e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 9 No. 2, Agustus, 2025, **Hal. 442-451**
DOI: 10.29408/edumatic.v9i2.30336

K-Means Clustering untuk Segmentasi Pelanggan: Mengungkap Pola Pembelian Strategi Pemasaran pada Sektor Ritel

Andrean Maulana Artiarno [1 *] , Pratomo Setiaji [1] , Fajar Nugraha [1]

[1] Program Studi Sistem Informasi, Universitas Muria Kudus, Indonesia
[*] Correspondence: 202153004@std.umk.ac.id

Received: 16 Mei 2025 | Revised: 2 Juni 2025 | Accepted: 29 Juni 2025 | Published: 12 Agustus 2025

Abstrak
Transformasi digital telah membawa perusahaan ritel menghadapi tantangan baru dalam memahami perilaku konsumen akibat volume data yang meningkat dan preferensi yang terus berubah. Penelitian ini bertujuan untuk mengungkap pola pembelian pelanggan ritel serta memberikan strategi pemasaran berbasis data melalui segmentasi pelanggan menggunakan algoritma *k-means cluster*ing. Jenis penelitian ini adalah kuantitatif eksploratif dengan menggunakan data sintetis sebanyak 3.900 entri dari *platform Kaggle* yang merepresentasikan transaksi ritel. Analisis difokuskan pada variabel usia, jenis kelamin, kategori produk, lokasi, jumlah pembelian, dan frekuensi transaksi. Proses analisis meliputi tahapan *preprocessing*, reduksi dimensi menggunakan PCA, serta segmentasi dengan algoritma *K-Means*. Jumlah klaster optimal ditentukan melalui *elbow method* dan *silhouette score*, sementara evaluasi kualitas klaster dilakukan menggunakan metrik internal, yaitu *calinski-harabasz score* (491,47) dan *davies-bouldin score* (2,02). Nilai tersebut menunjukkan struktur klaster yang baik dan dapat diandalkan. Hasil temuan kami menunjukkan adanya lima segmen pelanggan dengan karakteristik berbeda, mulai dari remaja dengan pembelian kecil dan berkala hingga pelanggan dewasa bernilai tinggi namun jarang bertransaksi. Temuan ini menjadi dasar bagi penyusunan strategi pemasaran seperti program loyalitas, promosi musiman, dan pendekatan eksklusif.

Kata kunci: analisis klaster; data transaksi; k-*means clustering;* segmentasi konsumen*; unsupervised learning*

Abstract
Digital transformation has posed new challenges for retail companies in understanding consumer behavior due to the increasing volume of data and continuously changing preferences. This study aims to uncover purchasing patterns among retail customers and to provide data-driven marketing strategies through customer segmentation using the K-Means Clustering algorithm. This research adopts a quantitative exploratory approach using 3,900 synthetic entries from the Kaggle platform, representing retail transactions. The analysis focuses on variables such as age, gender, product category, location, purchase amount, and transaction frequency. The analytical process includes data preprocessing, dimensionality reduction using PCA, and segmentation with the K-Means algorithm. The optimal number of clusters was determined using the Elbow Method and Silhouette Score, while the quality of the clustering was evaluated using internal metrics, namely the Calinski-Harabasz Score (491.47) and the Davies-Bouldin Score (2.02). These values indicate a well-structured and reliable clustering result. Our findings reveal five distinct customer segments with varying characteristics, ranging from teenagers with small and periodic purchases to high-value adult customers who transact infrequently. These insights serve as the foundation for developing marketing strategies such as loyalty programs, seasonal promotions, and exclusive approaches..

Keywords: *cluster analysis; consumer segmentation; k-means clustering; transaction data; unsupervised learning*

PENDAHULUAN

Di era digital saat ini, perusahaan menghadapi tantangan besar dalam memahami perilaku Konsumen akibat ledakan data transaksional dan meningkatnya ekspektasi pelanggan terhadap layanan yang bersifat personal dan relevan (Lega et al., 2024). Meskipun banyak bisnis telah memiliki akses terhadap big data, seperti data transaksi penjualan dan jejak digital pelanggan (Idham et al., 2024), pemanfaatannya masih terbatas. Sebuah laporan menyebutkan bahwa lebih dari 70% perusahaan mengandalkan data pelanggan dalam pengambilan keputusan (Harahap et al., 2022), tetapi hanya sekitar 30% yang mengimplementasikan segmentasi pelanggan secara efektif (Hermawan et al., 2024). Hal ini menunjukkan adanya kesenjangan signifikan antara potensi data dan penerapannya dalam strategi pemasaran yang tepat sasaran (Rusvinasari, 2025).

