

Analisis Perbandingan Algoritma K-Means, DBSCAN, dan Hierarchical Clustering untuk Segmentasi Pelanggan E-Commerce Berbasis RFM

Ahmad Farhan Ramadhansyah^{1*}, Wa Ode Cantika Nurhajratul Uswah², Rezky Aulia³,
Adha Mashur Sajiah⁴

^{1,2,3,4}Fakultas Teknik, Program Studi Teknik Informatika, Universitas Halu Oleo, Kota Kendari, Indonesia

Email: ^{1*}ramadhansyahfarhan96@gmail.com, ²waodecantikanurhajratuluswah12@gmail.com,

³Rezkyaulia.1275@gmail.com, ⁴adha.m.sajiah@uho.ac.id

(*Email Corresponding Author: ramadhansyahfarhan96@gmail.com)

Received: 29 Juni 2026 | Revision: 20 Agustus 2026 | Accepted: 30 September 2026

Abstrak

Pertumbuhan e-commerce di Indonesia telah mendorong akumulasi data transaksi dalam jumlah yang sangat besar, yang apabila dikelola dengan tepat dapat menjadi aset strategis bagi perusahaan dalam memahami perilaku pelanggannya. Salah satu pendekatan yang efektif adalah segmentasi pelanggan berbasis model Recency, Frequency, dan Monetary (RFM). Penelitian ini membandingkan tiga algoritma clustering, yaitu K-Means, DBSCAN, dan Hierarchical Clustering, dalam segmentasi pelanggan e-commerce berbasis RFM. Data yang digunakan adalah Online Retail Dataset dari Kaggle yang memuat 541.909 transaksi periode Desember 2010 hingga Desember 2011. Setelah prapemrosesan, sebanyak 397.884 baris data bersih diintegrasikan menjadi profil RFM untuk 4.338 pelanggan. Data ditransformasi menggunakan log1p, winsorizing 3×IQR, dan standarisasi StandardScaler sebelum proses clustering. Kualitas cluster dievaluasi menggunakan tiga metrik internal, yaitu Silhouette Score (SS), Davies-Bouldin Index (DBI), dan Calinski-Harabasz Index (CHI). Hasil evaluasi menunjukkan K-Means dengan K=3 unggul pada ketiga metrik (SS=0,337, DBI=1,048, CHI=3.632), diikuti Hierarchical Clustering (SS=0,320, DBI=1,105, CHI=3.275), dan DBSCAN dengan skor terendah (SS=0,294, CHI=2.395) meski tetap unggul dalam mendeteksi pelanggan VIP sebagai outlier. K-Means menghasilkan tiga segmen yang mudah diinterpretasikan, yaitu Champions (17,77%), Regular Customers (39,07%), dan Lost Customers (43,15%), menjadikannya algoritma paling optimal untuk segmentasi pelanggan e-commerce berbasis RFM.

Kata Kunci: Segmentasi Pelanggan, RFM, K-Means, DBSCAN, Hierarchical Clustering

Abstract

The growth of e-commerce in Indonesia has driven the accumulation of a huge amount of transaction data, which, if managed properly, can be a strategic asset for companies in understanding their customer behavior. One effective approach is customer segmentation based on the Recency, Frequency, and Monetary (RFM) model. This study compares three clustering algorithms, namely K-Means, DBSCAN, and Hierarchical Clustering, in RFM-based e-commerce customer segmentation. The data used is the Online Retail Dataset from Kaggle containing 541,909 transactions from December 2010 to December 2011. After preprocessing, a total of 397,884 rows of clean data were aggregated into RFM profiles for 4,338 customers. The data were transformed using log1p, winsorizing 3×IQR, and standardization StandardScaler before the clustering process. Cluster quality was evaluated using three internal metrics, namely the Silhouette Score (SS), Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). The evaluation results show that K-Means with K=3 excels in all three metrics (SS=0.337, DBI=1.048, CHI=3.632), followed by Hierarchical Clustering (SS=0.320, DBI=1.105, CHI=3.275), and DBSCAN with the lowest score (SS=0.294, CHI=2.395) although it still excels in detecting VIP customers as outliers. K-Means produces three easily interpretable segments, namely Champions (17.77%), Regular Customers (39.07%), and Lost Customers (43.15%), making it the most optimal algorithm for RFM-based e-commerce customer segmentation.

Keywords: Customer Segmentation, RFM Analysis, K-Means Clustering, DBSCAN, Hierarchical Clustering

1. PENDAHULUAN

Perkembangan teknologi digital telah mendorong pertumbuhan e-commerce di Indonesia secara masif dalam satu dekade terakhir. Platform perdagangan elektronik kini telah menjadi salah satu sektor ekonomi digital yang paling dominan, dengan jutaan transaksi terjadi setiap harinya dan melibatkan jutaan pelanggan dari berbagai segmen demografi. Kondisi ini semakin diperkuat oleh meningkatnya penetrasi internet, tingginya penggunaan perangkat *mobile*, serta perubahan perilaku konsumen yang semakin terbiasa berbelanja secara daring. Pertumbuhan yang pesat ini secara langsung menghasilkan akumulasi data transaksi dalam skala yang sangat besar, yang apabila dikelola dengan baik dapat menjadi aset strategis bagi perusahaan dalam memahami perilaku pelanggan dan merancang strategi bisnis yang lebih kompetitif [1].

Meskipun demikian, besarnya volume data transaksi juga menciptakan tantangan tersendiri bagi perusahaan e-commerce dalam memahami karakteristik setiap pelanggan secara individual. Pelanggan memiliki pola pembelian, frekuensi transaksi, dan nilai belanja yang sangat beragam, sehingga pendekatan pemasaran yang seragam tidak lagi efektif. Oleh

karena itu, segmentasi pelanggan menjadi strategi yang sangat penting dalam pengembangan bisnis yang berorientasi pada pelanggan. Segmentasi pelanggan adalah proses membagi pelanggan ke dalam beberapa kelompok berdasarkan kesamaan karakteristik atau perilaku, sehingga perusahaan dapat memberikan penawaran yang lebih personal, meningkatkan efisiensi promosi, serta memperkuat loyalitas pelanggan jangka panjang [2].

Salah satu pendekatan yang banyak digunakan dalam segmentasi pelanggan adalah model *Recency, Frequency, dan Monetary* (RFM). Model ini mampu mengukur perilaku pelanggan berdasarkan data historis transaksi. *Recency* mengukur seberapa baru pelanggan melakukan transaksi terakhir, *Frequency* mengukur seberapa sering pelanggan bertransaksi dalam periode tertentu, dan *Monetary* mengukur total nilai uang yang dikeluarkan pelanggan. Ketiga dimensi ini secara bersama-sama mampu menggambarkan profil pelanggan secara komprehensif, mulai dari pelanggan loyal yang aktif hingga pelanggan yang berpotensi beralih ke kompetitor, sehingga RFM menjadi fondasi yang ideal bagi proses segmentasi berbasis data [3].

Untuk mengolah atribut RFM menjadi segmen yang bermakna, digunakan teknik *clustering* sebagai metode *unsupervised learning*. Tiga algoritma yang paling umum diterapkan adalah K-Means yang mempartisi data berdasarkan kedekatan terhadap centroid [4], DBSCAN yang mengelompokkan data berdasarkan kepadatan dan mampu mendeteksi outlier secara otomatis tanpa perlu menentukan jumlah kluster di awal [5], serta *Hierarchical Clustering* yang membangun hierarki kluster secara bertahap dan menghasilkan dendrogram sebagai representasi visual pengelompokan [6]. Perbedaan karakteristik tersebut menjadikan ketiga algoritma memiliki keunggulan dan keterbatasan masing-masing sesuai dengan sifat data yang dianalisis.

