Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 4 October 2025, pp: 1377-1386
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Optimizing K-Means Performance with Hybrid RFM and Behavioral Analytics for Customer Segmentation

Optimasi Performa K-Means melalui *Hybrid Feature Engineering* RFM dan *Behavioral Analytics* untuk Segmentasi Pelanggan

Ilham B[1*], Rahmaddeni[2], Aldino Putra[3], Dimas Najario[4] , M. Aggie Fakhrizal[5]

[1,2,3,4,5]Program Studi Teknik Informatika, Universitas Sains dan Teknologi Indonesia, Indonesia

E-Mail: [1]durianbohong@gmail.com, [2]rahmaddeni@usti.ac.id , [3]2210031802078@sar.ac.id, [4]2210031802035@sar.ac.id, [5]2210031802112@sar.ac.id

Received Aug 04th 2025; Revised Sep 13th 2025; Accepted Oct 19th 2025; Available Online Oct 31th 2025
Corresponding Author: Ilham B

Abstract

Breaking down customers into segments is super important for successful marketing. In this study, we suggest a mixed method using the K-Means algorithm that's fine-tuned with Recency, Frequency, Monetary (RFM) metrics and insights from behavioral analytics. The main goal here is to see how adding these RFM features affects the quality of the segments we create. So, our approach includes cleaning up retail transaction data, creating behavioral features, and trying out two clustering methods: the regular K-Means and the enhanced RFM K-Means Aware++. We evaluated the results by looking at t-SNE visualizations, analyzing how the clusters were distributed, and checking things like the Silhouette Score and the Davies-Bouldin Index for internal validation. The findings show that the RFM-enhanced method gives us clusters that are more stable, well-separated, and representative of actual customer behavior. On the other hand, the model that skips RFM tends to have overlapping clusters and doesn't provide as meaningful segmentation. Overall, this study highlights that smart feature engineering plays a crucial role in boosting how well clustering algorithms perform.

Keyword: Behavioral Analytics, Clustering, Customer Segmentation, Feature Engineerin, K-Means, RFM.

Abstrak

Membagi pelanggan menjadi beberapa segmen itu krusial untuk kesuksesan strategi pemasaran. Dalam studi ini, kami mengusulkan metode kombinasi dengan memanfaatkan algoritma K-Means yang dipadukan dengan metrik *Recency, Frequency, Monetary* (RFM) serta wawasan dari analitik perilaku. Tujuan utama kami adalah untuk mengetahui seberapa besar pengaruh penambahan fitur RFM terhadap kualitas segmen yang dihasilkan. Untuk itu, pendekatan kami meliputi pembersihan data transaksi ritel, pembuatan fitur berbasis perilaku, dan penerapan dua metode klastering: K-Means standar dan versi yang ditingkatkan, yaitu RFM K-Means *Aware++*. Hasil klastering dievaluasi dengan visualisasi t-SNE, analisis distribusi klaster, dan pengukuran metrik validasi internal seperti Silhouette Score dan Davies-Bouldin Index. Temuan kami menunjukkan bahwa metode yang lebih baik dengan fitur RFM menghasilkan klaster yang lebih stabil, terpisah dengan baik, dan lebih akurat dalam mencerminkan perilaku pelanggan. Sebaliknya, model yang tidak menggunakan fitur RFM cenderung membentuk klaster yang tumpang tindih dan memberikan segmentasi yang kurang bermakna. Secara keseluruhan, studi ini menekankan bahwa rekayasa fitur yang tepat memiliki peran penting dalam meningkatkan performa algoritma klastering dan menawarkan segmentasi pelanggan yang lebih berharga.

Kata Kunci: *Behavioral Analytics, Clustering*, K-Means, Rekayasa Fitur, RFM, Segmentasi Pelanggan.

1. PENDAHULUAN

Segmentasi pelanggan merupakan komponen penting dalam strategi pemasaran modern yang berorientasi pada data. Dalam konteks ritel, kemampuan perusahaan untuk memahami perilaku pelanggan dan mengelompokkannya secara tepat menjadi fondasi utama bagi pengambilan keputusan berbasis data [1]. Keberhasilan strategi promosi, retensi pelanggan, serta personalisasi layanan sangat bergantung pada efektivitas model segmentasi yang digunakan [2].

Algoritma K-Means telah menjadi salah satu metode yang paling populer dalam *unsupervised learning* karena kesederhanaan dan efisiensinya dalam mengelompokkan data ke dalam klaster homogen [3]. Namun demikian, K-Means konvensional memiliki keterbatasan dalam menangkap kompleksitas perilaku pelanggan yang dinamis, terutama ketika hanya mengandalkan fitur numerik standar tanpa mempertimbangkan dimensi waktu, nilai, dan frekuensi transaksi [4]. Dalam banyak kasus, model ini gagal membedakan pelanggan yang memiliki nilai ekonomi tinggi namun dengan frekuensi transaksi rendah, atau sebaliknya.

Untuk mengatasi hal tersebut, pendekatan berbasis *feature engineering* seperti *Recency, Frequency, Monetary* (RFM) dikembangkan guna memperkaya representasi data pelanggan. RFM memungkinkan analisis yang lebih bermakna dengan mempertimbangkan kapan terakhir pelanggan bertransaksi, seberapa sering mereka melakukan pembelian, dan berapa besar nilai transaksi yang dihasilkan [5]. Beberapa penelitian juga menunjukkan bahwa penggabungan metode RFM dengan algoritma K-Means dapat menghasilkan segmentasi pelanggan yang lebih stabil dan mudah diinterpretasikan dalam konteks pemasaran digital [6].

