Segmentasi Pelanggan Berdasarkan Volume Pembelian Produk Menggunakan Algoritma K-Means Untuk Efektivitas Pemasaran

Dennis Ma'rifatul Kurnia[1*], Nurun Nihayatur R.A.[2],
Yulistiya Nur Hidayah[3], Cindy Avitaselly B.S.[4]
[1 - 4] Prodi Sistem Informasi, Fakultas Teknik dan Ilmu Komputer,
Universitas Nusantara PGRI Kediri
[*]email : denniskurnia33@gmail.com

ABSTRACT

The rapid development of information technology has changed the way products are sold, especially through online platforms that are increasingly in demand. In increasingly tight business competition, companies need to understand the differences in customer needs and behavior. Inability in this regard can make it difficult to design effective marketing strategies. Therefore, customer segmentation based on transaction data is an important solution to group customers based on similar purchasing patterns. This study aims to examine customer segmentation based on sales transactions to help companies understand customer characteristics and develop more targeted and adaptive marketing strategies. A quantitative approach is used by applying the K-Means Clustering algorithm and PCA dimension reduction to a dataset from Kaggle containing 3,900 entries with 9 attributes. Determination of the optimal number of clusters was carried out using the Elbow and Silhouette Score methods. The segmentation results show five optimal clusters with the highest Silhouette Score of 0.81. Cluster 0 is the most dominant. PCA visualization shows a fairly clear cluster separation although there is little overlap. This study has succeeded in grouping customers based on purchase volume. Limitations of the study include the uneven distribution of clusters and it is recommended to add demographic attributes and evaluate other algorithms such as DBSCAN.

Keywords: *customer segmentation, k-means, pca*

ABSTRAK

Pesatnya perkembangan teknologi informasi telah mengubah cara penjualan produk, terutama melalui platform daring yang semakin diminati. Dalam persaingan bisnis yang semakin ketat, perusahaan perlu memahami perbedaan kebutuhan dan perilaku pelanggan. Ketidakmampuan dalam hal ini dapat menyulitkan perancangan strategi pemasaran yang efektif. Oleh karena itu, segmentasi pelanggan berbasis data transaksi menjadi solusi penting untuk mengelompokkan pelanggan berdasarkan pola pembelian yang serupa. Penelitian ini bertujuan mengkaji segmentasi pelanggan berdasarkan transaksi penjualan guna membantu perusahaan memahami karakteristik pelanggan dan menyusun strategi pemasaran yang lebih terarah dan adaptif. Pendekatan kuantitatif digunakan dengan menerapkan algoritma *K-Means Clustering* dan reduksi dimensi PCA pada dataset dari *Kaggle* yang berisi 3.900 entri dengan 9 atribut. Penentuan jumlah klaster optimal dilakukan menggunakan metode Elbow dan *Silhouette Score*. Hasil segmentasi menunjukkan lima klaster optimal dengan Silhouette Score tertinggi sebesar 0,81. Klaster 0 menjadi yang paling dominan. Visualisasi PCA memperlihatkan pemisahan klaster yang cukup jelas meski terdapat sedikit tumpang tindih. Penelitian ini berhasil mengelompokkan pelanggan berdasarkan volume pembelian. Keterbatasan penelitian mencakup distribusi klaster yang kurang merata dan disarankan menambahkan atribut demografi serta mengevaluasi algoritma lain seperti DBSCAN.

Kata Kunci : segmentasi pelanggan, k-means, pca

PENDAHULUAN

Perkembangan teknologi digital telah merevolusi sistem penjualan barang, mendorong pelaku usaha untuk beradaptasi guna menjaga keberlangsungan bisnis. Penjualan secara daring kini menjadi pilihan utama karena memberikan kemudahan bagi konsumen tanpa perlu hadir langsung ke toko. Dalam persaingan yang semakin ketat, perusahaan dituntut untuk memahami kebutuhan dan perilaku pelanggan agar dapat merancang strategi pemasaran yang relevan dan tepat sasaran[1] [2]. Namun, perbedaan karakteristik pelanggan antar wilayah menimbulkan tantangan tersendiri. Strategi pemasaran tidak bisa bersifat seragam, melainkan harus disesuaikan dengan pola permintaan yang bervariasi. Tanpa pemahaman yang mendalam terhadap pola pembelian pelanggan, penyusunan strategi bisnis akan menjadi tidak efektif[2].