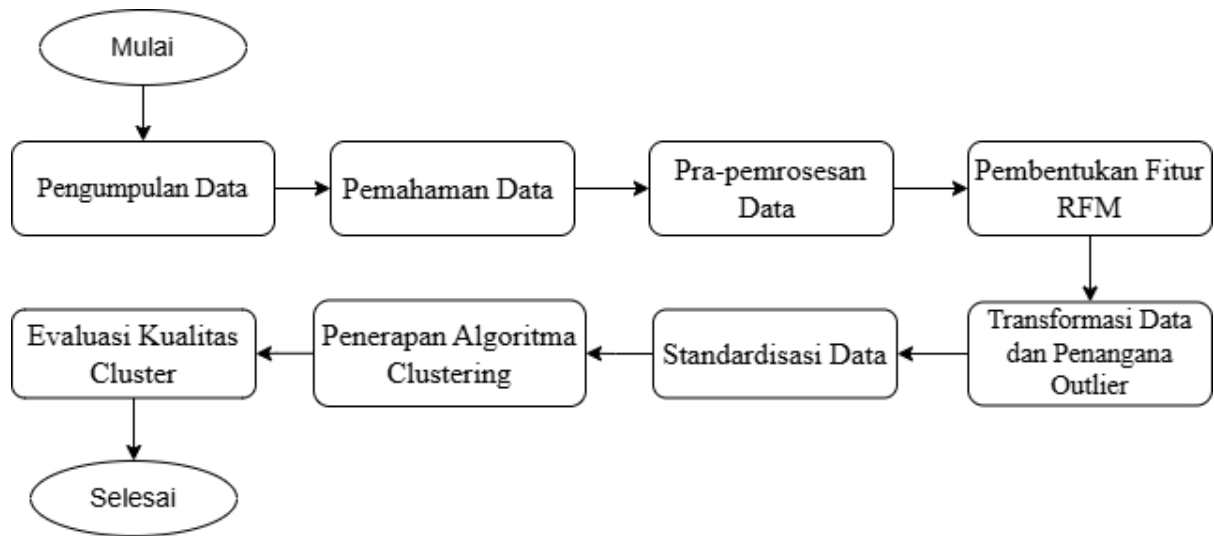
Beberapa penelitian sebelumnya telah menerapkan algoritma *clustering* berbasis RFM untuk segmentasi pelanggan. Fadhillah et al. (2025) membandingkan K-Means, *Agglomerative*, dan DBSCAN pada data perusahaan distribusi B2B dan menemukan K-Means sebagai algoritma terbaik, namun terbatas pada domain *non-e-commerce* [7]. Nugaha et al. (2025) menerapkan K-Means berbasis RFM pada perusahaan *payment gateway* dengan hasil DBI yang baik, tetapi hanya menggunakan satu algoritma tanpa perbandingan [8]. Saputra dan Yusuf (2024) membandingkan K-Means dan DBSCAN berbasis RFM, namun tidak menyertakan *Hierarchical Clustering* sehingga perbandingannya belum menyeluruh [9]. Wardani et al. (2023) mengkomparasi ketiga algoritma pada data penjualan umum tanpa integrasi model RFM [10], sedangkan Samuel (2025) melakukan kajian komparatif berbasis RFM pada skala internasional, namun belum mengevaluasi secara spesifik pada konteks *e-commerce* dengan metrik validasi yang lengkap [11]. Dari tinjauan tersebut tampak bahwa studi yang secara komprehensif membandingkan K-Means, DBSCAN, dan *Hierarchical Clustering* berbasis RFM pada domain *e-commerce* dengan evaluasi multi-metrik yaitu *Silhouette Score*, *Davies-Bouldin Index*, dan *Calinski-Harabasz Index* secara bersamaan masih sangat terbatas [12].

Oleh karena itu, penelitian ini bertujuan untuk menganalisis dan membandingkan performa algoritma K-Means, DBSCAN, dan *Hierarchical Clustering* dalam segmentasi pelanggan *e-commerce* berbasis RFM menggunakan tiga metrik validasi secara bersamaan. Hasil penelitian ini diharapkan dapat memberikan rekomendasi algoritma *clustering* yang paling optimal sebagai rujukan ilmiah bagi peneliti selanjutnya, sekaligus menjadi panduan praktis bagi perusahaan *e-commerce* dalam merancang strategi pemasaran berbasis data yang lebih efektif dan tepat sasaran.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif berbasis eksperimen komputasional, dengan tujuan membandingkan kinerja tiga algoritma *clustering*, yaitu K-Means, DBSCAN, dan *Hierarchical Clustering*, dalam melakukan segmentasi pelanggan menggunakan fitur RFM (*Recency, Frequency, Monetary*). Alur penelitian secara keseluruhan mencakup delapan tahapan sistematis, mulai dari pengumpulan data hingga evaluasi kualitas cluster, sebagaimana diilustrasikan pada Gambar 1. Penelitian ini dirancang untuk menghasilkan proses segmentasi pelanggan yang sistematis dan dapat direplikasi pada penelitian sejenis. Setiap tahapan dalam alur penelitian saling berkaitan, mulai dari pengolahan data transaksi pelanggan, pembentukan fitur RFM, hingga penerapan algoritma *clustering*. Pendekatan yang terstruktur diperlukan untuk memastikan bahwa hasil segmentasi yang diperoleh mampu merepresentasikan karakteristik pelanggan secara akurat dan konsisten.

Selain menghasilkan segmentasi pelanggan, penelitian ini juga bertujuan mengevaluasi efektivitas masing-masing algoritma dalam membentuk cluster yang berkualitas. Hasil evaluasi tersebut digunakan untuk mengidentifikasi algoritma yang paling sesuai dalam mengelompokkan pelanggan berdasarkan kemiripan perilaku transaksi. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi baik dari sisi akademis maupun praktis dalam penerapan teknik data mining untuk mendukung pengambilan keputusan bisnis berbasis data.



Gambar 1. Alur Penelitian

Berdasarkan Gambar 1, penjelasan alur pada setiap tahapan dijabarkan sebagai berikut.

2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah *Online Retail Dataset*, yaitu kumpulan data transaksi pelanggan dari sebuah perusahaan *retail* berbasis di Inggris yang bergerak di bidang penjualan produk hadiah dan kebutuhan rumah tangga secara daring yang diperoleh melalui platform Kaggle. Dataset tersebut mencakup transaksi yang berlangsung pada periode 1 Desember 2010 hingga 9 Desember 2011 dengan total data awal yang digunakan sebanyak 541.909 transaksi dengan 8 kolom atribut[6]. Pemilihan dataset ini didasarkan pada kelengkapan atribut transaksional yang mendukung pembentukan fitur RFM, serta relevansinya dalam konteks penelitian segmentasi pelanggan *e-commerce*[9]. Rincian atribut dataset disajikan pada Tabel 1 berikut.

Tabel 1. Deskripsi Atribut Dataset

No	Nama Atribut	Deskripsi
1	InvoiceNo	Nomor unik identifikasi setiap transaksi. Kode yang diawali huruf 'C' menandakan transaksi pembatalan (<i>cancellation</i>)
2	StockCode	Kode unik untuk setiap produk yang dijual
3	Description	Nama atau deskripsi produk yang dibeli
4	Quantity	Jumlah unit produk yang dibeli dalam satu baris transaksi. Nilai negatif menandakan transaksi retur
5	InvoiceDate	Tanggal dan waktu terjadinya transaksi
6	UnitPrice	Harga satuan produk dalam mata uang Pound Sterling (£)
7	CustomerID	Nomor unik identifikasi pelanggan. Bernilai kosong (<i>null</i>) apabila transaksi dilakukan tanpa login akun
8	Country	Nama negara tempat pelanggan berdomisili

Dataset memiliki 6 atribut utama berupa *InvoiceNo*, *CustomerID*, *Quantity* (Jumlah Produk), *InvoiceDate* (Tanggal Transaksi), *UnitPrice* (Harga Produk), dan *Country* (Negara). Keenam atribut tersebut secara kolektif menyediakan informasi yang diperlukan untuk menghitung nilai *Recency*, *Frequency*, dan *Monetary* pada setiap pelanggan, sehingga menjadi fondasi utama dalam proses segmentasi [13].

