R.Saputra dkk/ JIMI 11 (2) pp. 200-210

P-ISSN : 2549-7480
E-ISSN : 2549-6301

JURNAL ILMIAH INFORMATIKA
E-ISSN : 2549-6301
ISSN : 2549-7480
Faculty of Sains and Technology, Ibrahimy University
S4-Accredited : 105/E/KPT/2022
Journal page is available to : https://journal.ibrahimy.ac.id/index.php/JIMI

KOMPARASI METODE CLUSTERING K-MEANS, K-MEDOIDS, FUZZY C-MEANS, DAN DBSCAN UNTUK SEGMENTASI PELANGGAN E-COMMERCE

Rohman Saputra[1)], Risanto Darmawan[2)], Junaidi[3)]

[1,2,3] *Program Studi Teknik Informatika, Fakultas Teknik, Universitas Krisnadwipayana*

email: [1] rohmansaputra1203@gmail.com, [2] risantodarmawan@unkris.ac.id, [3] junaidi@unkris.ac.id

ARTICLE INFO

Article History:
Recieved : 22 Oktober 2026
Accepted : 1 November 2026
Published : 15 November 2026

Keywords:

Customer Segmentation;
E-Commerce;
RFM;
Clustering;
K-Means;

IEEE style in citing this article:

R. Saputra, R. Darmawan, Junaidi, "Komparasi Metode Clustering K-Means, K-Medoids, Fuzzy C-Means , dan DBSCAN untuk Segmentasi Pelanggan E-Commerce", *jurnal.ilmiah.informatika*, vol. 11, no. 2, pp. 200-211, Des. 2026.

Corresponding Author:
Rohman Saputra
Teknik Informatika, Universitas Krisnadwipayana

ABSTRACT

The rapid growth of e-commerce generates large volumes of customer transaction data that can be utilized to support data-driven marketing strategies. However, this transaction data needs to be processed properly to produce meaningful information through customer segmentation. This study aims to compare the performance of K-Means, K-Medoids, Fuzzy C-Means, and DBSCAN clustering methods in segmenting e-commerce customers based on RFM (Recency, Frequency, Monetary) features. The evaluation was conducted using three metrics: Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). The results show that K-Means achieved a Silhouette Score of 0.369997, a DBI of 0.919935, and a CHI of 38,226.138. While K-Medoids achieved the best DBI value of 0.914247, K-Means was considered the overall most optimal method due to its superiority in both Silhouette Score and CHI. The customer segmentation results successfully identified several distinct groups, including active and high-value customers, passive or at-risk customers, and new or regular customers. Therefore, RFM-based clustering methods can effectively help businesses understand customer characteristics to support more targeted marketing strategie .

© 2026 Jurnal Ilmiah Informatika (Scientific Informatics Journal) with CC BY NC licence

1. PENDAHULUAN

Perkembangan industri *e-commerce* dalam beberapa tahun terakhir menunjukkan pertumbuhan yang sangat signifikan, mendorong terciptanya ekosistem digital dengan akumulasi data transaksi pelanggan berskala masif [1] . Aktivitas operasional yang berlangsung tanpa henti ini menghasilkan volume data historis yang sangat besar, heterogen, dan kompleks [2]

Kendati demikian, banyak perusahaan *e-commerce* masih menghadapi kendala krusial dalam mengelola dan mengekstraksi tumpukan *dark data* tersebut menjadi wawasan analitis yang berharga untuk mendukung pengambilan keputusan strategis yang tepat guna atau *data-driven decision making* [3] .

Salah satu permasalahan paling mendasar adalah kurang efektifnya segmentasi pelanggan, yang menyebabkan strategi pemasaran berjalan tidak tepat sasaran, rendahnya tingkat retensi pelanggan, serta terjadinya inefisiensi anggaran promosi [1] , [4]

Segmentasi pasar secara konvensional yang hanya mengandalkan intuisi dinilai sudah tidak lagi relevan untuk merespons dinamika perilaku konsumen digital saat ini [1] . Oleh karena itu, pendekatan *data mining*, khususnya teknik *clustering*, diimplementasikan sebagai metodologi andal untuk membagi populasi pelanggan yang heterogen ke dalam beberapa sub-kelompok homogen (*cluster*) berdasarkan kemiripan pola interaksi dan transaksi tanpa memerlukan definisi kelas sebelumnya [5] , [6] .

Dalam merepresentasikan kerangka perilaku tersebut, analisis fitur RFM (*Recency, Frequency, Monetary*) digunakan secara luas karena terbukti efektif dalam memetakan tingkat keaktifan, intensitas loyalitas, dan kontribusi finansial pelanggan terhadap sebuah bisnis [7] , [8] . Kombinasi ekstraksi fitur RFM dengan algoritma *clustering* memungkinkan manajemen *e-commerce* mengidentifikasi mana segmen pelanggan yang bernilai tinggi (*high-value*) dan mana pelanggan yang pasif atau berisiko beralih ke kompetitor (*churn*) [9] , [10]

Berbagai algoritma *clustering* memiliki karakteristik komputasi, keunggulan, serta keterbatasan yang berbeda saat dihadapkan pada dimensi data *e-commerce* [11] . K-Means merupakan metode partisi berbasis *centroid* yang sangat populer karena efisiensi dan kecepatannya dalam memproses *dataset* berukuran masif, namun kelemahan utamanya terletak pada sensitivitas yang tinggi terhadap penentuan jumlah *cluster* secara manual dan rentan terhadap data pencilan (*outlier*) [11] , [12] .