Oleh karena itu, pendekatan berbasis data, seperti *clustering*, menjadi penting. *Clustering* memungkinkan perusahaan mengelompokkan pelanggan berdasarkan kemiripan perilaku, seperti kebiasaan belanja dan preferensi produk, sehingga strategi yang dirancang lebih sesuai untuk tiap segmen pasar[3] [4]. Metode *clustering* seperti *K-Means* sangat efektif untuk segmentasi pelanggan berdasarkan data transaksi. Penerapan teknik ini tidak hanya memudahkan dalam menyusun strategi promosi dan retensi, tetapi juga membantu dalam pengelolaan stok dan alokasi sumber daya[5] [6].

Dengan segmentasi yang tepat, pelanggan bernilai tinggi dapat diberi program loyalitas, sedangkan pelanggan bernilai rendah dapat diarahkan pada strategi reaktivasi [7]. Penelitian ini menggunakan algoritma K-Means sebagai metode utama segmentasi, dengan dukungan *Principal Component Analysis* (PCA) untuk reduksi dimensi dan visualisasi. Kelebihan K-Means terletak pada kemampuannya mengelompokkan data berdasarkan jarak dan kemiripan, sehingga cocok digunakan dalam pengolahan data transaksi pelanggan[8] [9]. Dalam konteks persaingan digital, strategi berbasis data seperti ini sangat relevan untuk menyusun kampanye pemasaran yang personal dan efisien[10]. Meskipun dataset yang digunakan dalam penelitian ini berasal dari sumber terbuka (Kaggle), penelitian ini tetap memiliki kebaruan dan diferensiasi yang signifikan dibandingkan studi-studi sebelumnya. Perbedaan utama terletak pada pendekatan evaluasi segmentasi yang lebih komprehensif**,** yaitu dengan menggabungkan tiga metode evaluasi klaster *Elbow Method, Silhouette Score,* dan *Davies-Bouldin Index* (DBI). Ketiga indikator tersebut memberikan gambaran kuantitatif mengenai kualitas klaster yang terbentuk. DBI secara khusus digunakan untuk mengukur pemisahan antar klaster dan kekompakan dalam klaster, memberikan validasi tambahan atas nilai K yang dipilih. Pendekatan ini memperkuat hasil segmentasi, dan mendukung temuan dari studi sebelumnya seperti A. Alamsyah [11] dan S. Sharyanto [12] Visualisasi hasil segmentasi dilakukan menggunakan PCA, yang memperjelas distribusi dan pemisahan antar klaster secara visual. Integrasi metode evaluasi dan visualisasi ini menjadi kekuatan utama dalam penelitian, karena menghasilkan pemetaan karakteristik pelanggan yang dapat langsung diterjemahkan ke dalam strategi pemasaran digital yang terarah dan responsif terhadap perubahan pasar.

METODE

Mengidentifikasi pola pembelian pelanggan, algoritma *K-Means* digunakan sebagai metode segmentasi. Hasil segmentasi ini bertujuan untuk membantu perusahaan dalam merumuskan strategi pemasaran yang lebih tepat sasaran berdasarkan pola pembelian pelanggan.

2.1 Dataset

Penelitian ini mengadopsi pendekatan kuantitatif dengan menggunakan data sekunder yang diperoleh dari platform *Kaggle.com*. Dataset tersebut berisi 3900 entri penjualan online dengan 9 atribut utama yang digunakan, Data ini dihimpun oleh pihak ketiga yang kredibel, meskipun identitas lengkap pengumpul data tidak diungkapkan untuk menjaga kerahasiaan sumber

Tabel 1. Klasterisasi Dataset

No	Table Dataset		
	Nama Kolom	**Jumlah**	**Keterangan**
1	Customer ID	3900	ID unik pelanggan
2	Age	53	Usia pelanggan (18-70 tahun)
3	Gender	2	Jenis kelamin (male, female)
4	Item Purchased	25	Nama barang yang dibeli
5	Category	4	Kategori barang (clothing, footwear,outerwear, accessories)
6	Purchase Amount (USD)	81	Jumlah pembelian dalam USD
7	Previous Purchases	50	Jumlah pembelian sebelumnya
8	Payment Method	6	Metode Pembayaran
9	Frequency of Purchases	7	Frekuensi Pembelian

2.2 CRISP-DM

Penelitian ini menggunakan pendekatan model proses CRISP-DM (Cross Industry Standard Process for Data Mining) sebagai kerangka metodologi utama [13] dalam pelaksanaan segmentasi pelanggan diperlukan beberapa tahapan sesuai dengan kasus yaitu :

2.2.1 Business Understanding

Tahap ini bertujuan untuk memahami situasi bisnis dan menetapkan tujuan data mining, seperti segmentasi pelanggan menggunakan teknik clustering. Indikator keberhasilan ditentukan, misalnya tingkat akurasi atau presisi model.

