclidean Distance*. Inisialisasi centroid dilakukan menggunakan metode K-

Means++ untuk meminimalkan risiko konvergensi ke solusi lokal yang suboptimal. Setiap titik data kemudian ditetapkan ke centroid terdekat berdasarkan jarak Euclidean, diikuti pembaruan posisi centroid menjadi rata-rata seluruh anggota klasternya. Proses assignment dan update ini diulang secara iteratif hingga penugasan klaster tidak berubah, yang menandai tercapainya konvergensi dan minimisasi nilai WCSS.

Eksekusi berhasil mempartisi 500 pelanggan ke dalam lima klaster dengan distribusi yang seimbang: Klaster 0 memuat 96 pelanggan, Klaster 1 memuat 100 pelanggan, Klaster 2 memuat 114 pelanggan, Klaster 3 memuat 91 pelanggan, dan Klaster 4 memuat 99 pelanggan. Rasio klaster terbesar terhadap terkecil hanya 1,25:1, mengindikasikan bahwa minimisasi WCSS berjalan efektif dan ruang fitur terpartisi secara proporsional.



Gambar 4. Ringkasan karakteristik untuk setiap cluster

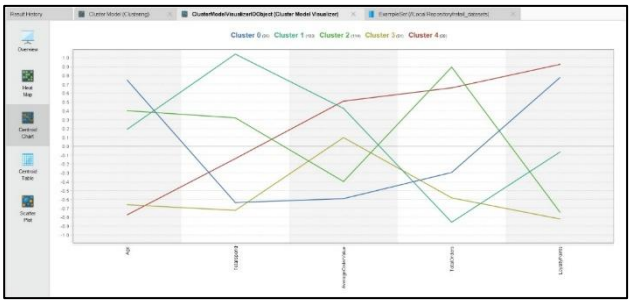
Analisis terhadap nilai centroid masing-masing klaster mengungkap profil perilaku yang distingtif. Klaster 0 dicirikan oleh nilai *Age* yang jauh di atas rata-rata (+42,90%) dan *LoyaltyPoints* sangat tinggi (+42,74%), namun *TotalSpend* justru berada di bawah rata-rata (−38,69%). Pola ini menggambarkan pelanggan senior yang loyal terhadap merek, di mana akumulasi poin yang tinggi berasal dari durasi keanggotaan yang panjang, bukan dari intensitas belanja.

Klaster 1 menampilkan karakteristik yang berlawanan secara diametral. *TotalSpend* pada klaster ini merupakan yang tertinggi di antara seluruh klaster (+63,35%) dengan *AverageOrderValue* yang juga besar, namun *TotalOrders* justru paling rendah (−46,56%). Pola ini menggambarkan segmen pelanggan premium yang jarang bertransaksi namun menghabiskan dana dalam jumlah masif setiap kali berbelanja, atau yang dalam literatur pemasaran disebut sebagai *high-value low-frequency customer*.

Klaster 2 merupakan kelompok dengan frekuensi transaksi tertinggi, ditandai oleh *TotalOrders* yang paling besar di antara semua klaster (+48,52%). Menariknya, meskipun sangat aktif bertransaksi, *LoyaltyPoints* kelompok ini justru rendah (−40,97%), mengindikasikan adanya ketidakselarasan antara aktivitas belanja dan keterlibatan dalam program loyalitas yang menjadi peluang strategis dari sisi retensi pelanggan.

Klaster 3 merepresentasikan segmen dengan performa terendah pada hampir seluruh metrik finansial. *TotalSpend* rendah (−44,05%), *LoyaltyPoints* rendah (−45,27%), dan kelompok ini didominasi oleh pelanggan dengan usia di bawah rata-rata (−37,86%). Profil ini menggambarkan segmen pelanggan pasif berusia muda yang belum menunjukkan keterlibatan signifikan baik dari segi nilai transaksi maupun loyalitas.

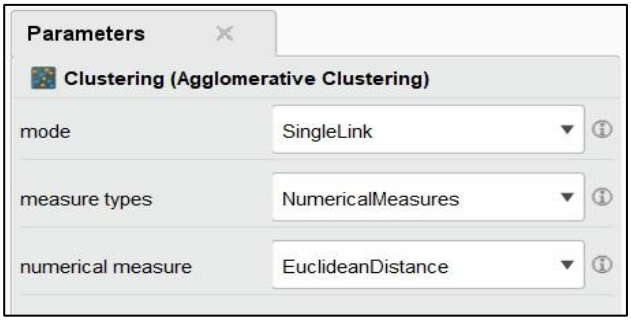
Klaster 4 merupakan kebalikan dari Klaster 0. Kelompok ini memuat pelanggan dengan usia termuda di antara semua klaster (−44,44%), namun menunjukkan *TotalOrders* yang tinggi (+35,81%) dan *LoyaltyPoints* tertinggi di seluruh klaster (+51,00%). Profil ini menggambarkan segmen pelanggan muda yang sangat aktif sekaligus paling loyal, yang merepresentasikan aset strategis jangka panjang dengan potensi *lifetime value* tertinggi bagi pelaku usaha ritel.