2.2 Pemahaman Data

Tahap pemahaman data bertujuan untuk mengevaluasi kondisi awal dataset sebelum diproses lebih lanjut. Pada tahap ini dilakukan beberapa pemeriksaan awal untuk memperoleh gambaran menyeluruh mengenai struktur dan kualitas data [14]. Kegiatan yang dilakukan mencakup.

- Pemeriksaan jumlah total data dan atribut yang tersedia dalam dataset
- Identifikasi keberadaan *missing value* pada setiap kolom, khususnya pada atribut *CustomerID* yang bersifat krusial dalam proses pengelompokan pelanggan.
- Pemeriksaan jumlah pelanggan unik berdasarkan *CustomerID* untuk memahami cakupan populasi data.
- Analisis rentang waktu transaksi guna memastikan konsistensi periode data yang digunakan.
- Deteksi transaksi tidak valid, yang meliputi data dengan nilai $Quantity \leq 0$ (mengindikasikan retur atau pembatalan pesanan) dan $UnitPrice \leq 0$ (mengindikasikan harga yang tidak wajar atau tidak valid) [15].

Hasil dari tahap ini menjadi acuan dalam menentukan langkah-langkah pembersihan dan transformasi data yang diperlukan pada tahap selanjutnya [14].

2.3 Prapemrosesan Data

Prapemrosesan data merupakan tahap yang sangat menentukan kualitas hasil analisis *clustering* [14]. Pada tahap ini dilakukan serangkaian proses pembersihan dan transformasi data, yang secara garis besar meliputi empat sub-proses yaitu, penghapusan *Missing Value*, penghapusan data tidak valid, pembuatan variabel total belanja, dan konversi format data [9][15][13].

2.4 Pembentukan dan Perhitungan Fitur RFM

Model RFM (*Recency, Frequency, Monetary*) merupakan kerangka analitik yang digunakan secara luas dalam segmentasi pelanggan berbasis perilaku transaksi [16]. Setelah proses pembersihan data selesai dilakukan, data transaksi digregasikan berdasarkan pelanggan (*CustomerID*) untuk menghasilkan tiga fitur utama, yaitu *Recency*, *Frequency*, dan *Monetary*. Ketiga fitur tersebut kemudian digabungkan ke dalam sebuah tabel RFM yang merepresentasikan profil perilaku masing-masing pelanggan. Tabel RFM ini menjadi masukan utama dalam proses segmentasi pelanggan menggunakan algoritma K-Means, DBSCAN, dan Hierarchical *Clustering* [4].

Persamaannya:

$$RFM = \{R, F, M\} \quad (1)$$

Dimana setiap pelanggan direpresentasikan sebagai sebuah vektor yang terdiri atas nilai *Recency*, *Frequency*, dan *Monetary*. Vektor tersebut selanjutnya digunakan sebagai dasar pembentukan cluster pelanggan berdasarkan kesamaan karakteristik transaksi.

2.4.1 Recency (R)

Recency digunakan untuk mengukur selisih waktu (dalam satuan hari) antara tanggal transaksi terakhir pelanggan dan tanggal referensi yang ditentukan. Nilai *Recency* yang lebih kecil menunjukkan bahwa pelanggan melakukan transaksi yang lebih baru [16].

Rumus *Recency* dapat dituliskan sebagai berikut:

$$R = D_{snapshot} - D_{last} \quad (2)$$

Keterangan:

- R = Nilai *Recency* pelanggan
- $D_{snapshot}$ = Tanggal referensi (biasanya tanggal analisis atau tanggal terakhir dalam dataset)
- D_{last} = Tanggal transaksi terakhir yang dilakukan oleh pelanggan

2.4.2 Frequency (F)

Frequency digunakan untuk menghitung jumlah total transaksi unik (*Invoice*) yang dilakukan oleh masing-masing pelanggan selama periode pengamatan. Nilai *Frequency* yang tinggi mengindikasikan tingkat keterlibatan pelanggan yang lebih aktif [16].

Rumus *Frequency* dapat dituliskan sebagai berikut:

$$F = \sum_{j=1}^n Invoice_{ij} \quad (3)$$

Atau dapat dituliskan secara sederhana sebagai:

$$F = \text{Count}(\text{Invoice Unique}) \quad (4)$$

Keterangan:

- F = Nilai *Frequency* pelanggan
- i = Indeks transaksi ke-1, ke-2, ke-3, ..., ke- n
- n = Jumlah total transaksi yang dilakukan pelanggan
- j = Jumlah transaksi pelanggan ke- i
- Invoice_{ij} = *Invoice* ke- j milik pelanggan ke- i

2.4.3 Monetary (M)

Monetary digunakan untuk menjumlahkan total pengeluaran seluruh transaksi masing-masing pelanggan. Nilai *Monetary* mencerminkan kontribusi finansial pelanggan terhadap perusahaan [16].

Rumus *Monetary* dapat dituliskan sebagai berikut:

$$M = \sum_{i=1}^n \text{Total}_i \quad (5)$$

Atau jika menggunakan harga dan kuantitas:

$$M = \sum_{i=1}^n (\text{Quantity}_i \times \text{UnitPrice}_i) \quad (6)$$

Keterangan:

- i = Indeks transaksi ke-1, ke-2, ke-3, ..., ke- n
- n = Jumlah total transaksi yang dilakukan pelanggan
- Quantity_i = Jumlah item pada transaksi ke- i
- UnitPrice_i = Harga satuan item pada transaksi ke- i
- TotalAmount_i = Total harga pada transaksi ke- i

2.5 Transformasi Data dan Penanganan Outlier

Data RFM umumnya memiliki distribusi yang sangat condong ke kanan (*right-skewed*) akibat adanya variasi perilaku pembelian pelanggan yang signifikan. Keberadaan *outlier* dapat memengaruhi hasil *clustering* secara substansial, sehingga perlu ditangani terlebih dahulu sebelum data dimasukkan ke dalam algoritma [17]. Transformasi logaritmik diterapkan pada setiap fitur RFM untuk mereduksi kemiringan distribusi data dan mendekatkan distribusi ke bentuk yang lebih simetris. Transformasi ini efektif dalam menekan dominasi nilai ekstrem tanpa menghilangkan informasi yang dikandung oleh data tersebut [17]. Teknik *winsorizing* diterapkan untuk menangani *outlier* yang masih tersisa pasca-transformasi. Metode ini bekerja dengan mengganti nilai yang melampaui batas persentil tertentu dengan nilai batas persentil yang ditetapkan, sehingga distribusi data menjadi lebih terkendali tanpa menghapus observasi secara permanen [18]. Penggunaan kombinasi transformasi logaritmik dan *winsorizing* terbukti efektif dalam mempersiapkan data RFM yang bersifat *skewed* sebelum memasuki tahap *clustering* [17].

2.6 Standarisasi Data

Setelah proses transformasi, seluruh fitur RFM distandarisasi menggunakan metode *StandardScaler*. Metode ini mengubah setiap fitur sehingga memiliki rata-rata (*mean*) bernilai nol dan simpangan baku (*standard deviation*) bernilai satu [19]. Standarisasi diperlukan karena algoritma *clustering* berbasis jarak, seperti K-Means dan Hierarchical Clustering, sangat sensitif terhadap perbedaan skala antar fitur; tanpa standarisasi, fitur dengan rentang nilai yang lebih besar akan mendominasi perhitungan jarak dan menghasilkan cluster yang bias [20].

2.7 Penerapan Algoritma Clustering

Pada tahap ini, tiga algoritma *clustering* diterapkan secara terpisah pada data RFM yang telah distandarisasi. Setiap algoritma memiliki pendekatan dan karakteristik yang berbeda dalam membentuk kelompok data.