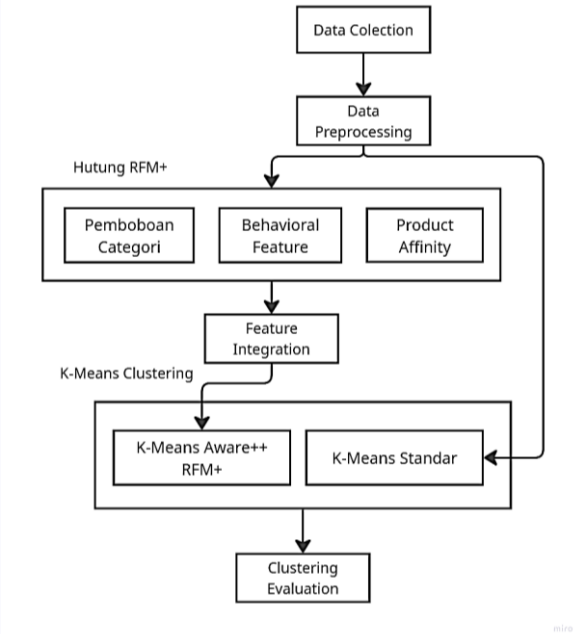
Meski begitu, sebagian besar penelitian sebelumnya masih terbatas pada integrasi fitur RFM tanpa memperhitungkan dimensi perilaku lain seperti sensitivitas terhadap diskon, waktu pembelian, atau preferensi lokasi. Studi oleh Rahmadhan (2022) [7] mengusulkan *Hybrid* K-Means berbasis nilai pelanggan (*Customer Lifetime Value*), namun belum mengintegrasikan aspek spasial maupun perilaku pelanggan secara menyeluruh.

Penelitian lain yang dilakukan oleh Ho et al. (2023) [4] dan Ufeli et al. (2025) [5] menunjukkan bahwa integrasi fitur non-transaksional dapat meningkatkan interpretabilitas hasil klasterisasi. Namun, masih terdapat *research gap* dalam penerapan pendekatan *hybrid feature engineering* yang mengombinasikan RFM dengan *behavioral analytics* secara sistematis, khususnya dengan inisialisasi klaster berbasis kepadatan (*Density-Aware* K-Means++).

Berdasarkan kondisi tersebut, penelitian ini berfokus pada pengembangan model K-Means *Aware++* yang mengintegrasikan rekayasa fitur hibrida berbasis RFM dan analitik perilaku pelanggan. Tujuannya adalah untuk mengevaluasi sejauh mana integrasi tersebut dapat meningkatkan kualitas hasil klasterisasi dibandingkan model K-Means standar. Evaluasi dilakukan melalui pendekatan visual menggunakan *t-distributed Stochastic Neighbor Embedding* (t-SNE) dan pendekatan kuantitatif menggunakan *Silhouette Score serta Davies-Bouldin Index* (DBI) untuk menilai separabilitas, stabilitas, dan konsistensi hasil klaster.

2. METODOLOGI PENELITIAN

Penelitian ini menerapkan pendekatan kuantitatif dengan pengolahan data sistematis untuk membangun segmentasi pelanggan berbasis *hybrid feature engineering*. Data transaksi daring internasional diambil dari Kaggle, mencakup atribut seperti ID pelanggan, produk, jumlah, harga, waktu transaksi, metode pembayaran, lokasi toko, kategori produk, diskon, dan total belanja. Fokus utama terletak pada integrasi dimensi transaksional, spasial, temporal, dan perilaku belanja guna membentuk fitur pelanggan yang komprehensif. Metodelogi penelitian dapat dilihat pada Gambar 1.



Gambar 1. Metodelogi Penelitian

2.1. *Data Collection*

Data collection merupakan proses sistematis dalam menghimpun informasi relevan untuk mendukung analisis dan pemodelan. Penelitian ini menggunakan *dataset* transaksi ritel fisik publik dari platform Kaggle, yang mencerminkan perilaku belanja pelanggan pada periode 29 April 2023 hingga 28 April 2024. *Dataset* mencakup atribut waktu transaksi, lokasi toko, produk, nilai belanja, dan variabel perilaku pelanggan lainnya. Total terdapat 95.215 pelanggan unik, 100.000 transaksi, dan lebih dari 500.000 interaksi perilaku. Jumlah dan keragaman data tersebut dinilai memadai untuk mendukung segmentasi berbasis perilaku dan spasio-temporal.

2.2. *Data Preprocessing*

Data preprocessing merupakan tahap krusial yang bertujuan membersihkan dan mentransformasi data mentah agar layak digunakan dalam analisis statistik dan pemodelan [3]. Proses ini mencakup penghapusan duplikasi, koreksi entri tidak valid, serta deteksi dan penanganan *outlier* menggunakan metode *Interquartile Range* (IQR). Tahapan ini sangat berpengaruh terhadap akurasi segmentasi dan validitas keputusan berbasis data. Implementasinya mencakup normalisasi skala menggunakan *RobustScaler* pada fitur transaksional dan *MinMaxScaler* pada fitur perilaku agar semua variabel memiliki skala seragam. Format waktu diselaraskan ke *UNIX time* untuk memudahkan sinkronisasi spasio-temporal. Transaksi yang terjadi pada hari Sabtu dan Minggu diberi bobot tambahan (×1.5) untuk menangkap kecenderungan pola belanja mingguan yang signifikan.

2.3. *Hybrid Feature Engineering*

Hybrid feature engineering merupakan inti dari metodologi ini, di mana fitur transaksional, perilaku pelanggan, dan spasial digabungkan untuk membentuk representasi pelanggan yang lebih kaya. Proses diawali dengan perhitungan RFM-*Plus*, yakni pengembangan dari pendekatan RFM klasik. Fitur tambahan meliputi *store affinity, discount sensitivity, temporal shopping pattern*, dan *product affinity*, sebagaimana pendekatan mendalam yang dikembangkan oleh Agustin et a[4].

Komponen *Recency+* dihitung berdasarkan hari sejak transaksi terakhir dengan peluruhan eksponensial ($\lambda = 0.05$), *Frequency+* mencerminkan jumlah transaksi enam bulan terakhir dengan bobot tambahan untuk transaksi akhir pekan, dan *Monetary+* dihitung dari total nilai transaksi yang dikalibrasi berdasarkan bobot lokasi toko. Setiap toko dikategorikan menjadi *Premium Mall, Downtown,* atau *Suburban*, dengan bobot masing-masing 1.3, 1.2, dan 0.9. Klasifikasi dilakukan melalui pendekatan *hybrid* yang menggabungkan *rule-based* dan *machine learning* [5].

Preferensi lokasi (*store affinity*) dihitung menggunakan indeks entropi berdasarkan keragaman kunjungan pelanggan ke toko. Sementara itu, *discount sensitivity* diukur dari rasio transaksi yang melibatkan diskon, dan temporal pattern dihitung dari persentase pembelian pada jam sibuk (16:00–20:00). Untuk menggambarkan afinitas produk, dibentuk *product affinity matrix* berdasarkan frekuensi pembelian tiap kategori, kemudian direduksi menggunakan *Principal Component Analysis* (PCA) menjadi tiga dimensi. Seluruh fitur ini kemudian digabungkan ke dalam satu vektor fitur hibrida sebagai masukan utama dalam proses *clustering*[6].