Gambar 5. Visualisasi Centroid Chart untuk analisis komparatif antar atribut

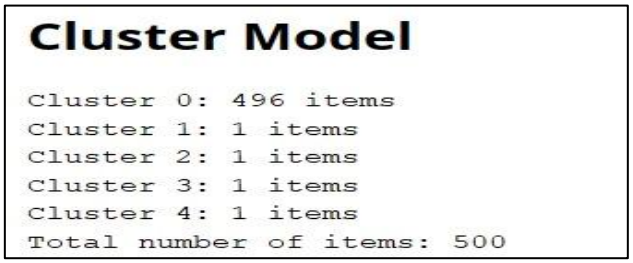
C. *Hasil Implementasi Agglomerative Hierarchical Clustering*

Algoritma Agglomerative Hierarchical Clustering dijalankan dengan konfigurasi *Single Linkage* dan k=5. Sesuai mekanisme *bottom-up* yang dijelaskan pada bagian sebelumnya, setiap dari 500 titik data mula-mula diperlakukan sebagai klaster mandiri, kemudian penggabungan bertahap dilakukan berdasarkan jarak minimum antar titik (*Single Linkage*) hingga jumlah klaster yang diinginkan tercapai.



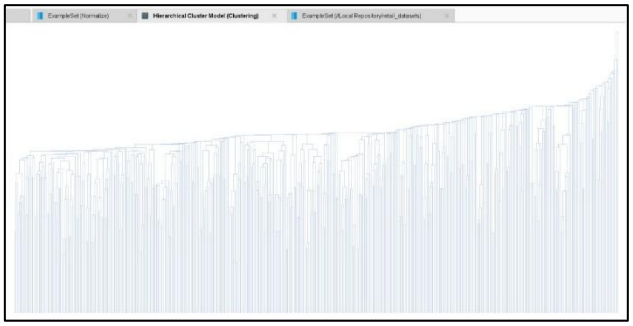
Gambar 6. Konfigurasi parameter operator Agglomerative clustering

Hasil yang diperoleh menunjukkan ketimpangan yang ekstrem. Dari total 500 pelanggan, sebanyak 496 pelanggan (99,2%) terpusat dalam satu klaster tunggal (Klaster 0), sementara Klaster 1, 2, 3, dan 4 masing-masing hanya berisi tepat 1 pelanggan. Ketimpangan ini merupakan manifestasi nyata dari *chaining effect*, yakni kelemahan inheren dari metode *Single Linkage*. Karena *Single Linkage* mengukur jarak antar klaster berdasarkan jarak minimum antar titik, algoritma cenderung menggabungkan data secara berantai satu per satu ke dalam satu rantai klaster raksasa, alih-alih membentuk kelompok yang terpisah secara tegas.



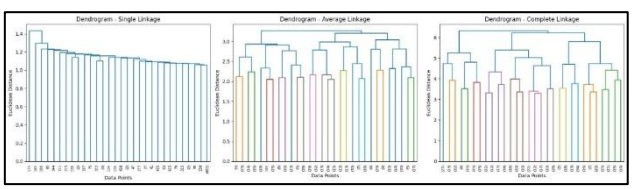
Gambar 7. Ringkasan distribusi jumlah pelanggan

Empat klaster minoritas yang masing-masing hanya berisi satu pelanggan bukanlah representasi segmen yang bermakna secara bisnis. Keempat titik data tersebut merupakan *outlier* yang karakteristik fiturnya sangat jauh dari mayoritas populasi pelanggan, sehingga algoritma *Single Linkage* mengisolasinya sebagai klaster tersendiri sebelum akhirnya bergabung ke dalam rantai utama. Visualisasi dendrogram pada RapidMiner mengonfirmasi fenomena ini, memperlihatkan satu blok hierarki masif di satu sisi struktur pohon yang menunjukkan mayoritas data bergabung pada jarak yang sangat kecil.