Kegagalan dalam mengenali kebutuhan unik setiap segmen pelanggan dapat menyebabkan inefisiensi anggaran pemasaran, rendahnya retensi pelanggan, serta potensi kehilangan pangsa pasar (Ariati et al., 2023). Segmentasi pelanggan menjadi salah satu solusi penting untuk menjawab permasalahan ini, dengan cara mengelompokkan pelanggan ke dalam segmen yang homogen berdasarkan karakteristik atau perilaku tertentu (Awalina & Rahayu, 2023). Sayangnya, adopsi teknologi segmentasi ini masih rendah, terutama pada skala bisnis kecil dan menengah di Indonesia, yang belum banyak menggunakan sistem analitik data untuk mendukung strategi pemasaran berbasis data.

Sebagai pendekatan yang banyak diadopsi, algoritma K-Means Clustering menawarkan solusi komputasional yang efisien dalam segmentasi pelanggan, dengan kapabilitas untuk mengelompokkan data numerik secara otomatis berdasarkan kesamaan karakteristik antar entitas (Perdana et al., 2022). *K-Means* memiliki keunggulan dalam hal efisiensi komputasi, kesederhanaan implementasi, dan kemudahan interpretasi hasil (Harani et al., 2020). Algoritma ini juga bersifat *scalable* untuk dataset berukuran besar, menjadikannya ideal dalam konteks data ritel (Pratama & Maharani, 2025). Namun demikian, *K-Means* juga memiliki keterbatasan, seperti sensitivitas terhadap *outlier* (Rahman et al., 2024), ketergantungan pada inisialisasi *centroid*, dan kurangnya kemampuan dalam menangani data non-numerik tanpa *preprocessing* tambahan (Hidayat et al., 2024). Oleh karena itu, penerapan *K-Means* perlu dilakukan dengan persiapan data yang matang dan evaluasi yang tepat (Prasetyawan et al., 2025).

Beberapa studi sebelumnya telah menggunakan K-Means dalam segmentasi pelanggan, namun masih menyisakan sejumlah celah penelitian. Pertama, minimnya fokus pada konteks industri lokal di Indonesia (Yahya & Kurniawan, 2025). Kedua, kurangnya eksplorasi terhadap implikasi strategis hasil segmentasi dalam mendukung pengambilan keputusan bisnis (Pujiono et al., 2024). Ketiga, terbatasnya penggunaan data publik yang terbuka dan dapat direplikasi oleh peneliti lain. Mayoritas studi terdahulu masih mengandalkan data internal perusahaan (Azhar et al., 2024), sehingga menyulitkan proses validasi atau perbandingan antarstudi (Fajar et al., 2024). Untuk menjawab celah-celah tersebut, penelitian ini menerapkan algoritma *K-Means Clustering* pada dataset publik sintetis dari *Kaggle* yang merepresentasikan transaksi pelanggan di sektor ritel. Pendekatan ini memungkinkan analisis segmentasi berdasarkan atribut usia, jenis kelamin, kategori produk, lokasi, jumlah pembelian, dan frekuensi transaksi.

Penelitian ini bertujuan untuk mengungkap pola pembelian pelanggan ritel melalui segmentasi berbasis algoritma *K-Means Clustering* pada data publik, serta memberikan strategi pemasaran kepada pelanggan loyal, pelanggan potensial, dan pelanggan yang berisiko *churn*. Temuan dari penelitian ini diharapkan memberikan dasar pengambilan keputusan yang lebih akurat dalam perancangan strategi pemasaran, kampanye promosi, dan program retensi pelanggan berbasis data, khususnya di sektor ritel yang semakin kompetitif.

METODE

Penelitian ini menggunakan pendekatan kuantitatif eksploratif, yang bertujuan untuk mengidentifikasi pola dan segmentasi pelanggan berdasarkan data transaksi yang tersedia. Proses penelitian dilakukan melalui beberapa tahapan utama, yaitu: pengumpulan dan eksplorasi data, *preprocessing* untuk mempersiapkan data analisis, penentuan jumlah klaster optimal menggunakan *elbow method* dan *silhouette score*, penerapan algoritma *k-means clustering* untuk segmentasi, serta analisis terhadap hasil klasterisasi guna merumuskan strategi pemasaran yang berbasis data.

Dataset yang digunakan dalam penelitian ini adalah *"Customer Shopping Trends Dataset"* yang diunduh dari *platform kaggle*. *Dataset* ini merupakan data sintetis yang disimulasikan berdasarkan pola transaksi pelanggan pada sektor ritel, dan tidak mencerminkan data transaksi nyata dari suatu perusahaan tertentu. Meskipun demikian, struktur dan variabel dalam *dataset* dirancang menyerupai data aktual, sehingga dapat digunakan untuk studi eksploratif dan pengembangan metode segmentasi pelanggan.

Tahap *preprocessing* meliputi penghapusan duplikat, konversi fitur kategorikal (*gender, category, location*) ke bentuk numerik menggunakan label *encoding*, serta normalisasi seluruh fitur numerik menggunakan *StandardScaler* agar setiap variabel memiliki skala yang sebanding (Setiaji et al., 2024). Karena data sudah bersih dari nilai kosong dan *outlier* ekstrem, proses imputasi tidak diperlukan (Sarimole & Hakim, 2024).