Sebagai pengembangan dari kelemahan tersebut, algoritma K-Medoids diusulkan karena kemampuannya menggunakan titik data observasi aktual (*medoid*) sebagai titik pusat *cluster*, sehingga lebih tangguh (*robust*) terhadap distorsi yang disebabkan oleh anomali atau *noise* [13] Di sisi lain, algoritma Fuzzy C-Means (FCM) menawarkan paradigma yang berbeda melalui pendekatan *soft clustering*, di mana sebuah titik data dimungkinkan untuk bergabung ke dalam beberapa *cluster* sekaligus dengan nilai derajat keanggotaan tertentu (*membership degree*) [14]

Karakteristik ini menjadikannya sangat fleksibel untuk menangani data perilaku konsumen yang batas antar kelompoknya kerap kali tumpang tindih [14] . Sementara itu, algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) bekerja berdasarkan ambang batas tingkat kerapatan data di ruang fitur, sehingga unggul dalam mengidentifikasi *cluster* yang berbentuk non-geometris sekaligus memisahkan pencilan data sebagai *noise*

DOI : https://doi.org/10.35316/jimi.v11i2.200-210
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

P-ISSN : 2549-7480
R.Saputra dkk/ JIMI 11 (2) pp. 200-210 E-ISSN : 2549-6301

secara otomatis tanpa harus menentukan jumlah kelompok sejak awal [15] ..

Walaupun beberapa penelitian terdahulu telah mengaplikasikan algoritma-algoritma tersebut untuk segmentasi pasar, literatur yang secara komprehensif mengkomparasikan keempat metode secara langsung—K-Means, K-Medoids, Fuzzy C-Means, dan DBSCAN—pada satu *dataset e-commerce* berbasis agregasi RFM masih sangat terbatas [1] . Berdasarkan identifikasi kesenjangan (*research gap*) tersebut, penelitian ini bertujuan untuk mengevaluasi dan membandingkan kinerja keempat algoritma *clustering* secara obyektif dalam memetakan segmentasi pelanggan *e-commerce* .

Kualitas struktur kelompok yang terbentuk akan dinilai secara kuantitatif menggunakan tiga metrik validasi internal utama, yaitu *Silhouette Score*, *Davies-Bouldin Index* (DBI), dan *Calinski-Harabasz Index* (CHI) [7] , [11] . Melalui komparasi evaluasi ini, diharapkan dapat ditemukan algoritma *clustering* yang paling optimal dalam mendukung perumusan strategi pemasaran yang dipersonalisasi serta meningkatkan efektivitas promosi di industri *e-commerce* [8] .

Sebagai kontribusi utama, penelitian ini tidak hanya berhenti pada tahap penemuan algoritma dengan kinerja metrik evaluasi terbaik, tetapi juga melangkah lebih jauh dengan menyajikan interpretasi bisnis secara komprehensif dari masing-masing segmen pelanggan yang terbentuk [4] , [8] . Hasil klasterisasi dari algoritma yang paling optimal akan dipetakan kembali ke dalam strategi *Customer Relationship Management* (CRM) yang konkret dan dapat ditindaklanjuti [3] , [6] . Sebagai contoh, luaran model akan digunakan untuk merumuskan program retensi eksklusif bagi pelanggan aktif bernilai tinggi, merancang skema peningkatan frekuensi dan *bundling* untuk pelanggan reguler, hingga merencanakan langkah reaktivasi melalui penawaran promosi bagi segmen pelanggan pasif yang rentan beralih ke kompetitor (*churn*) [1] , [14] . Dengan demikian, kerangka kerja pada penelitian ini diharapkan dapat memberikan wawasan berbasis data (*actionable insights*) bagi praktisi bisnis maupun tim pemasaran [4], [5] .

2. METODE PENELITIAN

Jenis Penelitian dan Tahapan

Penelitian ini tergolong penelitian terapan dengan pendekatan kuantitatif komparatif yang dijalankan melalui eksperimen komputasi. Sifat terapannya terletak pada tujuan hasilnya, yaitu membantu pelaku *e-commerce* memahami kelompok pelanggannya, sedangkan sifat komparatifnya berasal dari pengujian empat algoritma *clustering*, yaitu *K-Means*, *K-Medoids*, *Fuzzy C-Means* (FCM), dan DBSCAN, pada data yang sama. Karena *clustering* termasuk pembelajaran tanpa label, tidak ada kelas acuan yang dapat dipakai untuk menghitung akurasi. Mutu kelompok dinilai dari struktur data itu sendiri melalui metrik validasi internal.