Gambar 8. Visualisasi Dendrogram hasil Agglomerative Clustering pada RapidMiner

Sebagai kajian lanjutan, perbandingan visual antar metode *linkage* menunjukkan bahwa pilihan kriteria jarak sangat menentukan bentuk klaster yang dihasilkan. *Complete Linkage* dan *Average Linkage* terbukti mampu mengurangi *chaining effect* secara signifikan. Complete Linkage dan Average Linkage terbukti mampu mengurangi *chaining effect* secara signifikan dibandingkan Single Linkage. Ward Linkage, yang meminimalkan peningkatan total varians intra-klaster, menghasilkan distribusi ukuran klaster yang paling seimbang di antara ketiga varian tersebut. Namun demikian, evaluasi kuantitatif pada Tabel 1 menunjukkan bahwa ketiganya termasuk Ward Linkage masih berada di bawah performa K-Means baik dari sisi Silhouette Score maupun Davies-Bouldin Index.



Gambar 9. Perbandingan bentuk struktur hierarki berdasarkan kriteria Linkage yang berbeda

D. Evaluasi Kuantitatif dengan Metrik Clustering

Kualitas segmentasi kedua algoritma dievaluasi secara kuantitatif menggunakan *Silhouette Score* dan *Davies-Bouldin Index* (DBI) sebagaimana telah didefinisikan pada Bagian II. Hasil evaluasi disajikan pada Tabel 1.

Tabel 1. Perbandingan Nilai Metrik Evaluasi

Metode	Silhouette Score	Davies-Bouldin Index
K-Means (K=5)	0,165	1,592
Single Linkage	-0,123	0,777
Complete Linkage	0,082	1,931
Average Linkage	0,108	1,792
Ward Linkage	0,111	1,927

Silhouette Score K-Means yang lebih tinggi menunjukkan bahwa secara rata-rata setiap titik data lebih dekat ke anggota klasternya sendiri dibandingkan ke klaster terdekat lainnya, kluster konsisten dengan distribusi keanggotaan yang seimbang. Sebaliknya, *Silhouette Score* Agglomerative mendekati nol karena 99,2% data terkompresi dalam satu klaster, menyebabkan nilai a(i) hampir tidak dapat dibedakan dari b(i). Dari sisi DBI, nilai yang lebih kecil pada K-Means mengindikasikan klaster yang lebih kompak dan terpisah dengan baik, sedangkan DBI Agglomerative yang tinggi mencerminkan ketidakmampuan algoritma membentuk kelompok yang kohesif pada konfigurasi *Single Linkage* ini.

Pola menarik teramati pada nilai Davies-Bouldin Index Single Linkage, yang justru paling rendah (0,777) di antara seluruh konfigurasi meskipun mengalami chaining effect paling parah. Hal ini terjadi karena DBI menghitung dispersi internal klaster (si) yang bernilai mendekati nol pada klaster-klaster singleton hasil chaining effect, sehingga nilai DBI keseluruhan tampak rendah meskipun partisi yang dihasilkan tidak bermakna secara bisnis. Silhouette Score yang bernilai negatif (-0,123) justru lebih akurat mencerminkan kegagalan partisi ini, karena metrik tersebut mengevaluasi posisi relatif setiap titik data terhadap klaster tetangga, bukan hanya dispersi rata-rata di dalam klaster. Temuan ini menegaskan pentingnya menggunakan lebih dari satu metrik validasi internal secara bersamaan ketika mengevaluasi algoritma clustering pada data dengan distribusi ukuran klaster yang berpotensi sangat tidak seimbang.

E. Perbandingan Komprehensif Kedua Algoritma

Dari segi mekanisme pembentukan klaster, K-Means bekerja dengan memaksa setiap titik data masuk ke salah satu dari k klaster berdasarkan jarak terhadap centroid terdekat, sehingga menghasilkan klaster yang relatif bulat dan terdistribusi merata. Agglomerative Clustering sebaliknya menggabungkan data secara bertahap berdasarkan jarak antar titik, sehingga hasilnya sangat dipengaruhi oleh metode *linkage* yang dipilih dan tidak dapat direvisi setelah penggabungan terjadi (*non-reversible*).

Dari segi interpretabilitas bisnis, K-Means menawarkan keunggulan yang signifikan. Representasi centroid memungkinkan karakterisasi profil setiap klaster

secara langsung berdasarkan nilai rata-rata fitur, sehingga setiap segmen dapat diberi label yang bermakna dan langsung diterapkan sebagai dasar strategi pemasaran. Pada Agglomerative Clustering, interpretasi memerlukan langkah tambahan dan menjadi tidak relevan ketika distribusi klaster seektrem yang diamati dalam eksperimen ini.