Penentuan jumlah klaster optimal dilakukan menggunakan *elbow method*, yang menunjukkan bahwa penurunan nilai WCSS mulai melandai pada k = 5, serta divalidasi dengan *Silhouette Score*, di mana k = 5 memberikan keseimbangan terbaik antara konsistensi klaster dan kedalaman segmentasi. Klasterisasi dilakukan menggunakan algoritma *K-Means* dari *scikit-learn* di *Python*, dengan parameter *n_clusters=5*, *init='k-means++', random_state=42, n_init='auto', dan max_iter=300*.

Setelah proses segmentasi selesai, dilakukan analisis untuk mengidentifikasi pola perilaku pelanggan yang berbeda, seperti pelanggan dengan frekuensi belanja tinggi namun nominal rendah, serta pelanggan dengan transaksi besar tetapi jarang. Karena *K-Means* merupakan algoritma *unsupervised learning* yang tidak menggunakan *ground truth*, evaluasi berbasis *confusion matrix* tidak dapat diterapkan. Oleh karena itu, evaluasi jumlah klaster optimal ditentukan melalui *elbow method* dan *silhouette score,* sementara evaluasi kualitas klaster dilakukan menggunakan metrik evaluasi internal, yaitu *calinski-harabasz score* dan *davies-bouldin score*. Kedua metrik ini mengukur kepadatan dalam klaster (*intra-cluster cohesion*) dan pemisahan antar klaster (*inter-cluster separation*) untuk menilai seberapa baik struktur segmentasi yang terbentuk. Hasil segmentasi ini menjadi dasar dalam perumusan strategi pemasaran berbasis data, seperti pengembangan program loyalitas atau promosi yang lebih terarah.

HASIL DAN PEMBAHASAN

Hasil

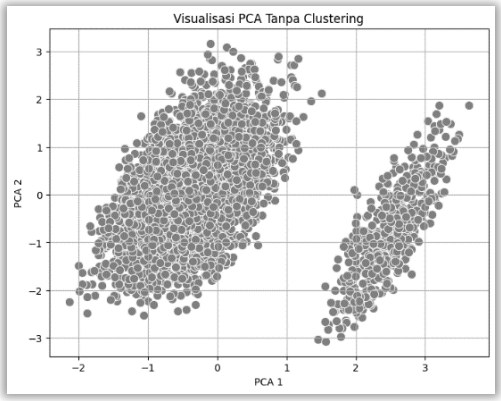
Tahap awal yang dilakukan adalah *preprocessing* yang mencakup penghapusan entri dengan *missing value* dan penambahan kolom *Age Group* yang dikategorikan manual menjadi Remaja, Dewasa, dan Lansia. Selanjutnya, fitur kategorikal seperti *Gender, Category, Location, dan Age Group* diubah menjadi numerik menggunakan Label *Encoding*. Metode ini dipilih karena lebih efisien dibanding *One-Hot Encoding*, terutama pada model berbasis jarak seperti *K-Means*, serta kompatibel dengan algoritma lain seperti *decision tree*.

Tabel 1 menunjukkan hasil dari dataset yang sudah melalui tahap *preprocessing,* Semua fitur numerik kemudian dinormalisasi menggunakan *StandardScaler* untuk menghindari dominasi fitur bernilai besar dan menjaga proporsi antar variabel. Dibanding *MinMaxScaler*, *StandardScaler* lebih sesuai karena menghasilkan distribusi dengan rata-rata nol dan standar

deviasi satu. *Dataset* hasil *preprocessing* ini kemudian digunakan untuk proses klasterisasi dan PCA.

Tabel 1. *Dataset* sudah di *preprocessing*

No	*Age*	*Gender*	*Category*	*Purchase Amount (USD)*	*Location*	*Previous Purchases*	*Frequency of Purchases*	*Age Group*
1	0.71891344	0.685994341	-0.002001925	-0.28562864	-0.576399475	-0.785830671	-0.522605097	-0.640224382
…	…	…	…	..	…	…	…	…
3900	0.521618217	-1.457737974	-1.117360219	0.896686187	-1.413134343	0.529478509	0.476003369	-0.640224382



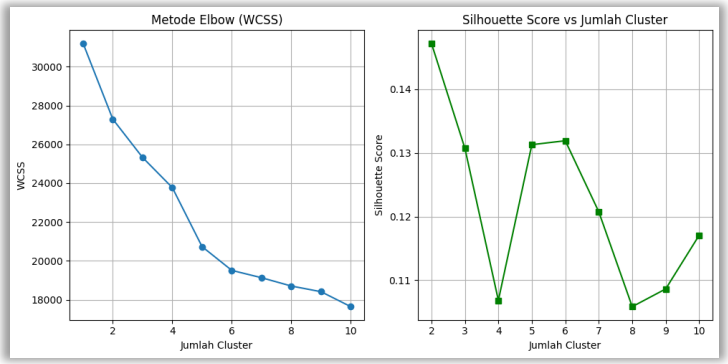
Gambar 1. Visualisasi PCA tanpa *clustering*