2.4. *Feature Integration*

Feature integration adalah proses penggabungan seluruh fitur yang telah direkayasa dari berbagai dimensi ke dalam satu kerangka data yang utuh dan konsisten. Langkah ini penting agar model *clustering* dapat memanfaatkan informasi komprehensif dari berbagai sumber fitur secara bersamaan. Dalam penelitian ini, seluruh fitur yang telah dikembangkan seperti RFM, bobot lokasi, sensitivitas diskon, dan pola waktu belanja digabungkan ke dalam matriks fitur akhir. Struktur data yang telah terintegrasi ini menjadi *input* utama dalam proses *clustering*, sehingga dapat menghasilkan segmentasi yang lebih akurat dan *actionable*[7].

2.5. *K-Means Clustering*

Algoritma K-Means *Clustering* merupakan salah satu metode *unsupervised learning* yang paling populer dan efisien dalam analisis segmentasi pelanggan karena kemampuannya mengelompokkan data berdasarkan kesamaan atribut numerik dengan kompleksitas yang relatif rendah [1], [4], [8]. Secara umum, tujuan utama algoritma ini adalah meminimalkan total jarak antara setiap data dengan pusat klaster (*centroid*) yang menaunginya, sebagaimana diformulasikan dalam fungsi objektif berikut:

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

Dengan: k adalah Jumlah klaster, x adalah vektor data ke j, C_i adalah himpunan anggota klaster ke i, dan μ_i adalah *centroid* (titik rata-rata) dari klaster ke i.

Proses klasterisasi berlangsung secara iteratif, dimulai dari inisialisasi *centroid* awal, perhitungan jarak *Euclidean* antara setiap data dengan seluruh *centroid*, hingga pembaruan posisi *centroid* berdasarkan rata-rata anggota klasternya. Proses berulang hingga perubahan posisi *centroid* mencapai konvergensi atau batas iterasi tertentu [15], [17].

2.6. *Clustering Evaluation*

Evaluasi klasterisasi dalam penelitian ini bertujuan untuk menilai kualitas dan konsistensi segmentasi yang dihasilkan dari dua pendekatan algoritmik: K-Means standar dan K-Means *Aware*++ yang diperkaya dengan fitur RFM dan *Behavioral Analytics*. Proses evaluasi dilakukan melalui pendekatan visual dan kuantitatif. Secara visual, digunakan algoritma t-SNE untuk memproyeksikan data berdimensi tinggi ke dalam dua dimensi, sehingga struktur dan pemisahan antar klaster dapat diamati secara intuitif. Di sisi lain, evaluasi kuantitatif dilakukan dengan membandingkan distribusi klaster silang antara kedua model, guna mengidentifikasi kestabilan serta sensitivitas klaster terhadap perubahan fitur. Evaluasi ini penting untuk memastikan bahwa metode segmentasi tidak hanya efektif secara teknis, tetapi juga mampu menghasilkan struktur klaster yang konsisten dan representatif terhadap perilaku pelanggan[11], [12].

2.7. *Literature Review*

Berbagai penelitian sebelumnya telah mengkaji penerapan algoritma K-Means dan metode RFM dalam segmentasi pelanggan untuk meningkatkan efektivitas strategi pemasaran berbasis data. Lathifah (2025) [1] mengembangkan model *AI-Driven Customer Segmentation* menggunakan K-Means untuk mengelompokkan pelanggan berdasarkan perilaku transaksi. Pendekatan ini menunjukkan bahwa *machine learning* dapat meningkatkan akurasi segmentasi pelanggan, namun masih terbatas pada dimensi perilaku transaksi tanpa mempertimbangkan nilai ekonomi pelanggan.

Penelitian Alzami et al. (2023) [2] memperluas pendekatan tersebut dengan mengintegrasikan metode RFM ke dalam algoritma K-Means dalam konteks *e-commerce*. Hasilnya menunjukkan peningkatan interpretabilitas hasil klasterisasi, tetapi model masih berfokus pada fitur transaksional murni dan belum memasukkan variabel perilaku yang lebih kompleks.

Selanjutnya, Ho et al. (2023) [4] mengembangkan model *Extended* RFM dengan menambahkan dimensi demografis dan perilaku pelanggan, yang terbukti meningkatkan ketajaman segmentasi meskipun menambah kompleksitas dalam perancangan fitur. Sejalan dengan itu, Ufeli et al. (2025) [5] menggunakan pendekatan *Factor Analysis of Mixed Data* (FAMD) untuk menggabungkan variabel kuantitatif dan kualitatif sebelum dilakukan klasterisasi menggunakan K-Means dan hierarchical *clustering*, menunjukkan pentingnya representasi fitur campuran dalam menjaga stabilitas segmentasi pelanggan.

Sementara itu, Rahmadhan (2022) [7] mengusulkan model *Hybrid* K-Means berbasis *Customer Lifetime Value* (CLV) yang memperkenalkan dimensi temporal untuk mengukur nilai pelanggan jangka panjang. Pendekatan ini membuka peluang bagi integrasi antara rekayasa fitur RFM dan *behavioral analytics* dalam membangun model klasterisasi yang lebih adaptif terhadap perubahan perilaku pelanggan.

Secara keseluruhan, hasil-hasil tersebut menunjukkan bahwa kombinasi antara RFM dan K-Means mampu meningkatkan kualitas segmentasi pelanggan. Namun, mayoritas penelitian sebelumnya masih terbatas pada pemrosesan fitur transaksional dan demografis tanpa memperhitungkan distribusi kepadatan data maupun perilaku pelanggan yang dinamis. Berdasarkan celah tersebut, penelitian ini mengusulkan model *Density-Aware* K-Means++ sebagai pendekatan yang mengintegrasikan *feature engineering* berbasis RFM dan analitik perilaku untuk menghasilkan segmentasi yang lebih stabil, representatif, dan kontekstual terhadap perilaku pelanggan dalam ekosistem ritel modern.