2.2.2 Data understanding

Fokus pada pemahaman struktur dan karakteristik data. Data ditinjau untuk mengidentifikasi atribut yang relevan, seperti usia, jumlah transaksi, dan riwayat pembelian. Juga dilakukan analisis distribusi data, korelasi antar variabel, serta deteksi *outlier* sebagai dasar untuk tahap selanjutnya.

2.2.3 Data preparation

Data diproses melalui seleksi, pembersihan duplikat, dan penanganan nilai kosong dengan teknik imputasi. Atribut turunan dapat dibuat untuk memperkaya informasi. Skala fitur numerik distandardisasi menggunakan *Z-score* karena *K-Means* sensitif terhadap perbedaan skala antar variabel.

2.2.4 Principal Component Analysis (PCA)

PCA atau *Principal Component Analysis* merupakan metode yang banyak dimanfaatkan dalam berbagai keperluan, seperti pengurangan dimensi data multivariat, kompresi informasi, pengenalan pola dalam sistem jaringan saraf, serta analisis statistik. Dalam praktiknya, PCA kerap digunakan sebagai tahap awal

pemrosesan data pada jaringan saraf tiruan, terutama dalam aplikasi klasifikasi maupun peramalan [14].

$$X_{\text{PCA}} = X_{\text{std}} \cdot W$$

- XPCA: Data pelanggan pada ruang dimensi baru sebagai klasterisasi.
- X_std: Data asli pelanggan yang sudah distandarisasikan.
- W: Matriks untuk menentukan arah dimensi baru berdasarkan variabilitas data.

Setelah pembersihan dan standarisasi, data direduksi dimensinya dengan PCA menjadi dua komponen utama (PC1 dan PC2). Reduksi ini mempercepat komputasi dan mempermudah visualisasi dua dimensi. Scatter plot hasil PCA menunjukkan pola penyebaran yang membentuk kelompok terpisah, mengindikasikan adanya klaster alami yang menjadi dasar segmentasi dengan *K-Means*.

2.2.4 Modeling K-Means Clustering

Rumus Penentuan jumlah *cluster* optimal (K) untuk modeling dalam algoritma *K-Means* sering dilakukan dengan metode *Elbow*, yang menggunakan rumus untuk menghitung *Within-Cluster Sum of Squares* (WCSS) atau *Sum of Squared Error* (SSE) seperti berikut:

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

- k: jumlah cluster.
- C_i: cluster ke-*i*.x: data poin dalam cluster C_i.
- μ_i: centroid dari cluster C_i.
- $\|x - \mu_i\|^2$: jarak kuadrat antara data poin x dan centroid μ_i.

2.2.5 Evaluation

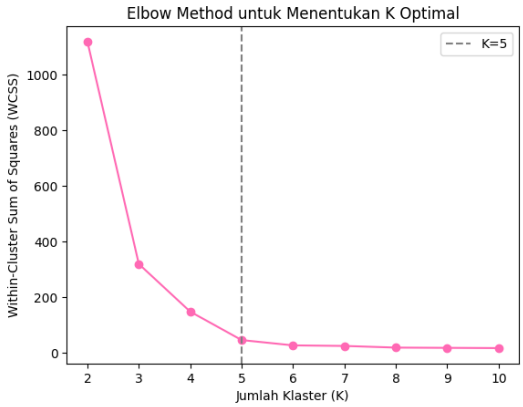
Tahap ini berfokus pada evaluasi hasil pemodelan dengan mengkaitkannya terhadap sasaran bisnis yang ingin dicapai. Pemahaman terhadap model yang dihasilkan sangat penting guna menyusun langkah strategis berikutnya, khususnya dalam hal segmentasi pelanggan. Proses evaluasi bertujuan untuk menilai efektivitas algoritma *clustering* yang digunakan yakni *K-Means* dalam merepresentasikan pola alami dalam data. Validasi hasil klaster tidak hanya dilihat dari kecocokannya dengan pemahaman domain, seperti frekuensi transaksi dan volume pembelian, tetapi juga diperkuat melalui visualisasi klaster menggunakan *scatter plot* hasil reduksi dimensi PCA. Selain itu, pada tahap ini dilakukan pengukuran performa model menggunakan metrik *Davies-Bouldin Index* (DBI) [15], sebagai salah satu indikator kuantitatif. DBI digunakan untuk menilai seberapa baik pemisahan antar klaster serta konsistensi dalam masing-masing kelompok. Nilai DBI yang rendah menunjukkan pemisahan yang jelas antar klaster dan kekompakan internal yang tinggi, sehingga dapat dijadikan acuan dalam menentukan jumlah klaster (K) yang optimal untuk segmentasi yang lebih akurat.

HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk melakukan segmentasi pelanggan berdasarkan volume pembelian produk dengan memanfaatkan algoritma *K-Means Clustering* sebagai metode utama dalam pengelompokan data. Dalam implementasinya, proses klasterisasi dilakukan melalui beberapa tahap penting, dimulai dari pra-pemrosesan data, penentuan jumlah klaster optimal, evaluasi kualitas pemodelan, hingga interpretasi dan visualisasi hasil klasterisasi. Seluruh tahapan tersebut dilakukan untuk memastikan bahwa hasil segmentasi yang diperoleh bersifat representatif dan dapat dijadikan dasar yang kuat dalam pengambilan keputusan strategis di bidang pemasaran.

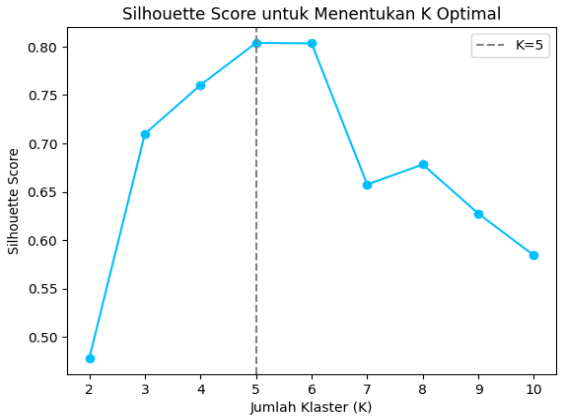
4.1. Penentuan Jumlah Klaster Optimal

Penentuan jumlah klaster yang optimal merupakan aspek krusial dalam metode *K-Means*, karena nilai K yang tidak sesuai akan menghasilkan segmentasi yang tidak mencerminkan struktur alami dari data. Oleh karena itu, dua pendekatan analisis yang lazim digunakan, yaitu metode *Elbow* dan metode *Silhouette Score*, diterapkan untuk mengevaluasi jumlah klaster yang paling tepat.

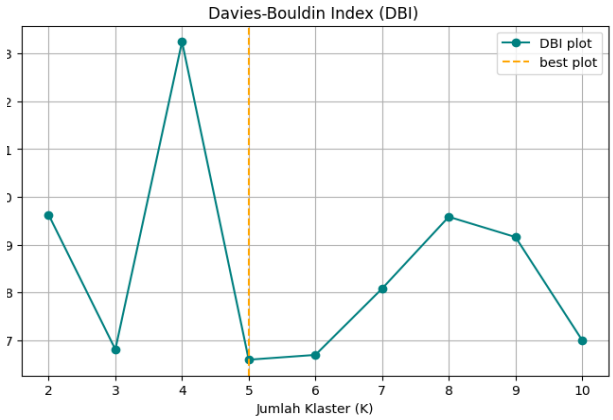


Gambar 1 : *Elbow Method* K Optimal

Pada analisis menggunakan *Elbow Method*, nilai *Within-Cluster Sum of Squares* (WCSS) dianalisis untuk sejumlah nilai K mulai dari 2 hingga 10. Dari grafik yang dihasilkan, terlihat bahwa penurunan WCSS signifikan terjadi hingga nilai K = 5, setelah itu penurunan menjadi relatif kecil. Titik ini dikenal sebagai elbow point, dan menunjukkan bahwa pada K = 5, struktur klaster mulai stabil dan penambahan klaster berikutnya tidak memberikan peningkatan signifikan dalam pemodelan.