Visualisasi PCA pada Gambar 1 memperlihatkan terbentuknya dua hingga tiga gugus besar dalam ruang dua dimensi berdasarkan sumbu PCA1 dan PCA2, yang menjelaskan 27,65% dari total variasi data. Meskipun belum dilakukan klasterisasi, pola ini menunjukkan adanya struktur laten yang berpotensi membentuk kelompok pelanggan. Secara hipotetik, perbedaan antar gugus mencerminkan variasi perilaku pembelian, seperti pelanggan yang sering berbelanja dalam jumlah kecil versus yang jarang berbelanja namun bernilai besar. Penyebaran ini juga kemungkinan dipengaruhi oleh faktor demografis seperti usia dan jenis kelamin.

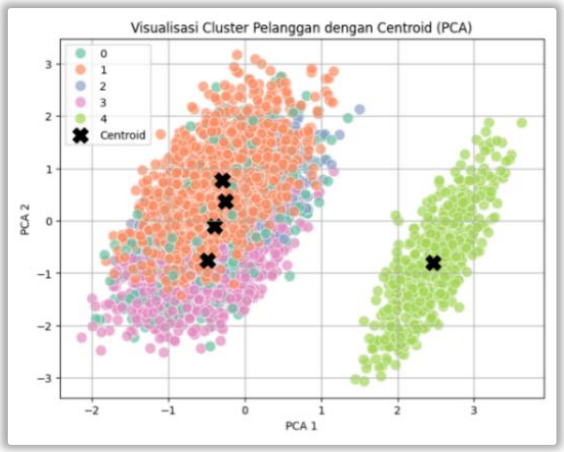
Penentuan jumlah klaster optimal dilakukan menggunakan *Elbow Method* dan *Silhouette Score* sebagai evaluasi pembanding. Nilai WCSS dihitung untuk k = 1 hingga 10 menggunakan algoritma *K-Means* dari pustaka *scikit-learn*, dan divisualisasikan dengan *matplotlib*. *Elbow* digunakan sebagai acuan utama, sementara *Silhouette Score* mendukung validitas pilihan jumlah klaster.

Hasil pada gambar 2 menampilkan grafik *Elbow* dan *Silhouette Score* untuk menentukan jumlah klaster optimal. Titik *elbow* terlihat pada k = 5, saat penurunan WCSS mulai melandai, menandakan bahwa penambahan klaster tidak lagi memberikan perbaikan signifikan. Sebagai pembanding*, Silhouette Score* menunjukkan bahwa meskipun k = 2 memiliki skor tertinggi, segmentasinya terlalu sederhana. Nilai pada k = 5 cukup tinggi dan stabil, sehingga dipilih sebagai klaster optimal karena memberikan keseimbangan antara detail segmentasi dan keterbacaan strategi. Setelah itu, algoritma *K-Means* diterapkan pada data yang telah

dinormalisasi dengan parameter *n_clusters*=5, *random_state*=42, dan n_init='auto', untuk menghasilkan segmentasi yang stabil dan representatif.



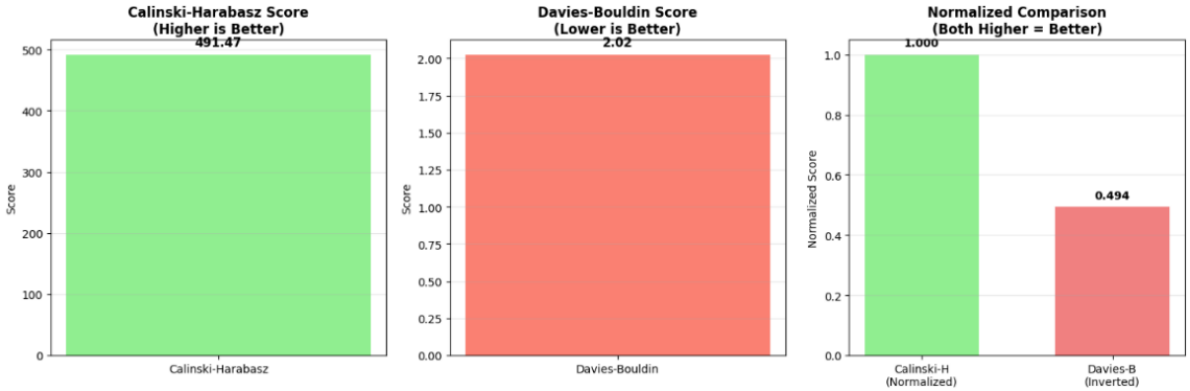
Gambar 2. Hasil *elbow* dan *silhouette score* untuk mencari klaster optimal