3. HASIL DAN PEMBAHASAN

3.1. *Data Collection*

Penelitian ini menggunakan data publik yang diambil dari platform Kaggle, yang merepresentasikan transaksi ritel fisik dalam skala besar. *Dataset* terdiri dari lebih dari 95.000 pelanggan unik dan 100.000 transaksi, mencakup variasi produk, jumlah pembelian, harga, metode pembayaran, waktu transaksi, dan lokasi toko. Struktur data ini dinilai cukup representatif untuk dianalisis dalam konteks segmentasi pelanggan berbasis perilaku dan spasio-temporal. Untuk keperluan analisis, data diseleksi dan disesuaikan agar fokus pada informasi yang relevan dengan proses segmentasi. Sampel struktur data yang digunakan ditunjukkan pada Tabel 1.

3.2. *Data Preprocessing*

Pada tahap awal, data disaring untuk memfokuskan analisis pada perilaku terkini. Dari 100.000 transaksi, diambil 20.000 transaksi terbaru berdasarkan waktu, agar segmentasi mencerminkan kondisi pembelian saat ini. Data tersebut dipilih karena dianggap cukup mewakili distribusi pelanggan aktif dan tetap relevan meski akan menyusut setelah proses pembersihan.

Penanganan *missing values* dilakukan dengan imputasi median pada fitur numerik, serta interpolasi temporal pada fitur waktu dan lokasi untuk menjaga kontinuitas data. Normalisasi skala dilakukan dengan *RobustScaler* pada fitur transaksional dan *MinMaxScaler* pada fitur perilaku untuk menyamakan rentang nilai antar fitur**.** Seluruh waktu transaksi diselaraskan ke *UNIX time* untuk integrasi spasio-temporal yang konsisten. *Outlier* seperti harga nol, jumlah produk ekstrem, dan lokasi di luar wilayah operasional dihapus menggunakan metode IQR. Transaksi akhir pekan diberikan bobot 1.5 untuk menangkap kecenderungan belanja yang lebih tinggi di hari Sabtu dan Minggu, sehingga memperkuat sensitivitas model terhadap pola mingguan pelanggan.

Table 1. Data Transaksi Pelanggan

No.	*Customer Id*	*Product Id*	*Qty*	*Price*	*Transaction Date*	*Payment Method*	*Store Location*	*Total Amount*
1	738026	C	9	70.8	4/28/2024 22:22	Debit Card	8717 Brown Mountain Apt. 590 East Heather, IL 11437	622.0
2	460190	C	6	64.2	4/28/2024 22:20	Credit Card	819 Lauren Ranch Aaronmouth, GA 12191	309.1
...	...	...	...	...	...	...	...	...
100.000	708242	D	4	53.5	4/29/2023 22:27	Credit Card	PSC 3757, Box 1899 APO AA 21818	176.1

3.3. *Hybrid Feature Engineering*

Proses dimulai dengan penghitungan RFM-*Plus* (lihat Tabel 2), yaitu pengembangan dari model klasik RFM yang telah disesuaikan dengan konteks spasial dan perilaku.

1. *Recency+* dihitung dengan fungsi *eksponensial decay* ($\lambda = 0{,}05$) untuk memberi bobot lebih pada transaksi terkini (Peter S. Fader faderp, Bruce G.S. Hardie bhardie, 2005).
2. *Frequency+* dihitung berdasarkan transaksi enam bulan terakhir, dengan bobot tambahan (×1,5) untuk transaksi akhir pekan.
3. *Monetary+* dihitung dari total nilai transaksi yang dikalibrasi menggunakan bobot lokasi, berdasarkan kategori strategis took.

Table 2. Hasil Hitung RFM *Plus*

No	*CustomerID*	*RecencyPlus*	*FrequencyPlus*	*MonetaryPlus*
0	59	0.197899	1.0	98.892.594
1	96	0.267135	1.5	194.356.816
2	135	0.339596	1.0	20.667.466
3	202	0.618783	1.5	174.726.716
4	289	0.115325	1.5	275.857.019
...	...	...	...	...
9937	999860	1.000.000	1.5	419.004.117

Selanjutnya, dilakukan ekstraksi fitur perilaku untuk memperkaya dimensi pemodelan:

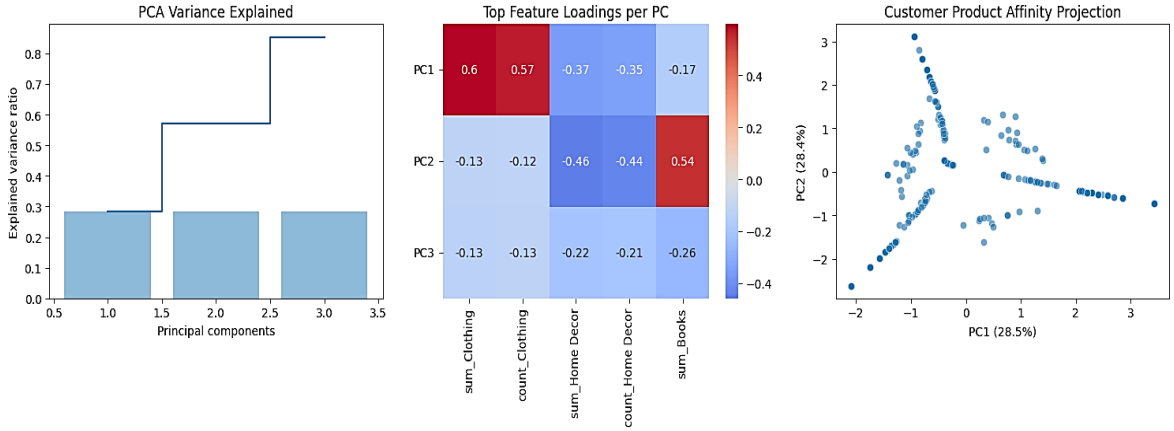
1. *Store Affinity* dihitung menggunakan indeks entropi berdasarkan distribusi lokasi toko tempat pelanggan bertransaksi. Nilai rendah menunjukkan konsistensi tinggi (*loyalty*) terhadap lokasi tertentu.
2. *Discount Sensitivity* dihitung dari rasio transaksi dengan diskon terhadap total transaksi, menggambarkan sensitivitas pelanggan terhadap promo.
3. *Temporal Pattern* dihitung sebagai proporsi transaksi yang terjadi pada jam sibuk (*peak hour*), yaitu pukul 16.00–20.00.
4. *Weekend Ratio* mencerminkan proporsi transaksi pelanggan yang terjadi di akhir pekan.
5. *Category Diversity* diukur dari rasio jumlah kategori produk unik terhadap total produk yang dibeli, mencerminkan keragaman preferensi produk pelanggan.