Gambar 2 : *Silhoutte Score*



Gambar 3 : *Davies-Bouldin Index (DBI)*

Untuk memperkuat hasil tersebut, dilakukan pula analisis menggunakan *Silhouette Score*, yang mengukur sejauh mana suatu objek sesuai dengan klaster tempatnya berada. Skor *silhouette* berada dalam rentang -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan bahwa objek sangat cocok dengan klaster-nya dan kurang sesuai jika dipindahkan ke klaster lain. Hasil analisis menunjukkan bahwa pada K = 5, diperoleh nilai *Silhouette Score* tertinggi sebesar 0,81, yang mengindikasikan kualitas segmentasi yang sangat baik.

Selain itu, evaluasi menggunakan *Davies-Bouldin Index (DBI)* juga dilakukan untuk menilai validitas pembentukan klaster berdasarkan rasio antara jarak antar klaster dan sebaran dalam klaster. Semakin kecil nilai DBI, semakin baik pemisahan antar klaster. Pada K = 5, diperoleh nilai DBI sebesar 2,66, yang menunjukkan pemisahan klaster yang cukup baik meskipun bukan yang paling optimal secara matematis.

Dengan mempertimbangkan kedua metrik tersebut, yaitu nilai *Silhouette Score* tertinggi dan nilai DBI yang masih berada dalam batas wajar pada K = 5, maka dapat disimpulkan bahwa jumlah klaster optimal dalam penelitian ini adalah sebanyak lima klaster (K = 5).

4.2. Statistik Deskriptif Pelanggan

Tahap ini bertujuan untuk memahami struktur dan karakteristik awal data yang digunakan dalam penelitian. Proses dilakukan melalui pengumpulan, eksplorasi, dan evaluasi data untuk mengidentifikasi atribut-atribut penting yang akan digunakan dalam segmentasi pelanggan. Fokus analisis ditujukan pada tiga variabel numerik utama, yaitu usia pelanggan (Age), jumlah transaksi pembelian (Purchase Amount), dan riwayat pembelian sebelumnya (Previous Purchases).

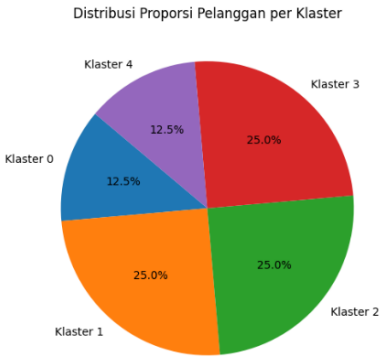
- Distribusi usia pelanggan berkisar antara 20 hingga 70 tahun, dengan rata-rata 48,6 tahun dan deviasi standar 14,2, mengindikasikan keberagaman umur yang cukup tinggi.
- Rata-rata nilai pembelian adalah $59,76 dengan deviasi standar $23,43, menandakan variasi transaksi antar pelanggan cukup besar.
- Riwayat pembelian menunjukkan rata-rata sebesar 25,4 dengan deviasi standar 14,5, mencerminkan bahwa sebagian besar pelanggan memiliki frekuensi pembelian yang tinggi meskipun tersebar luas.

Distribusi nilai pada *Purchase Amount* dan *Previous Purchases* cenderung miring ke kanan, yang menunjukkan adanya beberapa pelanggan dengan nilai transaksi yang jauh lebih tinggi dari mayoritas. Analisis korelasi menunjukkan bahwa ketiga variabel memiliki hubungan yang sangat lemah (nilai korelasi < 0,1), sehingga masing-masing dapat digunakan secara mandiri dalam proses segmentasi. Selain itu, pemeriksaan outlier menggunakan metode IQR tidak menemukan nilai ekstrem yang signifikan, sehingga keseluruhan data dinyatakan layak untuk digunakan dalam tahap pemodelan selanjutnya.

4.3. Distribusi Pelanggan Berdasarkan Klaster

Distribusi pelanggan dalam masing-masing klaster juga menjadi indikator penting untuk memahami proporsi dan komposisi dari tiap segmen yang terbentuk. Visualisasi

berbentuk diagram lingkaran (pie chart) menunjukkan bahwa distribusi antar klaster cukup seimbang, dengan klaster-klaster utama (Klaster 0, 1, dan 2) mencakup proporsi terbesar, masing-masing sekitar 25% dari keseluruhan populasi pelanggan.