Gambar 3. Sebaran hasil klaster

Selanjutnya, pada gambar 3 menunjukkan bahwa setiap pelanggan direpresentasikan sebagai titik berwarna berdasarkan klaster, dan *centroid* masing-masing ditunjukkan dengan tanda silang hitam. Sebagian besar klaster tampak berdekatan, namun Klaster 4 (hijau) terlihat paling terisolasi secara spasial dari klaster lainnya. Pola ini mengindikasikan bahwa pelanggan dalam klaster tersebut memiliki karakteristik yang berbeda secara signifikan. Secara hipotetik, perbedaan ini dapat disebabkan oleh faktor usia yang lebih muda, frekuensi belanja yang lebih jarang, atau dominasi kategori produk tertentu yang tidak muncul di klaster lain. Temuan ini menguatkan bahwa klaster tersebut merepresentasikan segmen pelanggan unik yang memerlukan strategi pemasaran tersendiri.

Gambar 4 menunjukkan hasil evaluasi kualitas klaster dilakukan menggunakan dua metrik internal, yaitu *calinski-harabasz score* dan *davies-bouldin score*. Nilai *calinski-harabasz* sebesar 491,47 menunjukkan struktur klaster yang baik dengan pemisahan yang jelas, sedangkan nilai *Davies-Bouldin* sebesar 2,02 mengindikasikan pemisahan klaster yang cukup. Untuk visualisasi perbandingan, kedua skor dinormalisasi: *calinski-harabasz* dibagi dengan nilainya sendiri (hasilnya 1.000), sementara *davies-bouldin* dibalik terlebih dahulu karena semakin kecil nilainya semakin baik, lalu dinormalisasi menggunakan rumus pembalikan skala. Hasil akhir menunjukkan skor normalisasi *davies-bouldin* sebesar 0,494, yang menandakan struktur klaster cukup layak dan mendukung segmentasi yang dapat diandalkan sebagai dasar strategi pemasaran.



Gambar 4. Evaluasi kualitas klaster

Tabel 2. Karakteristik masing-masing klaster

Cluster	Nama Segmentasi	*Age Group*	*Gender*	*Category*	*Purchase Amount (USD)*	*Location*	*Previous Purchases*	*Frequency of Purchases*
0	Praktis Berkala	Dewasa	*Male*	*Footwear*	57,3094	*North Dakota*	24,9247	*Every 3 Months*
1	Dominan Rutin	Dewasa	*Female*	*Clothing*	60,549	*Idaho*	25,1998	*Monthly*
2	*High-Spender* Stabil	Dewasa	*Male*	*Clothing*	65,1455	*West Virginia*	25,5711	*Quarterly*
3	Pasif Tahunan	Dewasa	*Male*	*Clothing*	54,8441	*Indiana*	26,1949	*Annually*
4	Remaja Periodik	Remaja	*Male*	*Clothing*	60,2016	*Alabama*	24,2243	*Every 3 Months*

Hasil pada tabel 2 menunjukkan bahwa masing-masing klaster memiliki ciri khas yang berbeda. Misalnya, Klaster 4 didominasi oleh pelanggan remaja laki-laki dengan kategori pembelian *clothing* dan pola pembelian setiap 3 bulan, sementara Klaster 2 menunjukkan pelanggan dewasa pria dengan rata-rata pengeluaran pembelian tertinggi. Sementara itu, Klaster 3 memiliki frekuensi pembelian tahunan, mengindikasikan kelompok pelanggan dengan keterlibatan yang lebih rendah. Informasi ini dapat digunakan untuk menyusun strategi pemasaran yang lebih tepat sasaran sesuai dengan segmen pelanggan masing-masing. Perbandingan silang antar klaster menunjukkan bahwa meskipun Klaster 2 memiliki nilai pembelian tertinggi, Klaster 1 lebih unggul dalam jumlah pelanggan. Di sisi lain, Klaster 4 memiliki karakteristik demografis yang sangat berbeda, yaitu kelompok remaja, dan jumlah pelanggan paling sedikit, menjadikannya sebagai segmen khusus.

Setelah mengidentifikasi karakteristik unik dari masing-masing segmen pelanggan hasil klasterisasi, langkah selanjutnya adalah merancang strategi pemasaran yang relevan, terfokus, dan berbasis data. Strategi ini disusun berdasarkan keterkaitan antara perilaku pembelian, faktor demografis, dan preferensi produk dari tiap klaster. Pendekatan ini memungkinkan perusahaan menyesuaikan promosi agar lebih tepat sasaran dan efektif.