Preferensi pelanggan terhadap kategori produk kemudian dimodelkan dalam bentuk *Product Affinity Matrix* (lihat Tabel 3), yang direduksi dimensinya menjadi tiga menggunakan PCA untuk efisiensi pemrosesan tanpa kehilangan informasi representatif.

Table 3. Matriks Afinitas Produk

CustomerId	PCA 1	PCA 2	PCA 3
59	-1.289.946	-1.614.339	-0.756865
96	2.335.848	-0.598976	-0.708634
...	...	...	...
999860	-1.481.220	-1.607.859	-0.745700

Akhirnya, semua fitur – RFM-*Plus*, *store affinity*, *discount sensitivity*, *temporal pattern*, *weekend ratio*, *category diversity*, dan *product embedding* – digabungkan ke dalam satu vektor fitur hibrida. Representasi ini memperkaya segmentasi pelanggan tidak hanya dari sisi frekuensi pembelian, tetapi juga dari aspek perilaku, preferensi produk, dan dimensi temporal secara simultan, sehingga menghasilkan model segmentasi yang lebih akurat dan kontekstual. Gambar 2 merupakan visualisasi *hybrid feature engineering*.



Gambar 2. Visualisasi *Hybrid Feature Engineering*

Semua fitur (RFM-*Plus*, *store affinity*, *discount sensitivity*, *temporal pattern*, dan *product embedding*) kemudian digabung menjadi satu vektor fitur hibrida. Representasi ini memperkaya segmentasi pelanggan dengan dimensi perilaku, spasial, dan temporal secara bersamaan.

3.4. *Feature Integration*

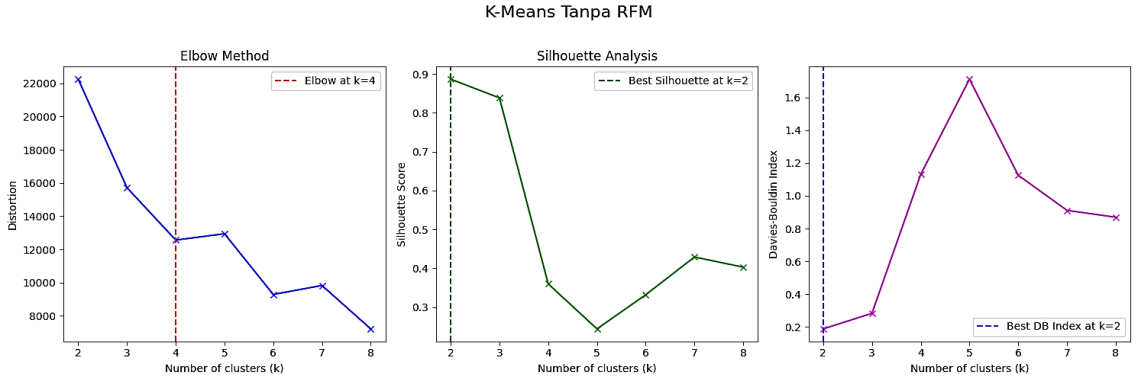
Setelah seluruh fitur utama—termasuk RFM-*Plus*, klasifikasi lokasi, *store affinity*, *discount sensitivity*, *temporal pattern*, *dan product embedding* berhasil dihitung melalui proses *hybrid feature engineering*, langkah selanjutnya adalah mengintegrasikannya ke dalam satu *dataset* terpadu. Proses ini bertujuan menyatukan dimensi spasial, temporal, dan perilaku pelanggan agar setiap entri mewakili satu vektor fitur lengkap yang siap digunakan dalam proses *clustering*. Dengan demikian, setiap pelanggan direpresentasikan secara holistik berdasarkan data yang telah terstandarisasi dan relevan secara analitik.

3.5. K-Means *Clustering*

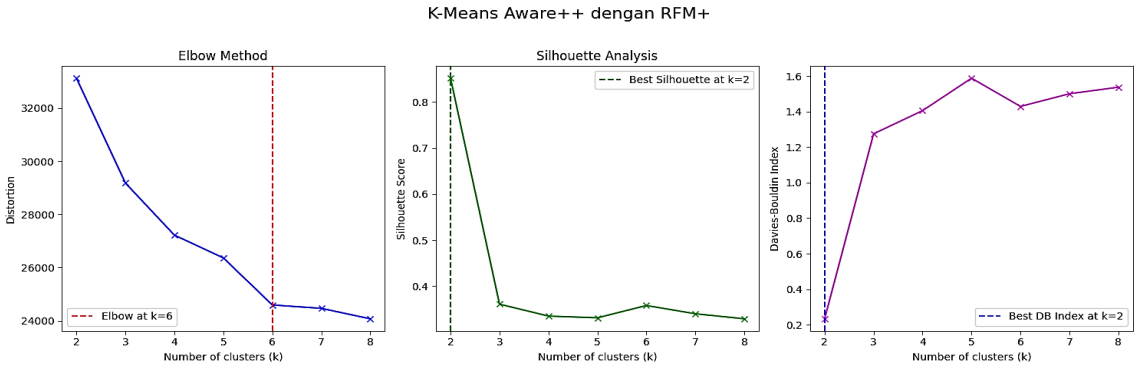
Algoritma K-Means digunakan untuk melakukan segmentasi pelanggan berdasarkan data perilaku dan fitur RFM ini selaras dengan [13]. Proses ini diawali dengan menentukan jumlah klaster optimal menggunakan tiga metrik evaluasi internal: *Elbow Method*, *Silhouette Score*, dan *Davies-Bouldin Index*. Dua pendekatan dibandingkan: K-Means standar dan versi yang telah ditingkatkan dengan fitur RFM dan pemodelan berbasis densita[14], [15].

Gambar 3 memperlihatkan hasil evaluasi K-Means standar terhadap nilai k dari 2 hingga 8. Grafik pertama menunjukkan nilai distortion yang digunakan dalam *Elbow Method* untuk mengidentifikasi titik infleksi. Pada grafik ini, terjadi penurunan yang signifikan hingga k = 2, yang ditandai dengan garis vertikal sebagai titik *elbow*. Grafik kedua memperlihatkan nilai Silhouette Score yang mengukur sejauh mana data dalam satu klaster memiliki kemiripan internal, dengan nilai tertinggi juga ditemukan pada k = 2. Grafik ketiga menunjukkan nilai *Davies-Bouldin Index*, yang lebih rendah menunjukkan pemisahan klaster yang lebih baik; nilai minimum juga terjadi pada k = 2 [16].