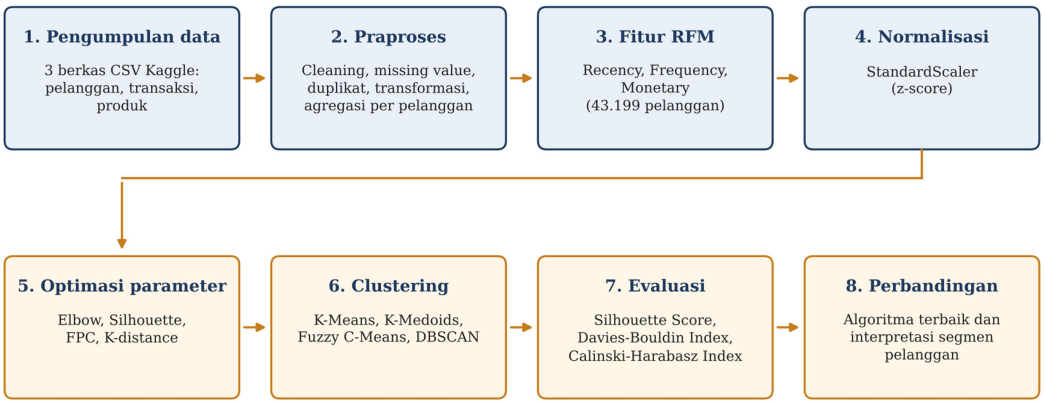
Agar perbandingan adil, seluruh algoritma menerima masukan yang identik, yaitu dataset RFM yang dinormalisasi dengan skema yang sama. Parameter tiap algoritma dioptimalkan secara terpisah, kemudian algoritma dibandingkan pada konfigurasi terbaiknya masing-masing. Jumlah *cluster* tidak disamakan secara paksa karena keempat algoritma membentuk kelompok dengan mekanisme yang berbeda. Seluruh eksperimen dijalankan dengan bahasa Python di Google Colaboratory.

Tahapan penelitian disusun mengikuti alur penambangan data yang selaras dengan kerangka CRISP-DM (*Cross-*

DOI : https://doi.org/10.35316/jimi.v11i2.200-210
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

Industry Standard Process for Data Mining) [16] , yang mencakup pemahaman data, persiapan data, pemodelan, dan evaluasi. Seperti terlihat pada Gambar 1, delapan tahap dijalankan berurutan. Tahap 1 dan 2 mengumpulkan serta membersihkan tiga berkas data dari Kaggle. Tahap 3 mengubah data transaksi menjadi fitur RFM untuk setiap pelanggan, dan tahap 4 menyeragamkan skala ketiga fitur tersebut.

Tahap 5 menentukan parameter optimal setiap algoritma, dan tahap 6 menjalankan keempat algoritma pada dataset RFM. Tahap 7 menilai hasil *clustering* dengan *Silhouette Score*, *Davies-Bouldin Index* (DBI), dan *Calinski-Harabasz Index* (CHI). Pada tahap 8, ketiga metrik dibandingkan untuk menetapkan algoritma terbaik, lalu *cluster* yang dihasilkannya ditafsirkan sebagai segmen pelanggan yang dapat ditindaklanjuti dalam strategi pemasaran. Hasil evaluasi akhirnya juga disajikan pada dasbor interaktif berbasis Streamlit yang dijalankan melalui Google Colaboratory.



Gambar 1. Tahapan penelitian segmentasi pelanggan *e-commerce* dengan empat algoritma *clustering*

Data dan Praproses

Penelitian ini memakai data sekunder berupa dataset *e-commerce* berformat CSV yang diunduh dari Kaggle. Dataset terdiri atas tiga berkas, yaitu data pelanggan, detail transaksi, dan produk (Tabel 1). Atribut identitas seperti *name*, *email*, *address*, dan *phone_number* dikeluarkan karena tidak berperan dalam pembentukan fitur RFM.

Tabel 1. Dataset penelitian

Berkas	Baris	Atribut
customers.csv	50.000	8
transaction_details.csv	200.000	6
products.csv	12	13

Praproses mencakup pemeriksaan nilai kosong, data duplikat, dan konsistensi format, seleksi atribut, serta transformasi data. Kolom tanggal transaksi diubah ke format *datetime* untuk menghitung *Recency*, sedangkan nilai total transaksi dibentuk dari perkalian kuantitas dan harga satuan. Setelah transaksi digabung per pelanggan, diperoleh 43.199 pelanggan dengan satu baris untuk setiap pelanggan, tanpa nilai kosong, duplikat, maupun nilai RFM yang tidak valid.

Ekstraksi Fitur RFM

Perilaku transaksi setiap pelanggan diringkas ke dalam tiga fitur RFM yang lazim dipakai pada segmentasi pelanggan *e-commerce*:

- *Recency*: selisih hari antara tanggal referensi penelitian dan transaksi terakhir pelanggan; nilai kecil

P-ISSN : 2549-7480
R.Saputra dkk/ JIMI 11 (2) pp. 200-210 E-ISSN : 2549-6301

menandakan pelanggan yang lebih aktif.
- *Frequency*: banyaknya transaksi yang dilakukan pelanggan selama periode pengamatan.
- *Monetary*: akumulasi nilai seluruh transaksi pelanggan pada periode yang sama.