Tabel 3. Strategi pemasaran sesuai dengan karakteristik klaster

Klaster	Nama Segmentasi	Karakteristik Utama	Strategi Pemasaran
0	Praktis Berkala	Pria dewasa, belanja sepatu tiap kuartal	Promosi musiman sepatu, *bundling* produk (sepatu + aksesori), kampanye per kuartal
1	Dominan Rutin	Wanita dewasa, belanja pakaian bulanan	*Membership* loyalitas, promosi melalui *influencer*, *reminder* bulanan fashion update
2	*High-Spender* Stabil	Pria dewasa, belanja pakaian kuartalan	Koleksi musiman pria, diskon kuartalan, kampanye gaya hidup maskulin
3	Pasif Tahunan	Pria dewasa, belanja tahunan	Promo tahunan besar, *bundling premium* tahunan, reminder menjelang *event* spesial
4	Remaja Periodik	Remaja pria, belanja pakaian tiap 3 bulan	Konten visual di *TikTok/Instagram*, diskon pelajar, gamifikasi dan voting desain

Pada tabel 3 menyajikan hasil segmentasi pelanggan ke dalam lima klaster yang menunjukkan variasi signifikan dalam karakteristik demografis dan perilaku belanja, sehingga memerlukan strategi pemasaran yang terpersonalisasi. Klaster 0 dan 2 sama-sama terdiri dari pria dewasa dengan pola belanja kuartalan, namun Klaster 2 memiliki daya beli lebih tinggi dan lebih responsif terhadap penawaran premium. Klaster 1 didominasi oleh wanita dewasa dengan kebiasaan belanja bulanan, menjadikannya target ideal untuk program loyalitas dan strategi berbasis tren. Klaster 3 mencerminkan konsumen pasif yang berbelanja tahunan, sehingga efektif dijangkau melalui promosi berskala besar pada momen-momen tertentu. Adapun Klaster 4 terdiri dari remaja pria dengan perilaku belanja periodik dan ketertarikan tinggi terhadap media sosial, yang menuntut pendekatan berbasis konten visual, diskon pelajar, dan kampanye interaktif. Segmentasi ini menegaskan pentingnya pendekatan diferensial berbasis data dalam menyusun strategi pemasaran yang lebih relevan, adaptif, dan bernilai strategis.

Pembahasan

Hasil penelitian ini menunjukkan bahwa algoritma *k-means clustering* berhasil mengelompokkan pelanggan ke dalam lima segmen dengan karakteristik yang berbeda. Hasil ini diperoleh berdasarkan atribut usia, jenis kelamin, kategori produk, nilai belanja, dan frekuensi transaksi. Keberhasilan ini didukung oleh *preprocessing* yang memastikan data berada dalam format numerik dan skala yang seragam, memungkinkan *k-means* bekerja secara optimal dalam menghitung jarak antar data. Visualisasi dua dimensi menggunakan PCA memperlihatkan adanya pemisahan alami antar kelompok, yang menguatkan validitas segmentasi. Pemilihan jumlah klaster sebanyak lima didasarkan pada titik siku (*elbow*) pada grafik WCSS dan diperkuat oleh *silhouette score* yang stabil pada k=5, karena nilai tersebut memberikan keseimbangan optimal antara kompleksitas model dan keterwakilan perilaku pelanggan. Pemilihan ini juga mempertimbangkan bahwa terlalu sedikit klaster (misal k=2 atau k=3) menyederhanakan segmentasi secara berlebihan, sedangkan terlalu banyak klaster (k>5) menghasilkan segmentasi yang terlalu *granular* dan sulit ditindaklanjuti.

Setiap klaster memiliki pola perilaku yang dapat dimanfaatkan secara strategis. Klaster 4 (Remaja Periodik) diisi oleh pelanggan muda pria yang rutin berbelanja pakaian setiap tiga bulan dengan nilai transaksi rendah. Pola ini kemungkinan besar disebabkan oleh keterbatasan

daya beli pada kelompok usia remaja, serta kecenderungan mereka untuk membeli produk saat promosi musiman atau tren tertentu muncul. Klaster 1 (dominan rutin) berisi wanita dewasa dengan frekuensi pembelian bulanan yang tinggi. Pola ini menunjukkan gaya hidup aktif dan konsumtif, yang biasanya dipengaruhi oleh minat terhadap tren fashion dan akses terhadap pendapatan tetap, sehingga membuat mereka responsif terhadap program loyalitas atau promosi berkala. Klaster 3 (pasif tahunan) terdiri dari pria dewasa yang hanya melakukan pembelian sekali dalam setahun. Perilaku ini dapat dikaitkan dengan preferensi belanja konservatif atau kebiasaan hanya membeli saat momen tertentu, seperti akhir tahun atau saat ada diskon besar. Klaster 2 (*high-spender* stabil) menampilkan pelanggan yang jarang berbelanja namun memiliki nilai transaksi tinggi. Hal ini bisa disebabkan oleh karakteristik pelanggan yang cenderung selektif, fokus pada kualitas dibanding kuantitas, serta lebih tertarik pada produk eksklusif atau premium. Terakhir, klaster 0 (praktis berkala) berisi pria dewasa yang secara konsisten membeli sepatu setiap kuartal. Pola ini mengindikasikan adanya kebutuhan fungsional dan kebiasaan rutin, kemungkinan karena aktivitas pekerjaan atau gaya hidup aktif yang menuntut penggantian sepatu secara berkala.