Evaluasi kemudian dilanjutkan dengan pendekatan K-Means yang dioptimasi menggunakan fitur *hybrid* RFM+ serta pembobotan berbasis densitas. Grafik pertama dalam Gambar 4 menunjukkan bahwa penurunan nilai distortion lebih gradual dibandingkan pendekatan sebelumnya, dengan titik *elbow* juga muncul pada k = 2. Grafik kedua menunjukkan bahwa nilai *Silhouette Score* tertinggi juga terdapat pada k = 2. Grafik ketiga menampilkan nilai *Davies-Bouldin Index* yang mencapai minimum pada k = 2, mengindikasikan konsistensi hasil antara semua *metric* [17], [18].

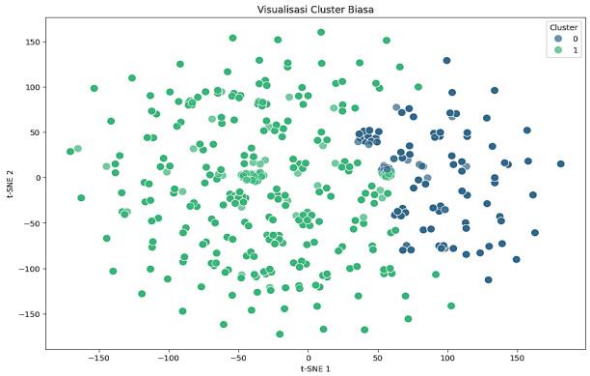


Gambar 3. Evaluasi K-Means Standar menggunakan *Elbow*, *Silhouette*, dan *DB Index*.



Gambar 4. Evaluasi K-Means dengan Fitur RFM+ menggunakan tiga metrik evaluasi internal.

Setelah nilai k optimal diperoleh, visualisasi hasil klasterisasi dilakukan dengan menggunakan teknik t-SNE untuk memproyeksikan data berdimensi tinggi ke dalam dua dimensi. Gambar 5 menunjukkan distribusi klaster dari hasil K-Means standar dengan k = 2, di mana tampak bahwa sebaran data antar klaster masih cukup berdekatan [19].



Gambar 5. Visualisasi Hasil Klasterisasi K-Means Standar (k = 2) dengan t-SNE.

Sebaliknya, hasil visualisasi dari pendekatan RFM+ ditampilkan pada Gambar 6 Tampak bahwa kedua klaster terbentuk dengan batas yang lebih jelas dan pemisahan yang lebih tegas, mengindikasikan bahwa penggunaan fitur RFM dan pembobotan densitas mampu meningkatkan separabilitas antar klaster.

3.6. *Clustering Evaluation*

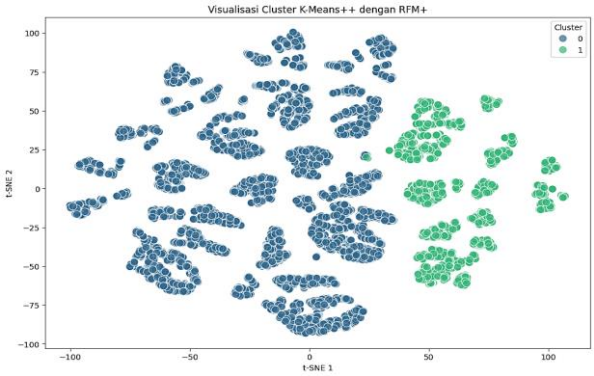
Evaluasi terhadap hasil klasterisasi dilakukan untuk membandingkan performa segmentasi antara model K-Means standar (tanpa fitur RFM) dan model K-Means *Aware++* yang disertai dengan rekayasa fitur berbasis RFM dan *Behavioral Analytics* [20]. Visualisasi dua dimensi menggunakan algoritma t-SNE ditampilkan pada Gambar 5 dan Gambar 6 untuk memberikan gambaran intuitif atas struktur klaster yang terbentuk.

Pada visualisasi hasil klasterisasi tanpa menggunakan fitur RFM (Gambar 5), distribusi data tampak acak dan tidak membentuk struktur spasial yang terpisah secara tegas. Beberapa klaster menunjukkan

tumpang tindih dan tidak adanya pembentukan batas klaster yang jelas. Hal ini mengindikasikan bahwa model tanpa fitur RFM mengalami kesulitan dalam membedakan karakteristik pelanggan secara mendalam[21].

Sebaliknya, pada hasil klasterisasi yang menggunakan pendekatan K-Means Aware++ dengan penambahan fitur RFM (Gambar 6), tampak formasi klaster yang lebih padat, terpisah, dan memiliki struktur yang lebih rapi. Klaster-klaster dalam visualisasi ini terbentuk secara lebih jelas dan menunjukkan pemisahan spasial yang signifikan [19], [22]. Hal ini memberikan indikasi kuat bahwa integrasi fitur RFM mampu meningkatkan representasi perilaku pelanggan dalam ruang fitur, yang berdampak pada pembentukan klaster yang lebih stabil dan informatif [23].

Evaluasi lanjutan dilakukan melalui analisis distribusi klaster antar dua model, yang disajikan dalam Tabel 4. Tabel ini menampilkan jumlah pelanggan pada masing-masing kombinasi klaster dari model dengan dan tanpa RFM.



Gambar 6. Visualisasi Hasil Klasterisasi K-Means RFM+ (k = 2) dengan t-SNE.

Table 4. Matriks Distribusi *Cluster_non*RFM terhadap *Cluster*_RFM

	*Cluster_non*RFM = 0	*Cluster_non*RFM = 1
*Cluster*_RFM = 0	0	7.559 pelanggan
*Cluster*_RFM = 1	2.379 pelanggan	0

Berdasarkan tabel 4, terlihat bahwa klaster yang terbentuk menggunakan fitur RFM mengalami pembalikan total ketika fitur tersebut dihilangkan. Seluruh pelanggan yang tergolong ke dalam klaster 0 pada model berbasis RFM berpindah sepenuhnya ke klaster 1 pada model tanpa RFM. Begitu pula sebaliknya, pelanggan pada klaster 1 (berbasis RFM) masuk sepenuhnya ke klaster 0 pada model tanpa RFM [24]. Fenomena ini mencerminkan terjadinya inversi klaster yang hampir sempurna [25].