Hasilnya adalah dataset RFM berisi 43.199 baris dengan empat kolom (*CustomerID*, *Recency*, *Frequency*, dan *Monetary*). Rentang ketiga fitur sangat timpang: *Frequency* hanya 1–10 transaksi, sedangkan *Monetary* berkisar dari Rp503.000 hingga Rp103.450.000. Karena itu, fitur distandardisasi dengan *StandardScaler* seperti pada Persamaan (1) sebelum *clustering* agar fitur bernilai besar tidak mendominasi perhitungan jarak.

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

dengan x adalah nilai asli fitur, μ rata-rata fitur, dan σ simpangan baku fitur.

Algoritma Clustering dan Optimasi Parameter

Keempat algoritma mewakili pendekatan pengelompokan yang berbeda. *K-Means* berbasis *centroid*: setiap data ditetapkan ke pusat terdekat dan pusat diperbarui berulang kali hingga konvergen [17] , dengan inisialisasi *k-means++* agar pusat awal terpilih lebih terarah [18] . Algoritma ini dipilih karena komputasinya efisien untuk data numerik berukuran besar. *K-Medoids* memakai objek data aktual sebagai pusat *cluster* sehingga lebih tahan terhadap pencilan [19] . Mengingat data yang besar, *medoid* dicari dengan pendekatan berbasis sampel bergaya CLARA[19] pada 10.000 data, lalu seluruh pelanggan diberi label berdasarkan jaraknya ke *medoid* akhir. FCM memberi setiap pelanggan derajat keanggotaan pada setiap *cluster* sehingga sesuai untuk batas segmen yang tidak tegas [20] . DBSCAN berbasis kepadatan [21] , tidak memerlukan jumlah *cluster* di awal, dan menandai titik di area yang jarang sebagai *noise*.

Parameter setiap algoritma dioptimalkan secara terpisah (Tabel 2). Pada *K-Means*, nilai K diuji dari 2 sampai 10. *Elbow Method* [22] berdasarkan inertia (Persamaan 2) hanya menghasilkan kandidat K, yang kemudian dibandingkan melalui *Silhouette Score*, DBI, dan CHI. Untuk menekan beban komputasi, *Silhouette Score* pada tahap penyetelan dihitung pada sampel acak maksimal 10.000 data, sedangkan DBI dan CHI model akhir dihitung pada seluruh data. Setiap algoritma dibandingkan pada konfigurasi terbaiknya masing-masing tanpa memaksakan jumlah *cluster* yang sama.

$$J = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2 \quad (2)$$

dengan K jumlah *cluster*, C_k himpunan anggota *cluster* ke-k, dan μ_k *centroid* dari *cluster* tersebut.

Tabel 2. Parameter yang dioptimalkan pada setiap algoritma

Algoritma	Parameter yang diuji	Dasar pemilihan
K-Means	K = 2–10	*Elbow*, *Silhouette*, DBI, CHI
K-Medoids	K = 2–6; jarak Euclidean	*Silhouette*
FCM	C = 2–6; m = 1,5; 2,0; 2,5	*Silhouette*, FPC
DBSCAN	*eps* (dari *K-distance plot*); *min_samples*	*Silhouette*, jumlah *cluster*, *noise*

Evaluasi

Karena *clustering* merupakan pembelajaran tanpa label, mutu *cluster* dinilai dengan tiga metrik validasi internal. *Silhouette Score* (Persamaan 3) membandingkan jarak rata-rata data ke anggota *cluster*-nya sendiri, a(i), dengan jarak rata-rata ke *cluster* terdekat lainnya, b(i). Nilainya berada pada rentang −1

DOI : https://doi.org/10.35316/jimi.v11i2.200-210
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

sampai 1, dan semakin tinggi semakin baik [23]

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (3)$$

Davies-Bouldin Index (DBI, Persamaan 4) [24] merata-ratakan rasio kemiripan terbesar setiap *cluster* terhadap *cluster* lain, dengan s_i penyebaran data dalam *cluster* i dan d_{ij} jarak antara pusat *cluster* i dan j. Nilai yang lebih rendah menunjukkan *cluster* yang lebih terpisah.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j\neq i} \frac{s_i+s_j}{d_{ij}} \quad (4)$$

Calinski-Harabasz Index (CHI, Persamaan 5) [25] adalah rasio dispersi antar-*cluster* terhadap dispersi dalam-*cluster*, dengan tr(B_k) dan tr(W_k) berturut-turut jejak matriks dispersi antar- dan dalam-*cluster*, n jumlah data, dan k jumlah *cluster*. Semakin tinggi nilainya, semakin baik.

$$CHI = \frac{\mathrm{tr}(B_k)(n-k)}{\mathrm{tr}(W_k)(k-1)} \quad (5)$$

Algoritma terbaik ditetapkan dari gabungan ketiga metrik, yaitu *Silhouette Score* dan CHI tertinggi serta DBI terendah. Hasil *clustering* algoritma terbaik kemudian diinterpretasikan menjadi segmen pelanggan berdasarkan profil *Recency*, *Frequency*, dan *Monetary* pada setiap *cluster*.