Segmentasi ini mengungkap wawasan yang tidak dapat diidentifikasi melalui analisis agregat biasa. Misalnya, ditemukan pelanggan yang sangat aktif namun memiliki nilai transaksi rendah, sementara disisi lain, ada pelanggan yang jarang belanja tetapi memberikan kontribusi besar secara finansial. Temuan ini menunjukkan pentingnya pendekatan *granular* dibandingkan analisis umum. Pendekatan ini memperkaya segmentasi dibanding model tradisional seperti RFM karena mempertimbangkan variabel demografis dan preferensi produk yang relevan dengan strategi kontemporer.

Hasil temuan oleh penelitian Wilbert et al. (2023) menunjukkan banyak menggunakan data internal dan pendekatan RFM, sedangkan penelitian kami memanfaatkan data publik yang terbuka dan dapat direplikasi, serta fokus pada atribut perilaku dan demografi. Hal ini memperluas cakupan segmentasi pelanggan dalam konteks industri ritel digital yang dinamis dan berbasis data terbuka. Selain itu, berbeda dengan penelitian yang dilakukan oleh Robbani et al. (2024) yang hanya mengevaluasi klaster menggunakan *silhouette score*, penelitian kami menerapkan dua metrik internal tambahan, yaitu *calinski-harabasz score* dan *davies-bouldin score*, untuk memperoleh validasi yang lebih objektif dan menyeluruh terhadap kualitas klaster. Pendekatan ini menghasilkan segmentasi yang lebih terukur dan dapat diandalkan dalam mendukung strategi pemasaran berbasis data.

SIMPULAN

Algoritma *k-means clustering* efektif dalam membentuk lima segmen pelanggan yang berbeda secara signifikan berdasarkan perilaku pembelian, frekuensi transaksi, dan karakteristik demografis. Hasil ini menunjukkan bahwa pelanggan tidak bersifat homogen, sehingga diperlukan pendekatan pemasaran yang lebih terarah. Visualisasi klaster melalui PCA memperkuat pemisahan antar segmen dan meningkatkan validitas segmentasi yang dilakukan. Segmentasi ini memberikan wawasan strategis yang dapat dimanfaatkan dalam pengambilan keputusan pemasaran, seperti promosi bulanan untuk pelanggan aktif, program loyalitas untuk pelanggan bernilai tinggi, atau kampanye digital untuk segmen muda. Sebagai kontribusi, penelitian ini menggunakan data publik untuk menghasilkan segmentasi yang dapat direplikasi dan diterapkan di berbagai konteks bisnis. Untuk penelitian selanjutnya, disarankan agar pendekatan serupa diterapkan pada data transaksi aktual dari industri tertentu guna meningkatkan relevansi praktis. Penelitian juga dapat dikembangkan dengan membandingkan algoritma lain seperti *DBSCAN* atau *hierarchical clustering*, atau dengan menambahkan dimensi perilaku lain seperti waktu pembelian atau jenis produk lebih rinci guna meningkatkan akurasi segmentasi pelanggan.

REFERENSI

Ariati, I., Norsa, R. N., Akhsan, L., & Heikal, J. (2023). Segmentasi Pelanggan Menggunakan K-Means Clustering Studi Kasus Pelanggan Uht Milk Greenfield. *Cerdika: Jurnal Ilmiah Indonesia*, *3*(7), 729–743. https://doi.org/10.59141/cerdika.v3i7.639

Awalina, E. F. L., & Rahayu, W. I. (2023). Optimalisasi strategi pemasaran dengan segmentasi pelanggan menggunakan penerapan K-means clustering pada transaksi online retail. *Jurnal Teknologi Dan Informasi*, *13*(2), 122–137. https://doi.org/10.34010/jati.v13i2.10090

Azhar, Z., Wulandari, C., Hanum, Z., Putra, W. A., & Saragih, Y. P. (2024). Implementasi Pengelompokan Persediaan Sepeda Motor Menggunakan Metode Clustering K-Means. *Explorer*, *4*(2), 69–76. https://doi.org/10.47065/explorer.v4i2.1255

Fajar, M., Rahaningsih, N., & Dana, R. D. (2024). Analisis Pola Penjualan Obat Di Apotek an-Naafi Menggunakan Metode K-Means Clustering. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(1), 486–492. https://doi.org/10.36040/jati.v8i1.8395

Harahap, M., Lubis, Y., & Situmorang, Z. (2022). Analisis Pemasaran Bisnis dengan Data Science: Segmentasi Kepribadian Pelanggan berdasarkan Algoritma K-Means Clustering. *Data Sciences Indonesia (DSI)*, *1*(2), 76–88. https://doi.org/10.47709/dsi.v1i2.1348

Harani, N. H., Prianto, C., & Nugraha, F. A. (2020). Segmentasi pelanggan produk digital service Indihome menggunakan algoritma K-Means berbasis Python. *Jurnal Manajemen Informatika (JAMIKA)*, *10*(2), 133–146. https://doi.org/10.34010/jamika.v10i2.2683