2.7.1 K-Means Clustering

K-Means merupakan algoritma clustering berbasis partisi yang bekerja dengan cara mengelompokkan data ke dalam sejumlah k cluster berdasarkan kedekatan jarak Euclidean terhadap centroid masing-masing cluster [21]. Penentuan jumlah cluster optimal dilakukan menggunakan metode Elbow yang mengamati penurunan nilai *inertia* [13]. Algoritma ini dipilih karena memiliki kompleksitas komputasi yang relatif rendah dan telah banyak digunakan dalam studi segmentasi pelanggan berbasis RFM [4]. Adapun prosedur algoritma K-Means dalam penelitian ini dijalankan dengan langkah-langkah berikut:

- a. *Input*: Dataset RFM yang telah distandarisasi.
- b. Tentukan jumlah cluster optimal (K) menggunakan Metode *Elbow* dan evaluasi *Silhouette Score*.
- c. Inisialisasi posisi awal K centroid secara cerdas menggunakan parameter *k-means++*.
- d. Perulangan (Iterasi) hingga posisi centroid konvergen (tidak berubah)
 1. Hitung jarak *Euclidean* dari setiap titik data RFM pelanggan ke semua centroid.
 2. Alokasikan setiap pelanggan ke cluster dengan centroid terdekat.
 3. Perbarui posisi centroid dengan menghitung nilai rata-rata (*mean*) dari seluruh titik data pada masing-masing cluster.
- e. *Output*: Himpunan K cluster pelanggan.

2.7.2 DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)

DBSCAN adalah algoritma clustering berbasis densitas yang mengelompokkan titik-titik data berdasarkan kepadatan distribusinya di ruang fitur. Algoritma ini memiliki kemampuan untuk mendeteksi cluster dengan bentuk arbitrer serta mengidentifikasi *noise* (outlier) yang tidak termasuk dalam cluster mana pun [15]. Penentuan parameter utama DBSCAN, yaitu epsilon (ϵ) dan *minimum samples*, dilakukan melalui analisis *k-distance graph* untuk memperoleh nilai yang optimal. Adapun prosedur algoritma DBSCAN dalam penelitian ini dijalankan dengan langkah-langkah berikut:

- a. *Input*: Dataset RFM yang telah distandarisasi, nilai epsilon (ϵ), dan nilai minimum sampel (*min_samples*).
- b. Tandai semua titik data pelanggan sebagai *unvisited* (belum dikunjungi).
- c. Untuk setiap titik data P dalam dataset:
 1. Jika P sudah dikunjungi, lanjutkan ke titik berikutnya.
 2. Tandai P sebagai *visited*.
 3. Cari semua titik data tetangga dalam radius ϵ dari P .
 4. Jika jumlah tetangga $< \text{min_samples}$, tandai P sebagai *Noise/Outlier* (label -1).
 5. Jika jumlah tetangga $\geq \text{min_samples}$, buat *Cluster* baru C , masukkan P ke C , dan perluas cluster dengan memeriksa tetangga-tetangganya yang memenuhi syarat kepadatan.
- d. *Output*: Himpunan cluster berbasis kepadatan dan kelompok pelanggan *noise* (Outlier/VIP).

2.7.3 Hierarchical Clustering

Hierarchical Clustering membentuk struktur hierarki cluster secara bertahap melalui pendekatan *agglomerative*, yaitu dengan menggabungkan pasangan cluster yang paling mirip secara berurutan dari bawah ke atas. Hasil pengelompokan divisualisasikan dalam bentuk dendrogram untuk membantu menentukan jumlah cluster yang paling representatif. Pendekatan ini tidak memerlukan penentuan jumlah cluster sejak awal, sehingga memberikan fleksibilitas dalam eksplorasi struktur data [22]. Adapun prosedur algoritma Hierarchical Clustering dalam penelitian ini dijalankan dengan langkah-langkah berikut:

- a. *Input*: Dataset RFM yang telah distandarisasi.
- b. Inisialisasi: Anggap setiap titik data pelanggan (N) sebagai satu cluster tunggal (sehingga terdapat N cluster awal).
- c. Hitung matriks jarak (*Euclidean distance*) antar semua cluster.
- d. Perulangan (Iterasi) hingga hanya tersisa 1 cluster utama yang menampung seluruh data:
 1. Temukan dua cluster dengan jarak terdekat menggunakan metode *Ward linkage* (meminimalkan variansi internal).
 2. Gabungkan kedua cluster tersebut menjadi satu cluster baru.
 3. Perbarui matriks jarak antar cluster.
- e. Visualisasikan proses penggabungan dalam bentuk *Dendrogram*.
- f. Lakukan pemotongan (thresholding) pada jarak tertentu di dendrogram untuk menentukan jumlah cluster akhir (K).
- g. *Output*: Himpunan K kelompok pelanggan hierarkis.

2.8 Evaluasi Kualitas Cluster

Kualitas hasil clustering dari ketiga algoritma dievaluasi secara kuantitatif menggunakan tiga metrik *internal evaluation* yang tidak memerlukan label data aktual (*ground truth*) [23]. Penggunaan lebih dari satu metrik evaluasi dilakukan untuk memperoleh penilaian yang lebih komprehensif terhadap kualitas cluster yang dihasilkan [24].

- Silhouette Score*: mengukur seberapa mirip sebuah objek dengan cluster-nya sendiri dibandingkan dengan cluster lain yang berdekatan. Nilai *Silhouette Score* berkisar antara -1 hingga 1, di mana nilai yang mendekati 1 menandakan pemisahan antar cluster yang baik [23].
- Davies-Bouldin Index (DBI)*: mengukur rasio rata-rata kemiripan antar cluster terhadap dispersi internal masing-masing cluster. Nilai DBI yang lebih rendah mengindikasikan hasil clustering yang lebih baik karena mencerminkan cluster yang lebih kompak dan terpisah dengan jelas [24].
- Calinski-Harabasz Index (CHI)*: mengukur rasio antara variansi antar cluster (*between-cluster dispersion*) dan variansi di dalam cluster (*within-cluster dispersion*). Nilai CHI yang lebih tinggi mengindikasikan cluster yang lebih padat dan lebih terpisah satu sama lain [24].

Hasil dari ketiga metrik evaluasi tersebut digunakan sebagai dasar perbandingan objektif antara *K-Means*, *DBSCAN*, dan Hierarchical Clustering, sehingga dapat ditentukan algoritma mana yang paling sesuai untuk segmentasi pelanggan e-commerce berbasis fitur RFM pada dataset yang digunakan dalam penelitian ini [24].

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pemahaman Data

Pemahaman data dilakukan melalui lima tahap pemeriksaan. Pertama, pemeriksaan struktur dataset menunjukkan bahwa data terdiri atas 541.909 baris transaksi dengan delapan atribut, yaitu InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, dan Country.

```
Jumlah Baris Kosong (Missing Values) per Kolom:
InvoiceNo      0
StockCode      0
Description    1454
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID    135080
Country        0
dtype: int64

Total Baris Transaksi : 541909
Total Pelanggan Unik : 4372
Transaksi Tanpa CustomerID: 135080 baris (24.93%)
```

Gambar 2. Identifikasi Missing Value dan Pemeriksaan Pelanggan Unik

Kedua, Gambar 2 mengidentifikasi missing value yang mengungkap bahwa kolom CustomerID memiliki 135.080 nilai kosong (24,93% dari total data), sedangkan kolom Description memiliki 1.454 nilai kosong. Ketidadaan CustomerID pada sebagian besar transaksi menjadi perhatian utama karena atribut ini bersifat krusial dalam proses pengelompokan pelanggan. Ketiga, pemeriksaan pelanggan unik berdasarkan CustomerID menghasilkan 4.372 pelanggan yang teridentifikasi, yang menjadi populasi dasar analisis segmentasi.

```
Rentang Waktu Transaksi:
Pesanan Pertama : 2010-12-01 08:26:00
Pesanan Terakhir: 2011-12-09 12:50:00

Jumlah baris dengan Quantity <= 0 (Barang Retur/Batal) : 10624
Jumlah baris dengan UnitPrice <= 0 (Gratis/Error)       : 2517
```

Gambar 3. Analisis Rentang Waktu Transaksi dan Deteksi Transaksi Tidak Valid

Keempat, Gambar 3 menganalisis rentang waktu transaksi yang menunjukkan bahwa data mencakup periode 1 Desember 2010 hingga 9 Desember 2011, memberikan cakupan temporal kurang lebih satu tahun yang dinilai memadai untuk menggambarkan pola perilaku pembelian pelanggan secara representatif. Tahap terakhir, deteksi transaksi tidak valid mengidentifikasi sebanyak 10.624 baris dengan nilai Quantity ≤ 0 yang mengindikasikan retur atau pembatalan pesanan, serta 2.517 baris dengan UnitPrice ≤ 0 yang mengindikasikan harga tidak wajar atau tidak valid. Seluruh temuan pada tahap ini menjadi landasan utama dalam merancang prosedur prapemrosesan data pada tahap berikutnya.