3. HASIL DAN PEMBAHASAN

Karakteristik Data RFM

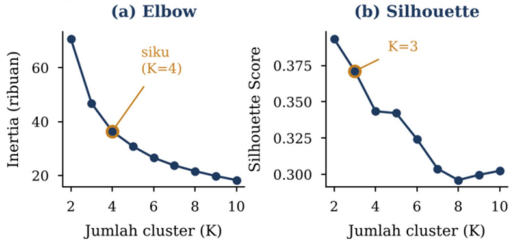
Dataset akhir memuat 43.199 pelanggan, masing-masing diwakili satu baris nilai RFM (Tabel 3). Rata-rata *Recency* mencapai 375,68 hari dengan simpangan baku 287,59 hari dan rentang 1–1.096 hari, sehingga data memuat pelanggan yang baru bertransaksi sekaligus pelanggan yang sudah lama tidak kembali. *Frequency* rata-rata hanya 2,31 transaksi (simpangan baku 1,26); sebagian besar pelanggan bertransaksi satu sampai tiga kali, dan jumlahnya menyusut pada frekuensi yang lebih tinggi. Nilai *Monetary* rata-rata sekitar Rp18,82 juta dengan simpangan baku Rp12,22 juta dan rentang Rp0,50 juta sampai Rp103,45 juta, sehingga kontribusi finansial antarpelanggan sangat beragam. Jauhnya selisih skala antarfitur ini menegaskan perlunya standardisasi sebelum *clustering*.

Tabel 3. Statistik deskriptif fitur RFM (n = 43.199)

Atribut	Rata-rata	Median	Min–maks
Recency (hari)	375,68	310	1–1.096
Frequency	2,31	2	1–10
Monetary (Rp juta)	18,82	16,38	0,50–103,45

Hasil Optimasi Parameter

Pada *K-Means*, inertia turun tajam dari K = 2 sampai K = 4 lalu melandai, sehingga titik siku terdeteksi pada K = 4 (Gambar 2a). Karena *Elbow Method* hanya berfungsi sebagai penunjuk kandidat, nilai K = 3, 4, dan 5 dibandingkan lebih lanjut. *Silhouette Score* tertinggi secara keseluruhan memang jatuh pada K = 2 (0,393), tetapi dua kelompok dinilai terlalu umum dan terletak sebelum titik siku. Di antara kandidat di sekitar siku, K = 3 unggul pada *Silhouette Score* (0,371) serta menghasilkan DBI terendah (0,920) dan CHI tertinggi (38.226,23) (Gambar 2b), sehingga K = 3 ditetapkan sebagai konfigurasi akhir.



Gambar 2. Penyetelan K pada *K-Means*: (a) inertia dan (b) *Silhouette Score*

K-Medoids mencapai *Silhouette Score* tertinggi pada K = 2 (0,404), dan skornyamenurun ketika cluster ditambah, sehingga dua cluster dipilih. Pada FCM,

P-ISSN : 2549-7480
R.Saputra dkk/ JIMI 11 (2) pp. 200-210 E-ISSN : 2549-6301

pasangan C = 2 dan m = 1,5 memberikan *Silhouette Score* tertinggi (0,390) dengan FPC 0,860, yang menandakan keanggotaan data cukup tegas. Untuk setiap jumlah cluster, menaikkan m selalu menurunkan FPC dan skor *Silhouette*. Pada DBSCAN, kandidat *eps* diambil dari *K-distance plot*, dan kombinasi *eps* = 0,2221 dengan *min_samples* = 20 menghasilkan *Silhouette Score* tertinggi (0,175), dengan enam cluster dan 742 titik *noise* (1,72%). Konfigurasi optimal keempat algoritma dirangkum pada Tabel 4.

Tabel 4. Parameter optimal setiap algoritma

Algoritma	Parameter optimal	Cluster
K-Means	K = 3; *k-means++*; *n_init* = 20	3
K-Medoids	K = 2; jarak Euclidean; sampel 10.000	2
FCM	C = 2; m = 1,5	2
DBSCAN	*eps* = 0,2221; *min_samples* = 20	6

Hasil Clustering Setiap Algoritma

Komposisi cluster berbeda cukup jauh antaralgoritma (Tabel 5). *K-Means* membagi pelanggan ke dalam tiga cluster beranggotakan 12.600 (29,17%), 19.046 (44,09%), dan 11.553 (26,74%) pelanggan. Pembagian ini paling proporsional, dan visualisasi PCA memperlihatkan batas antarkelompok yang relatif jelas. *K-Medoids* membentuk dua cluster, yaitu 12.521 (28,98%) dan 30.678 (71,02%) pelanggan, sehingga satu cluster mendominasi dan segmentasinya lebih umum. FCM juga menghasilkan dua cluster (26.083 atau 60,38% dan 17.116 atau 39,62%) berdasarkan keanggotaan tertinggi. Karena setiap pelanggan memiliki derajat keanggotaan pada semua cluster, batas FCM lebih lentur, yang cocok untuk perilaku transaksi yang saling tumpang tindih.