Secara keseluruhan, hasil evaluasi kuantitatif pada Tabel 1 menunjukkan bahwa K-Means konsisten mengungguli seluruh varian Agglomerative Clustering dari sisi Silhouette Score (0,165 berbanding 0,082–0,111), dan mengungguli Complete, Average, serta Ward Linkage dari sisi Davies-Bouldin Index (1,592 berbanding 1,792–1,931). Pengecualian hanya terjadi pada DBI Single Linkage (0,777) yang tampak lebih rendah, namun nilai ini merupakan artefak klaster-klaster *singleton* akibat *chaining effect* dan tidak mencerminkan kualitas partisi yang baik, sebagaimana ditegaskan oleh Silhouette Score-nya yang negatif. Dengan demikian, untuk tujuan segmentasi pelanggan ritel pada *dataset* dengan distribusi data yang padat dan homogen, K-Means terbukti lebih unggul secara konsisten dibandingkan seluruh konfigurasi Agglomerative Clustering yang diuji dalam penelitian ini. Meskipun demikian, Agglomerative Clustering khususnya dengan konfigurasi Complete, Average, atau Ward Linkage tetap memiliki nilai eksplorasi yang tinggi karena kemampuannya mengungkap struktur hierarkis data dan mendeteksi *outlier* secara natural tanpa memerlukan spesifikasi jumlah klaster di awal, meskipun performa kuantitatifnya pada dataset ini belum menyamai K-Means.

IV. KESIMPULAN

Penelitian ini telah mengimplementasikan dan membandingkan dua algoritma klaster, yaitu K-Means Clustering dan Agglomerative Hierarchical Clustering, dalam melakukan segmentasi pelanggan pada *Synthetic Customer Dataset* yang memuat 500 data pelanggan menggunakan platform RapidMiner. Berdasarkan serangkaian proses yang meliputi praproses data, implementasi algoritma, dan analisis komparatif, diperoleh beberapa kesimpulan sebagai berikut.

K-Means Clustering terbukti jauh lebih efektif dan fungsional dalam mempartisi dataset ini dibandingkan Agglomerative Hierarchical Clustering dengan konfigurasi *Single Linkage*. Algoritma K-Means berhasil membagi 500 pelanggan ke dalam lima klaster dengan distribusi keanggotaan yang seimbang dan representatif, berkisar antara 91 hingga 114 pelanggan per klaster. Sebaliknya, Agglomerative Clustering mengalami *chaining effect* yang menyebabkan 496 pelanggan atau 99,2% dari total data tergumpal menjadi satu klaster tunggal, sementara empat klaster lainnya hanya berisi satu data pencilan masing-masing, sehingga gagal menghasilkan segmentasi yang bermakna secara bisnis.

Evaluasi lanjutan terhadap tiga varian *linkage* lain (Complete, Average, dan Ward) menegaskan bahwa K-Means tetap konsisten unggul secara kuantitatif (Silhouette Score 0,165 berbanding 0,082–0,111 dan DBI 1,592 berbanding 1,792–1,931), meskipun ketiga varian tersebut berhasil menghindari *chaining effect* ekstrem yang terjadi pada Single Linkage. Dengan demikian, kesimpulan keunggulan K-Means pada penelitian ini didukung oleh evaluasi terhadap seluruh konfigurasi Agglomerative Clustering yang umum digunakan, bukan hanya terbatas pada satu konfigurasi *linkage* saja.

Melalui K-Means, berhasil diidentifikasi lima profil pelanggan yang distingtif. Klaster 0 menggambarkan pelanggan senior yang loyal namun hemat, dengan usia dan poin loyalitas di atas rata-rata tetapi total belanja yang rendah. Klaster 1 merepresentasikan pelanggan premium atau *high spender* yang jarang bertransaksi namun menghabiskan dana dalam jumlah besar setiap kali melakukan pembelian. Klaster 2 terdiri dari pelanggan aktif dengan frekuensi belanja tertinggi namun poin loyalitas yang masih rendah, mengindikasikan ketidakselarasan antara aktivitas transaksi dan keterlibatan program loyalitas. Klaster 3 menunjukkan segmen pelanggan pasif berusia muda dengan nilai transaksi dan loyalitas terendah di antara seluruh klaster. Klaster 4 mencerminkan pelanggan muda yang paling aktif sekaligus paling loyal, dengan frekuensi belanja tinggi dan poin loyalitas tertinggi, menjadikannya aset strategis jangka panjang yang paling bernilai bagi pelaku usaha ritel.