3.2 Hasil Prapemrosesan Data

Tahap prapemrosesan dilaksanakan secara sistematis melalui empat prosedur utama. Rincian prosedurnya disajikan pada Tabel 2.

Tabel 2. Hasil Prapemrosesan Data

Tahap Pembersihan	Baris Terhapus	Keterangan
Data Awal	—	541.909 baris transaksi
Hapus <i>CustomerID</i> kosong	135.080	24,93% dari total data
Hapus <i>Quantity</i> ≤ 0	10.624	Transaksi retur dan batal
Hapus <i>UnitPrice</i> ≤ 0	2.517	Harga tidak valid atau gratis
Data bersih final	—	397.884 baris

Tabel 2 menjelaskan ringkasan hasil tahap prapemrosesan data dengan melalui empat prosedur utama:

- penghapusan seluruh baris yang tidak memiliki *CustomerID*,
- eliminasi transaksi retur dengan nilai *Quantity* kurang dari atau sama dengan nol,
- penghapusan entri dengan *UnitPrice* kurang dari atau sama dengan nol,
- konversi format *CustomerID* dari tipe float menjadi string. Selain itu, ditambahkan kolom *TotalSum* yang dihitung dari perkalian *Quantity* dan *UnitPrice* sebagai dasar perhitungan nilai *Monetary*.

Hasil akhir prapemrosesan diperoleh sebanyak 397.884 baris data bersih yang siap diolah lebih lanjut.

3.3 Hasil Pembentukan dan Analisis Fitur RFM

Fitur RFM dibangun melalui agregasi data per pelanggan dengan menetapkan snapshot date pada 10 Desember 2011, yaitu satu hari setelah tanggal transaksi terakhir. *Recency* dihitung sebagai selisih hari antara snapshot date dan tanggal belanja terakhir pelanggan; *Frequency* merupakan jumlah faktur unik; dan *Monetary* adalah total akumulasi nilai pembelanjaan. Proses ini menghasilkan tabel RFM untuk 4.338 pelanggan.

Untuk memperoleh gambaran karakteristik data yang terbentuk, dilakukan analisis statistik deskriptif terhadap ketiga variabel RFM. Analisis ini bertujuan untuk mengetahui pola distribusi data, tingkat penyebaran nilai, serta mengidentifikasi adanya variasi perilaku pelanggan berdasarkan aktivitas transaksinya. Statistik yang dihitung meliputi nilai minimum, kuartil, median, rata-rata, simpangan baku, dan nilai maksimum untuk setiap variabel.

Ringkasan hasil statistik deskriptif fitur RFM disajikan pada Gambar 4. Informasi ini menjadi dasar untuk memahami karakteristik pelanggan sekaligus menentukan kebutuhan transformasi dan standarisasi data sebelum dilakukan proses clustering.

```

=== Statistik Deskriptif Tabel RFM ===
      Recency  Frequency  Monetary
count  4338.00    4338.00    4338.00
mean    92.54         4.27    2054.27
std    100.01         7.70   8989.23
min      1.00         1.00      3.75
25%     18.00         1.00     307.41
50%     51.00         2.00     674.48
75%    142.00         5.00    1661.74
max    374.00        209.00  280206.02

=== Analisis Loyalitas Pelanggan ===
Total pelanggan keseluruhan : 4338 orang
Pelanggan yang beli 1 kali   : 1493 orang
Persentase One-Time Buyer   : 34.42%
  
```

Gambar 4. Statistik Deskriptif Fitur RFM

Nilai rata-rata *Recency* sebesar 92,54 hari dengan simpangan baku 100,01 menunjukkan distribusi yang sangat tersebar, mencerminkan perbedaan perilaku pembelian yang mencolok. Rata-rata *Frequency* 4,27 kali (std 7,70) mengindikasikan frekuensi pembelian yang relatif rendah, sementara rata-rata *Monetary* GBP 2.054,27 (std 8.989,23) menandakan ketimpangan nilai transaksi yang signifikan. Analisis loyalitas mengungkap bahwa 1.493 pelanggan atau

34,42% dari total hanya melakukan satu kali pembelian sepanjang periode observasi, mengisyaratkan tantangan nyata dalam hal retensi pelanggan.

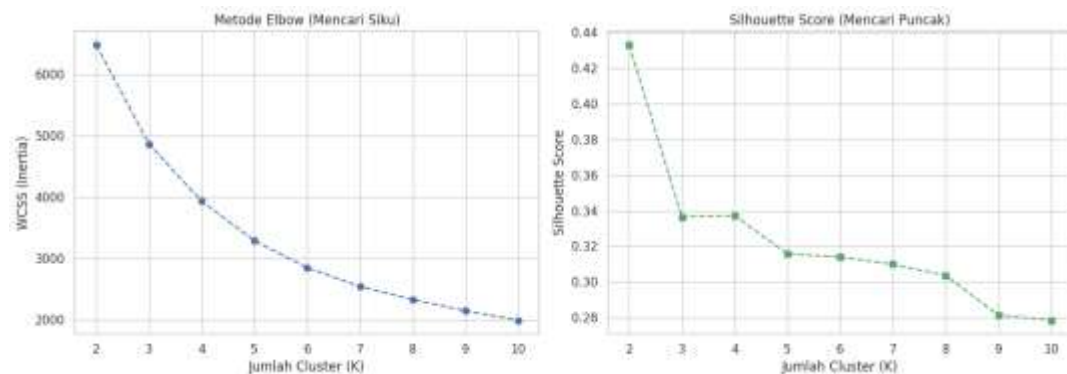
3.4 Hasil Transformasi dan Standarisasi Data

Ketiga fitur RFM menunjukkan distribusi *right-skewed* yang signifikan sehingga berpotensi mengganggu kinerja algoritma berbasis jarak. Oleh karena itu, diterapkan tiga tahap transformasi secara berurutan. Pertama, transformasi logaritma menggunakan fungsi $\log_{10}$ pada R, F, dan M untuk menekan pengaruh nilai ekstrem. Kedua, *Winsorizing* dengan batas $3 \times \text{IQR}$ diterapkan pada hasil transformasi untuk membatasi nilai anomali. Terakhir, standarisasi data menggunakan *StandardScaler* digunakan untuk menormalisasi data ke rata-rata nol dan simpangan baku satu. Data hasil transformasi ini selanjutnya disebut *rfm_scaled* dan menjadi input bagi seluruh algoritma clustering.

3.5 Penentuan Parameter Clustering

3.5.1 Algoritma K-Means

Nilai K optimal ditentukan melalui Metode Elbow dan *Silhouette Score* pada rentang K=2 hingga 10 dengan inisialisasi k-means++ dan n_init=10.