Tabel 5. Komposisi hasil *clustering* setiap algoritma

Algoritma	Cluster	Terbesar	*Noise*
K-Means	3	19.046 (44,09%)	0
K-Medoids	2	30.678 (71,02%)	0
FCM	2	26.083 (60,38%)	0
DBSCAN	6	13.663 (31,63%)	742 (1,72%)

DBSCAN menghasilkan enam cluster dengan ukuran yang sangat timpang, dari 438 hingga 13.663 pelanggan, ditambah 742 pelanggan yang ditandai sebagai *noise*. Pada visualisasi PCA, sebagian cluster DBSCAN tampak berdekatan dengan batas yang kurang tegas, kondisi yang tercermin pada nilai evaluasinya yang paling rendah.

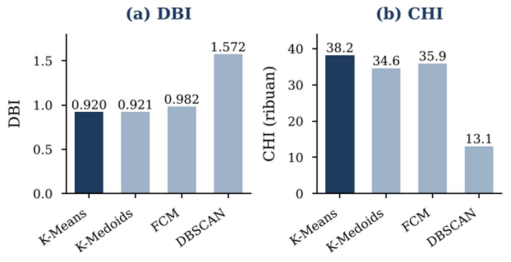
Perbandingan Hasil Evaluasi

Perbandingan akhir pada konfigurasi optimal ditunjukkan oleh Tabel 6 dan Gambar 3. DBI dan CHI menjadi penentu utama, sedangkan *Silhouette Score* berperan sebagai informasi pendukung dari tahap optimasi. *K-Means* memperoleh DBI terendah (0,9196) dan CHI tertinggi (38.226,23), yang berarti cluster-nya paling kompak sekaligus paling terpisah. *K-Medoids* hampir menyamai pada DBI (0,9210) dan mencatat *Silhouette Score* tertinggi (0,4035), tetapi CHI-nya (34.639,20) lebih rendah dan hanya membentuk dua cluster. FCM tampil cukup baik, namun masih di bawah *K-Means* pada DBI (0,9822) maupun CHI (35.928,81).

Tabel 6. Perbandingan evaluasi pada konfigurasi optimal (cetak tebal menandai nilai terbaik)

Algoritma	Silhouette	DBI	CHI
K-Means	0,3712	**0,9196**	**38.226,23**
K-Medoids	**0,4035**	0,9210	34.639,20
FCM	0,3896	0,9822	35.928,81
DBSCAN	0,1753	1,5724	13.079,73

DOI : https://doi.org/10.35316/jimi.v11i2.200-210 © 2026 Jurnal Ilmiah Informatika with CC BY NC licence



Gambar 3. Perbandingan (a) DBI (makin rendah makin baik) dan (b) CHI (makin tinggi makin baik)

DBSCAN berada di posisi terakhir pada ketiga metrik, dengan *Silhouette Score* 0,1753, DBI 1,5724, dan CHI 13.079,73. Dengan demikian *K-Means* ditetapkan sebagai algoritma terbaik pada data pelanggan *e-commerce* berbasis RFM ini. Selain unggul pada metrik utama, tiga cluster yang dibentuknya lebih informatif untuk menafsirkan perilaku pelanggan dibandingkan dua kelompok milik *K-Medoids* dan FCM.

Interpretasi Segmen Pelanggan

Interpretasi dilakukan pada hasil *K-Means* dengan membaca nilai rata-rata *Recency*, *Frequency*, dan *Monetary* setiap cluster (Tabel 7). Cluster 0 (12.600 pelanggan atau 29,17%) memiliki *Recency* sekitar 744,52 hari, jauh lebih tinggi daripada cluster lain, dengan *Frequency* 1,42 transaksi dan *Monetary* sekitar Rp11,08 juta. Pelanggan ini sudah lama tidak bertransaksi dan nilai belanjanya rendah, sehingga digolongkan sebagai pelanggan pasif atau berisiko hilang.

Tabel 7. Profil rata-rata segmen hasil *K-Means* (R = *Recency* dalam hari, F = *Frequency*, M = *Monetary* dalam Rp juta)

Segmen	Anggota	R	F	M
0: pasif / berisiko hilang	12.600 (29,17%)	744,52	1,42	11,08
1: reguler / potensial	19.046 (44,09%)	224,10	1,93	14,47
2: aktif / bernilai tinggi	11.553 (26,74%)	223,28	3,93	34,42

Cluster 1 (19.046 pelanggan atau 44,09%) merupakan kelompok terbesar, dengan *Recency* 224,10 hari, *Frequency* 1,93, dan *Monetary* sekitar Rp14,47 juta. Pelanggan ini masih aktif tetapi intensitas dan nilai belanjanya belum tinggi, sehingga disebut pelanggan reguler atau potensial. Cluster 2 (11.553 pelanggan atau 26,74%) memiliki *Recency* yang hampir sama dengan cluster 1 (223,28 hari), tetapi *Frequency* (3,93) dan *Monetary* (sekitar Rp34,42 juta) tertinggi. Karena itu, cluster ini dikategorikan sebagai pelanggan aktif dan bernilai tinggi. Perbedaan cluster 1 dan 2 terutama terletak pada frekuensi serta nilai belanja, bukan pada kebaruan transaksi.