Hermawan, A., Jayanti, N. R., Saputra, A., Tambunan, C., Baihaqi, D. M., Syahreza, M. A., & Bachtiar, Z. (2024). Optimalisasi Strategi Pemasaran Melalui Analisis RFM pada Dataset Transaksi Ritel Menggunakan Python. *Jurnal Manajemen Riset Inovasi*, *2*(4), 254–267. https://doi.org/10.55606/mri.v2i4.3246

Hidayat, R. S., Muttaqin, M. R., & Irmayanti, D. (2024). Pengelompokan Daerah Rawan Bencana Di Jawa Tengah Menggunakan Algoritma K-Means Clustering. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(5), 10035–10042. https://doi.org/10.36040/jati.v8i5.10880

Idham, I., Rosika, H., & Yuliadi, Y. (2024). Implementasi Rapidminer Untuk Clestering Data Penjualan Pakaian Menggunakan Metode K-Means. *JUTECH: Journal Education and Technology*, *5*(1), 221–231. https://doi.org/10.31932/jutech.v5i1.3642

Lega, A., Adytia, P., & Lailiyah, S. (2024). Penerapan Algoritma K-means Clustering untuk Klasterisasi Penjualan Smartphone pada Carin Cell. *STMIK Widya Cipta Dharma*.

Perdana, S. A., Florentin, S. F., & Santoso, A. (2022). Analisis Segmentasi Pelanggan Menggunakan K-Means Clustering Studi Kasus Aplikasi Alfagift. *Sebatik*, *26*(2), 446–457. https://doi.org/10.46984/sebatik.v26i2.1991

Prasetyawan, D., Mulyanto, A., & Gatra, R. (2025). Pemetaan Lintasan Karir Alumni Berdasarkan Analisis Cluster: Kombinasi K-Means dan Reduksi Dimensi Autoencoder. *Edumatic: Jurnal Pendidikan Informatika*, *9*(1), 198–207. https://doi.org/10.29408/edumatic.v9i1.29713

Pratama, R. F. P., & Maharani, W. (2025). Comparative Analysis of Naive Bayes and SVM for Improved Emotion Classification on Social Media. *Edumatic: Jurnal Pendidikan Informatika*, *9*(1), 11–20. https://doi.org/10.29408/edumatic.v9i1.29087

Pujiono, S., Astuti, R., & Basysyar, F. M. (2024). Implementasi Data Mining Untuk Menentukan Pola Penjualan Produk Menggunakan Algoritma K-Means Clustering. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(1), 615–620. https://doi.org/10.36040/jati.v8i1.8360

Rahman, F. D., Mulki, M. I. Z., & Taryana, A. (2024). Clustering dan klasifikasi data cuaca Cilacap dengan menggunakan metode K-Means dan Random Forest. *Jurnal SINTA:*

Sistem Informasi Dan Teknologi Komputasi, *1*(2), 90–97. https://doi.org/10.61124/sinta.v1i2.15

Robbani, M. A., Firmansyah, G., Widodo, A. M., & Tjahjono, B. (2024). Clustering of Child Stunting Data in Tangerang Regency Using Comparison of K-Means, Hierarchical Clustering and DBSCAN Methods. *Asian Journal of Social and Humanities*, *2*(12), 3105–3115. https://doi.org/10.59888/ajosh.v2i12.422

Rusvinasari, D. (2025). Analisis Klasterisasi Pola Penjualan Menu Makanan pada Rumah Makan menggunakan Metode K-Means Clustering. *Jurnal Informatika: Jurnal Pengembangan IT*, *10*(2), 398–409. https://doi.org/10.30591/jpit.v10i2.8511

Sarimole, F. M., & Hakim, L. (2024). Klasifikasi barang menggunakan metode clustering K-Means dalam penentuan prediksi stok barang. *Jurnal Sains Dan Teknologi*, *5*(3), 846–854. https://doi.org/10.55338/saintek.v5i3.2709

Setiaji, P., Adi, K., & Surarso, B. (2024). Development of Classification Method for Determining Chicken Egg Quality Using GLCM-CNN Method. *Ingenierie Des Systemes d'Information*, *29*(2), 397–407. https://doi.org/10.18280/isi.290201

Wilbert, H. J., Hoppe, A. F., Sartori, A., Stefenon, S. F., & Silva, L. A. (2023). Recency, Frequency, Monetary Value, Clustering, and Internal and External Indices for Customer Segmentation from Retail Data. *Algorithms*, *16*(9), 396. https://doi.org/10.3390/a16090396

Yahya, A., & Kurniawan, R. (2025). Implementasi Algoritma K-Means untuk Pengelompokan Data Penjualan Berdasarkan Pola Penjualan: Implementation of K-Means Algorithm for Clustering Sales Data Based on Sales Patterns. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(1), 350–358. https://doi.org/10.57152/malcom.v5i1.1773