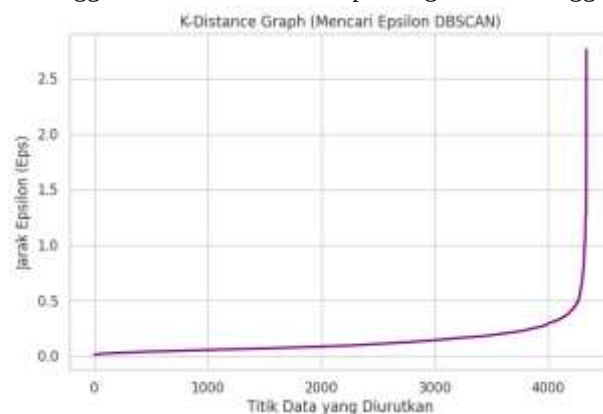


Gambar 5. Penentuan Parameter *Clustering K-Means*

Gambar 5 menunjukkan sebuah grafik, yaitu grafik *Elbow* dan *Silhouette Score*. Grafik *Elbow* menampilkan penurunan WCSS yang tajam hingga K=4 kemudian melandai, dengan titik siku paling jelas pada K=3. *Silhouette Score* tertinggi diperoleh pada K=2 sebesar 0,433, namun dengan mempertimbangkan keseimbangan antara kualitas cluster dan interpretabilitas bisnis, K=3 ditetapkan sebagai parameter optimal.

3.5.2 Algoritma DBSCAN

Nilai epsilon optimal dicari menggunakan K-Distance Graph dengan k=4 tetangga terdekat.



Gambar 6. Penentuan Parameter *Clustering DBSCAN*

Berdasarkan Gambar 6, kurva menunjukkan kenaikan tajam pada jarak sekitar 0,5, sehingga $\epsilon=0,5$ ditetapkan. Parameter $\text{min_samples}=10$ dipilih untuk memastikan setiap cluster memiliki populasi yang bermakna secara bisnis.

3.5.3 Algoritma Hierarchical Clustering

Dendrogram metode Ward dibangun dari sampel acak 3.000 titik data untuk efisiensi memori.



Gambar 7. Penentuan Parameter *Hierarchical Clustering*

Berdasarkan Gambar 7, analisis visual menunjukkan tiga kelompok utama terbentuk pada *threshold* jarak *Euclidean* sekitar 40, sehingga K=3 digunakan secara konsisten.

3.6 Segmentasi Pelanggan Menggunakan Algoritma Clustering

Ketiga algoritma dijalankan pada *rfm_scaled* dengan parameter yang telah ditentukan Distribusi lengkap disajikan pada Tabel 3.

Tabel 3. Distribusi Pelanggan per Cluster pada Setiap Algoritma

Algoritma	Cluster	Jumlah Pelanggan (N)	Persentase (%)	Recency	Frequency	Monetary (GBP)
K-Means	0 (Champions)	771	17,77	17,10	13,33	7.901,75
	1 (Lost Customers)	1.872	43,15	167,36	1,35	362,54
	2 (Regular Customers)	1.695	39,07	44,21	3,38	1.262,83
Hierarchical	0 (Lost Customers)	2.344	54,03	139,74	1,48	392,97
	1 (Regular Customers)	1.304	30,06	46,09	3,96	1.695,56
	2 (Champions)	690	15,91	19,95	14,33	8.375,76
DBSCAN	-1 (Noise/Outlier)	79	1,82	57,61	28,44	36.514,05
	0 (Mass Market)	2.780	64,08	59,18	5,33	1.981,15
	1 (Inactive)	1.479	34,09	157,11	1,00	351,05

Berdasarkan hasil Tabel 3, K-Means dan Agglomerative Clustering (*linkage=ward*) berhasil mengelompokkan seluruh 4.338 pelanggan ke dalam tiga cluster tanpa pengecualian. DBSCAN menghasilkan dua cluster inti serta satu kelompok *noise* (label -1) yang menampung 79 pelanggan dengan perilaku transaksi yang sangat atipikal.

3.7 Evaluasi dan Perbandingan Algoritma Clustering

Evaluasi kinerja dilakukan menggunakan tiga metrik internal: *Silhouette Score* (SS) yang mengukur kepadatan dan pemisahan cluster dengan rentang nilai -1 hingga 1, *Davies-Bouldin Index* (DBI) yang mengukur rasio dispersi intra-cluster terhadap jarak antar centroid di mana nilai lebih rendah lebih baik, serta *Calinski-Harabasz Index* (CHI) yang merupakan rasio dispersi antar dan dalam cluster di mana nilai lebih tinggi lebih baik. Titik noise DBSCAN dikecualikan dari perhitungan.

Tabel 4. Hasil Evaluasi Perbandingan Algoritma *Clustering*

Algoritma	Jumlah Cluster	Silhouette Score	DBI	CHI
K-Means	3	0,337	1,048	3.632
Hierarchical Clustering	3	0,320	1,105	3.275
DBSCAN	2	0,294	1,061	2.395

Berdasarkan Tabel 4, K-Means menunjukkan performa terbaik dengan *Silhouette Score* tertinggi (0,337), DBI terendah (1,048), dan CHI tertinggi (3.632), mengindikasikan cluster yang paling kompak dan terpisah. Hierarchical Clustering berada di posisi kedua dengan *Silhouette Score* 0,320, DBI 1,105 dan CHI 3.275. Sementara itu, DBSCAN dengan

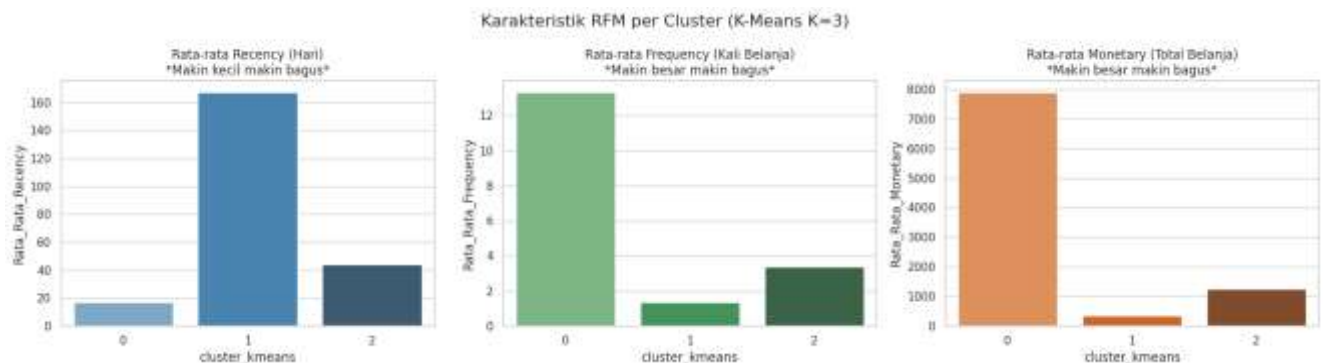
Silhouette Score 0,294 dan *CHI* 2.395 yang menunjukkan performa terendah karena distribusi data RFM yang bersifat kontinu dan tumpang tindih menyebabkan algoritma ini kesulitan membentuk cluster yang bermakna.

3.8 Profiling Segmen Pelanggan

Profiling dilakukan terhadap hasil *clustering* dari ketiga algoritma dengan menghitung nilai rata-rata *Recency*, *Frequency*, dan *Monetary* pada setiap *cluster* menggunakan nilai asli (tidak diskalakan). Tujuannya adalah untuk memahami karakteristik perilaku pembelian yang mencirikan setiap segmen. Profil lengkap segmen algoritma K-Means disajikan pada Tabel 5 dan Gambar 8.