Hasil segmentasi ini dapat diterjemahkan menjadi strategi yang berbeda untuk tiap kelompok. Cluster 0 dapat disasar dengan program reaktivasi, misalnya voucher atau promosi untuk berbelanja kembali. Cluster 1 dapat didorong agar lebih sering dan lebih banyak berbelanja melalui program loyalitas atau paket *bundling*. Cluster 2 perlu dipertahankan melalui hadiah dan penawaran eksklusif.

4. KESIMPULAN

Penelitian ini membandingkan *K-Means*, *K-Medoids*, *Fuzzy C-Means* (FCM), dan DBSCAN untuk menyegmentasi 43.199 pelanggan *e-commerce* berdasarkan fitur RFM. Hasilnya menunjukkan bahwa *K-Means* merupakan algoritma yang paling sesuai untuk data ini. Dengan tiga cluster, *K-Means* mencatat DBI terendah (0,9196) dan CHI tertinggi (38.226,23), sehingga kelompok yang dihasilkannya paling kompak dan paling terpisah. *K-Medoids* hampir menyamai pada DBI (0,9210) dan unggul pada *Silhouette Score* (0,4035), tetapi hanya membentuk dua kelompok yang lebih umum. FCM menunjukkan kinerja yang mendekati *K-Means*, sedangkan DBSCAN memperoleh

nilai terendah pada ketiga metrik karena cluster-nya saling berdekatan dan berukuran timpang.

K-Means membagi pelanggan menjadi tiga segmen. Segmen pasif atau berisiko hilang (29,17%) memiliki *Recency* sekitar 744,52 hari. Segmen reguler atau potensial (44,09%) adalah kelompok terbesar. Segmen aktif dan bernilai tinggi (26,74%) paling sering bertransaksi, rata-rata 3,93 kali, dengan belanja sekitar Rp34,42 juta. Segmen-segmen ini dapat menjadi dasar strategi pemasaran, yaitu reaktivasi bagi pelanggan pasif, peningkatan frekuensi dan nilai belanja bagi pelanggan reguler, serta retensi bagi pelanggan bernilai tinggi. Dengan demikian, kombinasi RFM dan *clustering* terbukti dapat dipakai untuk memahami karakter pelanggan secara objektif.

Penelitian ini masih terbatas pada satu dataset *e-commerce* dan tiga fitur RFM. Penelitian lanjutan disarankan memakai data yang lebih besar dan beragam serta menambahkan variabel seperti kategori produk, lokasi pelanggan, metode pembayaran, atau tingkat diskon. Ruang parameter yang lebih luas juga layak dijajaki, terutama *eps* dan *min_samples* pada DBSCAN, bersama pembandingan dengan metode lain seperti *Hierarchical Clustering*, *Gaussian Mixture Model*, dan *Spectral Clustering*. Selain itu, dasbor yang telah dibuat dapat dikembangkan agar lebih mudah dimanfaatkan oleh pihak bisnis dan analis data.

5. REFERENSI

[1] M. Alves and G. Tobias, *for personalized customer targeting in e - commerce use cases*, vol. 21, no. 3. Springer Berlin Heidelberg, 2023. doi: 10.1007/s10257-023-00640-4.

[2] S. Anwar and E. Iskandar, "Comparison Of Clustering Algorithms In Analyzing E-Commerce Data From Kaggle," *International Journal of Social Research*, vol. 4, no. 1, pp. 38–52, Feb. 2026, doi: 10.59888/INSIGHT.V4I1.105.

[3] O. Akande, E. O. Asani, and B. Dautare, "Customer Segmentation Through RFM Analysis and K-Means Clustering: Leveraging Data-Driven Insights for Effective Marketing Strategy," *Ceddi Journal of Information System and Technology (JST)*, vol. 3, no. 1, pp. 14–25, Apr. 2024, doi: 10.56134/JST.V3I1.81.

[4] R. M. Fauzan and G. Alfian, "Segmentasi Pelanggan E-Commerce Menggunakan Fitur Recency, Frequency, Monetary (RFM) dan Algoritma Klasterisasi K-Means," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 9, no. 3, pp. 170–177, Sep. 2024, doi: 10.14421/JISKA.2024.9.3.170-177.

[5] C. G. Wong, G. K. Tong, and S. C. Haw, "Exploring Customer Segmentation in E-Commerce using RFM Analysis with Clustering Techniques," *Journal of Telecommunications and the Digital Economy*, vol. 12, no. 3, pp. 97–125, Sep. 2024, doi: 10.18080/JTDE.V12N3.978.

[6] P. Vijay Vade, "Comparative analysis of clustering algorithms for customer segmentation and improved marketing strategies," 2024. Accessed: Oct. 30, 2026. [Online]. Available: https://hdl.handle.net/10788/4566

[7] A. Nowak-Brzezińska and I. Gaibei, "How the Outliers Influence the Quality of Clustering?," *Entropy*, vol. 24, no. 7, Jul. 2022, doi: 10.3390/E24070917.

[8] B. N. Yulisasih, B. N. Yulisasih, H. Herman, S. Sunardi, and H. Yuliansyah, "Evaluation of K-

DOI : https://doi.org/10.35316/jimi.v11i2.200-210
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

Means Clustering Using Silhouette Score Method on Customer Segmentation," *ILKOM Jurnal Ilmiah*, vol. 16, no. 3, pp. 330–342, Dec. 2024, doi: 10.33096/ilkom.v16i3.2325.330-342.