Temuan komparatif penelitian ini menegaskan bahwa keberhasilan algoritma Agglomerative Hierarchical Clustering sangat bergantung pada pemilihan metode *linkage* yang tepat. Penggunaan *Single Linkage* terbukti tidak sesuai untuk dataset dengan distribusi data yang padat dan homogen, dan disarankan untuk digantikan dengan *Ward Linkage* atau *Complete Linkage* pada penelitian selanjutnya guna meminimalkan *chaining effect* dan menghasilkan segmentasi hierarkis yang lebih bermakna. Selain itu, perluasan dimensi fitur melalui penyertaan atribut kategoris yang dikuantifikasi, seperti metode pembayaran dan tingkat keanggotaan, serta penerapan metrik RFM (*Recency*, *Frequency*, *Monetary*) secara eksplisit, berpotensi menghasilkan profil segmen yang lebih kaya dan informatif pada penelitian mendatang.

DAFTAR PUSTAKA

[1] E. F. L. Awalina and W. I. Rahayu, "Optimalisasi strategi pemasaran dengan segmentasi pelanggan menggunakan penerapan K-Means clustering pada transaksi online retail," *Jurnal Teknologi dan Informasi*, vol. 13, no. 2, pp. 122–137, Aug. 2023.

[2] S. Kumar, R. Rani, S. K. Pippal, and R. Agrawal, "Customer segmentation in e-commerce: K-means vs hierarchical clustering," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 23, no. 1, pp. 119–128, Feb. 2025.

[3] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.

[4] N. A. Hidayati and U. Khasanah, "Shopping pattern segmentation: HAC versus K-Means performance analysis," *Bulletin of Applied Mathematics and Mathematics Education*, vol. 5, no. 2, pp. 97–102, Oct. 2025.

[5] D. Ramadhani, Q. A'yuniyah, W. Elvira, N. Nazira, I. Ambarani, and S. F. Intan, "Analisa algoritma Naïve Bayes Classifier (NBC) untuk prediksi penjualan alat kesehatan," *Indonesian Journal of Informatic Research and Software Engineering (IJIRSE)*, vol. 3, no. 2, pp. 119–126, Sep. 2023.

[6] Y. N. F. D. Arifin and A. Zubair, "Segmentasi pelanggan e-commerce berdasarkan pola pembatalan dan pengembalian pesanan menggunakan K-Means," *JUSIFOR: Jurnal Sistem Informasi dan Informatika*, vol. 4, no. 2, pp. 192–203, Dec. 2025.
[7] J. K. Sihombing and B. A. Wijaya, "Implementasi algoritma clustering dan classification dalam data mining: Systematic literature review terhadap tren dan tantangan terkini," *Jurnal Publikasi Sistem Informasi dan Manajemen Bisnis*, vol. 4, no. 3, pp. 372–386, Sep. 2025.
[8] B. Apriyanto and S. L. M. Sitio, "Penerapan K-Means dalam menganalisis pola pembelian pelanggan pada data transaksi e-commerce," *bit-Tech*, vol. 7, no. 3, pp. 790–797, Apr. 2025.
[9] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, NJ, USA: Wiley & Sons, 2005.
[10] B. Ilham, Rahmaddeni, A. Putra, D. Najario, and M. A. Fakhrizal, "Segmentasi pelanggan: Optimizing K-Means performance with hybrid RFM and behavioral analytics for customer segmentation," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 4, Oct. 2025.
[11] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2011.
[12] A. Saputra and R. Yusuf, "Perbandingan algoritma DBSCAN dan K-MEANS dalam segmentasi pelanggan pengguna transportasi publik Transjakarta menggunakan metode RFM," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 4, Aug. 2024.
[13] Y. I. Muasaroh and Z. Fatah, "Implementasi RapidMiner dalam optimasi pembentukan kelas unggulan menggunakan K-Means clustering," *Journal of Advanced Research in Informatics*, vol. 3, no. 1, 2024.
[14] P. Septiana Batirsi, "KMeans clustering menggunakan RapidMiner dalam segmentasi pelanggan dengan evaluasi Davies Bouldin Index untuk menentukan jumlah cluster paling optimal," *Bahari Teknik Informatika dan Sistem Informasi*, vol. 6, no. 2, pp. 8–13, 2023
[15] U. Lathifah and R. D. Dana, "Implementasi metode linear regression untuk prediksi harga properti real estate menggunakan RapidMiner," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 1, pp. 1129–1137, Feb. 2024.