Tabel 5. *Profiling* Segmen Pelanggan K-Means

Cluster	Jumlah Pelanggan (N)	Persentase (%)	Rata-rata Recency	Rata-rata Frequency	Rata-rata Monetary
0	771	17,77	17,10 hari	13,33 kali	GBP 7.901,75
1	1.872	43,15	167,36 hari	1,35 kali	GBP 362,54
2	1.695	39,07	44,21 hari	3,38 kali	GBP 1.262,83

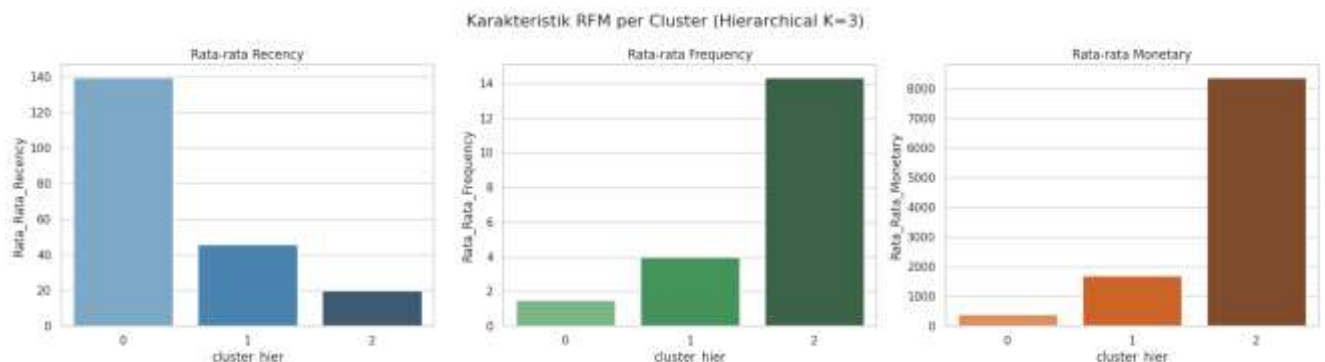


Gambar 8. Karakteristik RFM per Cluster (K-Means K=3)

Berdasarkan Tabel 5 dan Gambar 8, pada K-Means, menghasilkan tiga segmen pelanggan. Cluster 0 (17,77%) merupakan pelanggan paling aktif dan bernilai tinggi dengan *Recency* rendah serta *Frequency* dan *Monetary* tertinggi. Cluster 1 (43,15%) mencerminkan pelanggan yang sudah lama tidak bertransaksi dengan nilai transaksi rendah. Cluster 2 (39,07%) terdiri atas pelanggan aktif dengan nilai transaksi menengah.

Tabel 6. *Profiling* Segmen Pelanggan Hierarchical Clustering

Cluster	Jumlah Pelanggan (N)	Persentase (%)	Rata-rata Recency	Rata-rata Frequency	Rata-rata Monetary
0	2.344	54,03	139,74 hari	1,48 kali	GBP 392,97
1	1.304	30,06	46,09 hari	3,96 kali	GBP 1.695,56
2	690	15,91	19,95 hari	14,33 kali	GBP 8.375,76



Gambar 9. Karakteristik RFM per Cluster (Hierarchical k_graph=3)

Berdasarkan Tabel 6 dan Gambar 9, pada Hierarchical Clustering juga membentuk tiga segmen. *Cluster 0* (54,03%) didominasi pelanggan dengan aktivitas dan nilai transaksi rendah. *Cluster 1* (30,06%) menunjukkan karakteristik pelanggan menengah, sedangkan *Cluster 2* (15,91%) merupakan kelompok pelanggan paling bernilai dengan rata-rata *Monetary* tertinggi.

Tabel 7. Profiling Segmen Pelanggan DBSCAN

<i>Cluster</i>	Jumlah Pelanggan (N)	Persentase (%)	Rata-rata <i>Recency</i>	Rata-rata <i>Frequency</i>	Rata-rata <i>Monetary</i>
-1 (<i>Noise</i>)	79	1,82	57,61 hari	28,44 kali	GBP 36.514,05
0	2.780	64,08	59,18 hari	5,33 kali	GBP 1.981,15
1	1.479	34,09	157,11 hari	1,00 kali	GBP 351,05

Berdasarkan Tabel 7, pada DBSCAN menghasilkan dua *cluster* utama dan satu kelompok *noise*. Kelompok *noise* (1,82%) memiliki *Frequency* dan *Monetary* yang sangat tinggi, menunjukkan pelanggan bernilai tinggi yang teridentifikasi sebagai *outlier*. Sementara itu, *Cluster 0* (64,08%) dan *Cluster 1* (34,09%) merepresentasikan segmen pelanggan umum dan pelanggan yang relatif kurang aktif.

3.9 Pelabelan Otomatis Bisnis Segmen Pelanggan

Pelabelan segmen dilakukan secara otomatis berdasarkan karakteristik RFM. Untuk K-Means dan Hierarchical Clustering, *Cluster* dengan nilai *Monetary* tertinggi diberi label *Champions*, *cluster* dengan *Recency* tertinggi diberi label *Lost Customers*, sedangkan *cluster* lainnya diberi label *Regular Customers*.

Pada K-Means, *Cluster 0* (771 pelanggan) dilabeli *Champions* yang memerlukan program loyalitas, *Cluster 1* (1.872 pelanggan) sebagai *Lost Customers* yang membutuhkan strategi *win-back*, dan *Cluster 2* (1.695 pelanggan) sebagai *Regular Customers* yang berpotensi ditingkatkan melalui *upselling* dan *cross-selling*.

Pada Hierarchical Clustering menghasilkan label serupa namun dengan segmen *Champions* yang lebih eksklusif (690 pelanggan, 15,91%) dan *Lost Customers* yang lebih besar (2.344 pelanggan, 54,03%), mencerminkan batas pemisahan *cluster* yang lebih ketat.

Pada DBSCAN, kelompok *noise* (79 pelanggan) justru merepresentasikan pelanggan bernilai tertinggi yang dapat diperlakukan sebagai segmen *VIP/Key Account*, sementara *Cluster 0* (2.780 pelanggan) menjadi sasaran pemasaran massal dan *Cluster 1* (1.479 pelanggan) merepresentasikan pelanggan berpotensi tidak aktif.

Secara keseluruhan, K-Means terbukti paling sesuai untuk segmentasi RFM berkat keunggulan metrik evaluasi dan kemudahan interpretasi. Hierarchical Clustering merupakan alternatif yang layak, sedangkan DBSCAN memberikan nilai tambah khusus dalam deteksi otomatis pelanggan *VIP* bernilai tinggi.

4. KESIMPULAN

Penelitian ini telah berhasil mencapai tujuan sebagaimana yang diharapkan dan dinyatakan dalam bagian "Pendahuluan", yang mana pembuktiannya telah diperoleh secara komprehensif pada bagian "Hasil dan Pembahasan". Kinerja tiga algoritma clustering, yaitu K-Means, DBSCAN, dan Hierarchical Clustering, telah berhasil dibandingkan dalam konteks segmentasi pelanggan e-commerce menggunakan model RFM pada Online Retail Dataset (4.338 pelanggan). Dengan mengevaluasi ketiga algoritma secara bersamaan menggunakan Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Index, penelitian ini menghasilkan perbandingan yang objektif. Dari hasil evaluasi, K-Means terbukti menjadi algoritma yang paling unggul dengan nilai Silhouette Score 0,337, Davies-Bouldin Index 1,048, dan Calinski-Harabasz Index 3.632. Lebih dari sekadar unggul secara metrik, K-Means menghasilkan tiga segmen yang mudah dipahami dalam konteks bisnis, yaitu *Champions* (17,77%), *Regular Customers* (39,07%), dan *Lost Customers* (43,15%). Pelanggan *Champions* layak mendapat program loyalitas eksklusif, *Regular Customers* didorong melalui *upselling*, sementara *Lost Customers* perlu didekati kembali melalui kampanye *win-back*. Di sisi lain, Hierarchical Clustering menunjukkan performa kompetitif di posisi kedua, sedangkan DBSCAN memperlihatkan keunggulan unik dalam mendeteksi otomatis 79 pelanggan *noise* (*VIP*) dengan rata-rata nilai transaksi mencapai GBP 36.514. Secara keseluruhan, penelitian ini merekomendasikan K-Means sebagai algoritma utama segmentasi RFM, dengan DBSCAN sebagai pelengkap deteksi pelanggan *VIP*. Sebagai prospek pengembangan aplikasi lebih jauh, model segmentasi algoritma ini dapat diintegrasikan secara *real-time* ke dalam sistem dasbor *Customer Relationship Management* (CRM) pada perusahaan e-commerce untuk mengotomatisasi penargetan promosi. Sementara untuk prospek kajian ilmiah berikutnya, disarankan mengeksplorasi model RFMT, menerapkan reduksi dimensi PCA, serta menguji dataset yang lebih beragam agar temuan dapat digeneralisasi lebih luas.