[9] B. A. Pimentel, R. De Amorim Silva, and J. C. S. Costa, "Fuzzy C-Means Clustering Algorithms with Weighted Membership and Distance," *International Journal of Uncertainty, Fuzziness and Knowldege-Based Systems*, vol. 30, no. 4, pp. 567–594, Aug. 2022, doi: 10.1142/S0218488522500143.

[10] M. V. Rajesh, S. R. Chintalapudi, and M. H. M. K. Prasad, "A Comparative Analysis of RFM-based Customer Segmentation with K-Means and BIRCH Clustering Techniques," pp. 977–989, 2024, doi: 10.2991/978-94-6463-471-6_94.

[11] J. Heidari, N. Daneshpour, and A. Zangeneh, "A novel K-means and K-medoids algorithms for clustering non-spherical-shape clusters non-sensitive to outliers," *Pattern Recognit.*, vol. 155, Nov. 2024, doi: 10.1016/J.PATCOG.2024.110639.

[12] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (N. Y).*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/J.INS.2022.11.139.

[13] A. Sundari, I. S. Putra, and N. Sibuea, "Application of The RFM Model and K-Means Clustering for Customer Segmentation in E-Wallet Top-Up Services," *Infomatek*, vol. 28, no. 1, pp. 203–210, May 2026, doi: 10.23969/INFOMATEK.V28I1.42246.

[14] N. F. Fahrudin and R. Rindiyani, "Comparison of K-Medoids and K-Means Algorithms in Segmenting Customers based on RFM Criteria," *E3S Web of Conferences*, vol. 484, Feb. 2024, doi: 10.1051/E3SCONF/202448402008.

[15] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability 2022, Vol. 14, Page 7243*, vol. 14, no. 12, p. 7243, Jun. 2022, doi: 10.3390/SU14127243.

[16] V. Plotnikova, M. Dumas, and F. Milani, "Adaptations of data mining methodologies: A systematic literature review," *PeerJ Comput. Sci.*, vol. 6, pp. 1–43, May 2020, doi: 10.7717/PEERJ-CS.267/SUPP-2.

[17] A. Vouros, S. Langdell, M. Croucher, and E. Vasilaki, "An empirical comparison between stochastic and deterministic centroid initialisation for K-means variations," *Machine Learning 2021 110:8*, vol. 110, no. 8, pp. 1975–2003, Jul. 2021, doi: 10.1007/S10994-021-06021-7.

[18] S. Harris and R. C. De Amorim, "An Extensive Empirical Comparison of k-means Initialization Algorithms," *IEEE Access*, vol. 10, pp. 58752–58768, 2022, doi: 10.1109/ACCESS.2022.3179803.

[19] Z. Wu, L. Jin, J. Zhao, L. Jing, and L. Chen, "Research on Segmenting E-Commerce Customer through an Improved K-Medoids Clustering Algorithm," *Comput. Intell. Neurosci.*, vol. 2022, no. 1, p. 9930613, Jan. 2022, doi: 10.1155/2022/9930613.

[20] M. B. Ferraro, "Fuzzy k-Means: history and applications," *Econom.*

DOI : https://doi.org/10.35316/jimi.v11i2.200-210
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

P-ISSN : 2549-7480
R.Saputra dkk/ JIMI 11 (2) pp. 200-210 E-ISSN : 2549-6301

Stat., vol. 30, pp. 110–123, Apr. 2024, doi: 10.1016/J.ECOSTA.2021.11.008.

[21] R. J. G. B. Campello, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 2, p. e1343, Mar. 2020, doi: 10.1002/WIDM.1343;REQUESTEDJOURNAL:JOURNAL:19424795;PAGEGROUP:STRING:PUBLICATION.

[22] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP Journal on Wireless Communications and Networking 2021 2021:1*, vol. 2021, no. 1, pp. 31-, Feb. 2021, doi: 10.1186/S13638-021-01910-W.

[23] A. M. Bagirov, R. M. Aliguliyev, and N. Sultanova, "Finding compact and well-separated clusters: Clustering using silhouette coefficients," *Pattern Recognit.*, vol. 135, p. 109144, Mar. 2023, doi: 10.1016/J.PATCOG.2022.109144.

[24] M. Gagolewski, M. Bartoszuk, and A. Cena, "Are cluster validity measures (in) valid?," *Inf. Sci. (N. Y).*, vol. 581, pp. 620–636, Dec. 2021, doi: 10.1016/J.INS.2021.10.004.

[25] B. Mardhotillah, A. Fadli, E. Elisa, and Z. Zurweni, "Indeks Calinski Harabasz Analisis Fuzzy C Means dan K Means Cluster Kabupaten/Kota di Provinsi Jambi Menurut Potensi Pertambangan, Penggalian, Pengadaan Listrik, dan Gas," vol. 2, no. 1, pp. 16–22, Jun. 2023, doi: 10.22437/MULTIPROXIMITY.V2I1.25693.

DOI : https://doi.org/10.35316/jimi.v11i2.200-210 © 2026 Jurnal Ilmiah Informatika with CC BY NC licence