Ketidakstabilan ini menandakan bahwa model tanpa RFM gagal mempertahankan struktur segmentasi pelanggan yang bermakna secara perilaku. Tidak ditemukan minor *cluster* yang hilang karena jumlah klaster dibatasi dua, namun perubahan ekstrem distribusi ini menunjukkan bahwa fitur RFM memainkan peran penting dalam menjaga stabilitas dan konsistensi pemisahan antar kelompok pelanggan.

Dari visualisasi dan distribusi ini, dapat disimpulkan bahwa penambahan fitur RFM secara signifikan meningkatkan kualitas segmentasi pelanggan, baik secara visual maupun struktural. Model dengan fitur RFM berhasil mengurangi *overlap semantik* antar pelanggan serta memberikan klasterisasi yang lebih representatif terhadap perilaku aktual pelanggan dalam konteks bisnis ritel[26].

4. DISKUSI

Hasil penelitian menunjukkan bahwa penggunaan fitur RFM dalam proses klasterisasi pelanggan memberikan dampak signifikan terhadap kualitas segmentasi. Perbandingan antara model K-Means standar dan K-Means Aware++ mengindikasikan adanya peningkatan yang jelas baik dalam aspek visualisasi maupun distribusi anggota klaster. Visualisasi dua dimensi dengan t-SNE memperlihatkan bahwa klaster yang dibentuk oleh K-Means *Aware++* tampak lebih terpisah dan stabil dibandingkan model standar yang cenderung tumpang tindih dan tidak memiliki struktur spasial yang jelas. Hal ini menandakan bahwa rekayasa fitur berbasis perilaku pelanggan mampu meningkatkan representasi data dalam ruang fitur yang digunakan algoritma K-Means.

Distribusi klaster antara kedua model juga menunjukkan fenomena inversi yang ekstrem. Pelanggan yang termasuk dalam klaster tertentu pada model RFM berpindah secara menyeluruh ke klaster berbeda pada model non-RFM. Hal ini mencerminkan ketergantungan tinggi struktur klaster terhadap fitur yang digunakan dalam pembentukan representasi. Ketika dimensi perilaku dihilangkan, model gagal mempertahankan

segmentasi yang bermakna secara bisnis. Artinya, informasi transaksi historis yang tercermin dalam fitur RFM bersifat krusial dalam membedakan kelompok pelanggan dengan perilaku berbeda.

Fenomena ini sejalan dengan literatur terdahulu yang menyatakan bahwa kualitas fitur memegang peranan penting dalam proses klasterisasi. Model pembelajaran tanpa fitur yang relevan berisiko menghasilkan klaster artifisial yang tidak mewakili pola nyata dalam data. Dalam konteks bisnis, hal ini berpotensi menurunkan efektivitas strategi pemasaran berbasis segmentasi karena kampanye tidak lagi diarahkan pada kelompok yang homogen dalam perilaku dan nilai ekonomis.

Secara konseptual, pendekatan K-Means *Aware++* memberikan arah baru dalam pengembangan segmentasi pelanggan berbasis data. Tidak hanya mengandalkan pemisahan spasial, tetapi juga memperhatikan representasi fitur yang bermakna dari sisi perilaku konsumen. Implikasi praktisnya adalah bahwa strategi rekayasa fitur yang kontekstual dan berbasis domain sangat menentukan keberhasilan model dalam menghasilkan klaster yang dapat ditindaklanjuti secara bisnis.

5. KESIMPULAN

Penelitian ini membuktikan bahwa integrasi rekayasa fitur berbasis RFM dan *Behavioral Analytics* secara nyata meningkatkan performa segmentasi pelanggan dengan algoritma K-Means. Hasil perbandingan antara model standar dan model yang ditingkatkan (K-Means *Aware++*) menunjukkan peningkatan signifikan dalam kualitas klaster ketika dimensi perilaku pelanggan dimasukkan dalam pembentukan fitur. Evaluasi visual menggunakan t-SNE memperlihatkan bahwa klaster yang dihasilkan dari model berbasis RFM lebih terpisah, padat, dan terstruktur dibandingkan model tanpa RFM, sedangkan analisis distribusi klaster menunjukkan adanya pergeseran struktural yang menandakan peran penting RFM dalam membedakan perilaku pelanggan.

Dari sisi metodologis, pendekatan hibrida antara rekayasa fitur dan algoritma *unsupervised learning* terbukti efektif dalam menangkap variasi perilaku pelanggan dan memperkaya konteks segmentasi. Integrasi RFM tidak hanya berkontribusi terhadap stabilitas dan interpretabilitas hasil klasterisasi, tetapi juga meningkatkan relevansi bisnis karena mampu menghasilkan kelompok pelanggan yang lebih representatif terhadap perilaku aktual. Dengan demikian, penelitian ini menegaskan bahwa kualitas segmentasi pelanggan sangat ditentukan oleh kedalaman informasi perilaku yang terintegrasi dalam fitur.

Hasil temuan ini dapat menjadi dasar bagi pengembangan sistem segmentasi pelanggan yang lebih presisi dan adaptif, khususnya dalam implementasi strategi pemasaran berbasis data dan personalisasi layanan. Meskipun demikian, penelitian ini memiliki keterbatasan pada ruang lingkup data yang masih berfokus pada transaksi ritel offline dan belum mengakomodasi interaksi pelanggan lintas kanal seperti *e-commerce* atau media sosial. Penelitian lanjutan diharapkan dapat mengembangkan pendekatan serupa dengan memasukkan dimensi perilaku omnichannel, serta mengeksplorasi integrasi metode berbasis *deep clustering* untuk meningkatkan kemampuan adaptasi model terhadap dinamika perilaku pelanggan di masa mendatang.

REFERENSI

[1] Z. F. A. Syfa Nur Lathifah, "AI-Driven Customers Segmentation Using K-Means Clustering," vol. 9, 2025, [Online]. Available: https://doi.org/10.70609/gtech.v9i1.6202

[2] F. Alzami *et al.*, "Implementation of RFM Method and K-Means Algorithm for Customer Segmentation in E-Commerce with Streamlit," *Ilk. J. Ilm.*, vol. 15, no. 1, pp. 32–44, 2023, doi: 10.33096/ilkom.v15i1.1524.32-44.

[3] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, "Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 273–281, 2024, doi: 10.57152/malcom.v4i1.1085.

[4] T. Ho, S. Nguyen, H. Nguyen, N. Nguyen, D. S. Man, and T. G. Le, "An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry," *Bus. Syst. Res.*, vol. 14, no. 1, pp. 26–53, 2023, doi: 10.2478/bsrj-2023-0002.