REFERENCES

- [1] R. M. Fauzan and G. Alfian, "Segmentasi pelanggan e-commerce menggunakan fitur recency, frequency, monetary (RFM) dan algoritma klasterisasi K-Means," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 9, no. 3, pp. 170–177, 2024.
- [2] Alamsyah et al., "Customer segmentation using the integration of the recency frequency monetary model and the K-Means cluster algorithm," *Sci. J. Informatics*, vol. 9, no. 2, pp. 189–196, 2022, doi: 10.15294/sji.v9i2.39437.
- [3] P. Anitha and M. M. Patil, "RFM model for customer purchase behavior using K-Means algorithm," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 5, pp. 1785–1792, 2022, doi: 10.1016/j.jksuci.2019.12.011.
- [4] B. T. Kristanti, A. Junaidi, and E. P. Mandyartha, "Implementasi K-Means clustering dalam segmentasi pelanggan berdasarkan usia, pendapatan, dan model RFM (studi kasus: Lantikya Store Jombang)," *JITET (Jurnal Inform. dan Tek. Elektro Terapan)*, vol. 12, no. 3, 2024.
- [5] R. Windy, R. Rita, and Mustafid, "Segmentasi pelanggan e-money dengan menggunakan algoritma DBSCAN (Density Based Spatial Clustering Applications with Noise) di Provinsi DKI Jakarta," *J. GAUSSIAN*, vol. 9, pp. 162–169, 2020.
- [6] I. K. Y. Triantara, M. A. Raharja, and I. B. G. Sarasvananda, "Segmentasi pelanggan berbasis RFMT menggunakan K-Means dan hierarchical clustering," *J. Nas. Teknol. Inf. dan Apl.*, vol. 3, no. M, pp. 917–924, 2025.
- [7] M. F. Fadhillah, S. Aldo Lovely Arief, and I. Puspitasari, "Segmentasi pelanggan dengan algoritma clustering berdasarkan atribut recency, frequency dan monetary (RFM)," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. January, pp. 48–56, 2025.
- [8] D. S. Nugaha, I. Thoib, N. Sururi, F. B. F., and B. P. Candra, "Segmentasi pelanggan berbasis analisis RFM menggunakan algoritma K-Means clustering," *J. Komput. Teknol. Inf. Sist. Komput.*, vol. 4, no. 2, pp. 1361–1369, 2025.
- [9] A. Saputra and R. Yusuf, "Perbandingan algoritma DBSCAN dan K-Means dalam segmentasi pelanggan pengguna transportasi publik Transjakarta menggunakan metode RFM," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. October, pp. 1346–1361, 2024.
- [10] S. D. K. Wardani, A. S. Ariyanto, M. Umroh, and D. Rolliawati, "Perbandingan hasil metode clustering K-Means, DBSCAN & hierarchical untuk analisa segmentasi pasar," *JIKO (Jurnal Inform. dan Komputer)*, vol. 7, no. 2, pp. 191–201, 2023.
- [11] T. Z. Fate, D. S. Ezra, and I. I. Samuel, "Comparative analysis of clustering techniques for customer segmentation: Evaluating K-Means, hierarchical, and DBSCAN models alongside RFM frameworks to enhance marketing strategies through behavioral, demographic, and transactional insights," *Int. J. Adv. Eng. Manag.*, vol. 7, no. 04, pp. 34–43, 2025, doi: 10.35629/5252-07043443.
- [12] J. M. John, O. Shobayo, and Bayode, "An exploration of clustering algorithms for customer segmentation in the UK retail market," *Analytics*, vol. 2, no. 4, pp. 809–823, 2023.
- [13] P. I. Pangestu, T. I. Hermanto, and D. Irmayanti, "Analisis segmentasi pelanggan berbasis model recency frequency dan monetary (RFM)," *JITET (Jurnal Inform. dan Tek. Elektro Terapan)*, vol. 11, no. 3, pp. 930–937, 2023.
- [14] Widyawati, W. Laksito Yuly Saptomo, and Y. R. W. Utami, "Penerapan agglomerative hierarchical clustering untuk segmentasi pelanggan," *J. Ilm. Sinus*, vol. 18, no. 1, pp. 75–87, 2020.
- [15] I. R. C. Pata, Statiswaty, and N. Ransi, "Segmentasi pemetaan pelanggan potensial menggunakan algoritma DBSCAN dengan RFM model berbasis web," *ANIMATOR*, vol. 1, no. 2, pp. 63–71, 2023.
- [16] A. D. Savitri, F. A. Bachtiar, and N. Y. Setiawan, "Segmentasi pelanggan menggunakan metode K-Means clustering berdasarkan model RFM pada klinik kecantikan (studi kasus: Belle Crown Malang)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 9, pp. 2957–2966, 2018.
- [17] P. R. Sihombing, D. A. Sunarjo, Y. Paulus, and A. C. Yuda, "Identifikasi data outlier (pencilan) dan kenormalan data pada data univariat serta alternatif penyelesaiannya," *J. Ekon. dan Stat. Indones.*, vol. 2, no. 3, pp. 307–315, 2022, doi: 10.11594/jesi.02.03.07.
- [18] I. F. Ashari, E. D. Nugroho, R. Baraku, I. N. Yanda, and R. Liwardana, "Analysis of elbow, silhouette, Davies-Bouldin, Calinski-Harabasz, and rand-index evaluation on K-Means algorithm for classifying flood-affected areas in Jakarta," *J. Appl. Informatics Comput.*, vol. 7, no. 1, pp. 95–103, 2023.
- [19] I. Permana and F. N. Salisah, "Pengaruh normalisasi data terhadap performa hasil klasifikasi algoritma backpropagation," *IJIRSE Indones. J. Inform. Res. Softw. Eng.*, vol. 2, no. 1, pp. 67–72, 2022.
- [20] S. I. Murpratiwi, I. G. A. Indrawan, and A. Aranta, "Analisis pemilihan cluster optimal dalam segmentasi," *J.*

Pendidik. Teknol. dan Kejuru., vol. 18, no. 2, pp. 152–163, 2021.

- [21] R. W. B. S. Berahmana, F. A. Mohammed, and K. Chairuang, “Customer segmentation based on RFM model,” *LONTAR Komput.*, vol. 11, no. 1, pp. 32–43, 2020.
- [22] K. E. Setiawan and A. Kurniawan, “Pengelompokan rumah sakit di Jakarta menggunakan model DBSCAN, Gaussian mixture, dan hierarchical clustering,” *J. Inform. Terpadu*, vol. 9, no. 2, pp. 149–156, 2023.
- [23] Y. Hasan, “Pengukuran silhouette score dan Davies-Bouldin index pada hasil cluster K-Means,” *JITET (Jurnal Inform. dan Tek. Elektro Terapan)*, vol. 12, no. 3, pp. 3517–3536, 2024.
- [24] I. R. Effendy, Y. A. Manurung, and M. I. Hujaifah, “Implementasi algoritma K-Means clustering untuk mencari preferensi pelanggan toko online Tazeee Clothes,” *J. MEDIA Inform. [JUMIN]*, vol. 7, no. 1, pp. 69–76, 2026.