[5] C. P. Ufeli, M. U. Sattar, R. Hasan, and S. Mahmood, "Enhancing Customer Segmentation Through Factor Analysis of Mixed Data (FAMD)-Based Approach Using K-Means and Hierarchical Clustering Algorithms," *Information*, vol. 16, no. 6, p. 441, 2025, doi: 10.3390/info16060441.

[6] V. Holý and O. Sokol, "Clustering retail products based on customer behaviour," vol. 60, 2017, [Online]. Available: https://doi.org/10.1016/j.asoc.2017.02.004

[7] M. W. Rahmadhan, Radit, "Segmentation using Customers Lifetime Value: Hybrid K-means Clustering and Analytic Hierarchy Process," vol. 8, 2022, [Online]. Available: https://doi.org/10.20473/jisebi.8.2.130-141

[8] G. Ramkumar, J. Bhuvaneswari, S. Venugopal, S. Kumar, C. K. Ramasamy, and R. Karthick, "Enhancing customer segmentation: RFM analysis and K-Means clustering implementation," *Hybrid Adv. Technol.*, pp. 70–76, 2025, doi: 10.1201/9781003559139-9.

[9] A. M. A. Zamil and T. G. Vasista, "Customer Segmentation Using RFM Analysis: Realizing Through Python Implementation," *Pacific Bus. Rev. Int.*, vol. 13, no. 11, pp. 24-36 WE-Emerging Sources Citation Index (ESCI), 2021.

[10] A. Julian, A. Faqih, and G. Dwilestari, "Segmentasi Pelanggan Menggunakan Algoritma K-Means Pada Jaringan Telekomunikasi Untuk Optimalisasi Strategi Pemasaran," *J. Teknol. Terpadu*, vol. 13, no. 1, pp. 94–100, 2025.

[11] N. Suhartanto, S. Wijoyo, and W. Purnomo, "Strategi Peningkatan SDM Berdasarkan Pengelompokan Kualitas Kinerja Pegawai CV Mediatama Perkasa Bogor Menggunakan K-Means Clustering," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 5, pp. 2548–964, 2025, [Online]. Available: http://j-ptiik.ub.ac.id

[12] S.-C. H. Wong Chun-Gee, Gee-Kok Tong, "Exploring Customer Segmentation in E-Commerce using RFM Analysis with Clustering Techniques," vol. 12, 2024, [Online]. Available: https://doi.org/10.18080/jtde.v12n3.978

[13] E. W. Agustin, K. Uthami, and A. I. Ulfa, "Optimization of Customer Segmentation in the Retail Industry Using the K-Medoid Algorithm," vol. 5, no. July, pp. 766–775, 2025.

[14] H. Safitri, S. P. L. Geni, F. Merry, M. Wati, and Haviluddin, "Penerapan K-Means Clustering untuk Segmentasi Konsumen E-Commerce Penerapan K-Means Clustering untuk Segmentasi Konsumen E-Commerce Berdasarkan Pola Pembelian," *JUKI J. Komput. dan Inform.*, vol. 7, pp. 89–99, 2025.

[15] Ahmed Mohamed Ahmed Serwah, K. W. KHAW, Cheang Sharon Peck Yeng, and Alhamzah Alnoor, "Customer analytics for online retailers using weighted k-means and RFM analysis," *Data Anal. Appl. Math.*, vol. 4, no. 1, pp. 1–7, 2023, doi: 10.15282/daam.v4i1.9171.

[16] J. M. John, O. Shobayo, and B. Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," *Analytics*, vol. 2, no. 4, pp. 809–823, 2023, doi: 10.3390/analytics2040042.

[17] N. F. Fahrudin and R. Rindiyani, "Comparison of K-Medoids and K-Means Algorithms in Segmenting Customers based on RFM Criteria," *E3S Web Conf.*, vol. 484, 2024, doi: 10.1051/e3sconf/202448402008.

[18] K. Das, S. Gupta, and A. Kumar, "CAS Condensed and Accelerated Silhouette: An Efficient Method for Determining the Optimal K in K-Means Clustering," pp. 1–13, 2025, [Online]. Available: http://arxiv.org/abs/2507.08311

[19] V. Dawane, P. Waghodekar, and J. Pagare, "RFM Analysis Using K-Means Clustering to Improve Revenue and Customer Retention," *SSRN Electron. J.*, no. Icsmdi, 2021, doi: 10.2139/ssrn.3852887.

[20] Z. Yang, Y. Chen, and J. Corander, "T-SNE Is Not Optimized to Reveal Clusters in Data," pp. 1–20, 2021, [Online]. Available: http://arxiv.org/abs/2110.02573

[21] J. T. Anggelina, "Market Segmentation via K-Means Algorithm and RFM Analysis (Case Study: Microbusiness Sales Transactions)," vol. 1, 2024, [Online]. Available: https://proceeding.unesa.ac.id/index.php/iconbit/article/view/4208/1058

[22] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong, "Towards K-means-friendly spaces: Simultaneous deep learning and clustering," *34th Int. Conf. Mach. Learn. ICML 2017*, vol. 8, pp. 5888–5901, 2017.

[23] G. W. N. W. Zahro, Nafissatus, Nadia Annisa Maori, "Integration of RFM Method and K-Means Clustering for Customer Segmentation Effectiveness," vol. 5, 2025, [Online]. Available: https://doi.org/10.20895/dinda.v5i1.1649

[24] H. J. Wilbert, A. F. Hoppe, A. Sartori, S. F. Stefenon, and L. A. Silva, "Recency, Frequency, Monetary Value, Clustering, and Internal and External Indices for Customer Segmentation from Retail Data," *Algorithms*, vol. 16, no. 9, 2023, doi: 10.3390/a16090396.

[25] R. Gustriansyah, N. Suhandi, and F. Antony, "Clustering optimization in RFM analysis based on k-means," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 18, no. 1, pp. 470–477, 2019, doi: 10.11591/ijeecs.v18.i1.pp470-477.

[26] B. E. Adiana, I. Soesanti, and A. E. Permanasari, "Analisis Segmentasi Pelanggan Menggunakan Kombinasi Rfm Model Dan Teknik Clustering," *J. Terap. Teknol. Inf.*, vol. 2, no. 1, pp. 23–32, 2018, doi: 10.21460/jutei.2018.21.76.