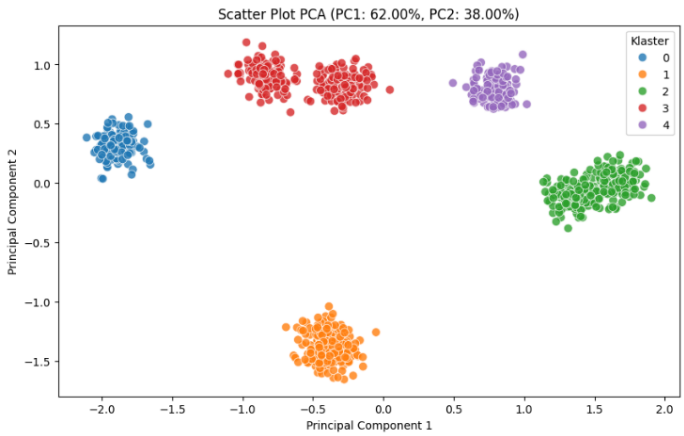


Gambar 4. Distribusi Pelanggan Per Cluster

Sementara itu, Klaster 3 dan 4 memiliki jumlah pelanggan yang lebih sedikit namun tetap signifikan untuk dipertimbangkan dalam pengambilan keputusan strategis. Distribusi yang seimbang ini mengindikasikan bahwa proses klasterisasi tidak menghasilkan dominasi dari salah satu klaster tertentu, sehingga memungkinkan penerapan strategi pemasaran yang proporsional dan merata pada seluruh segmen pelanggan.

4.4. Visualisasi Hasil Klasterisasi Menggunakan PCA

Agar hasil klasterisasi dapat divisualisasikan secara dua dimensi dan mudah diinterpretasikan, digunakan metode Principal Component Analysis (PCA) untuk melakukan reduksi dimensi. PCA digunakan untuk mereduksi sejumlah fitur dalam data asli menjadi dua komponen utama yang paling banyak menjelaskan variansi data.



Gambar 5. Scatter Plot PCA

Hasil visualisasi scatter plot menunjukkan bahwa dua komponen utama PCA (PC1 dan PC2) mampu menjelaskan 100% variasi total dalam data, dengan PC1 berkontribusi sebesar 62,00% dan PC2 sebesar 38,00%. Dalam visualisasi tersebut, empat klaster yang berbeda ditandai dengan warna biru, oranye, hijau, dan ungu. Warna pada scatter plot mewakili klaster yang berbeda, dengan klaster 0 (biru) dan klaster 1 (oranye) terletak

di sisi kiri plot, sedangkan klaster 2 (hijau) dan klaster 3 (ungu) berada di sisi kanan. Posisi ini menunjukkan adanya perbedaan karakteristik yang jelas antara klaster, dengan PC1 yang mendominasi variasi horizontal dan PC2 yang mendukung variasi vertikal. Distribusi titik dalam setiap klaster juga menunjukkan konsistensi internal, yang mengindikasikan bahwa data dalam setiap klaster memiliki kesamaan yang tinggi.

4.5. Interpretasi Karakteristik Klaster

Interpretasi karakteristik klaster dilakukan berdasarkan hasil segmentasi yang telah diperoleh melalui penerapan algoritma K-Means dan didukung visualisasi menggunakan Principal Component Analysis (PCA). Tabel berikut menyajikan ringkasan statistik deskriptif dari masing-masing klaster berdasarkan empat atribut utama: usia rata-rata, jumlah pembelian, riwayat pembelian sebelumnya, dan frekuensi pembelian.

Klaster	Rata-rata Usia	Rata-rata Jumlah Pembelian (USD)	Rata-rata Pembelian Sebelumnya	Rata-rata Frekuensi
0	50,3 tahun	75,2	38,7	6,1
1	47,8 tahun	58,4	24,1	4,5
2	46,9 tahun	60,5	26,7	5,2
3	43,2 tahun	48,9	19,3	3,8
4	51,1 tahun	39,2	11,4	2,6

Berdasarkan data pada tabel tersebut, masing-masing klaster menunjukkan pola perilaku yang berbeda, yang dapat dijadikan dasar penyusunan strategi pemasaran yang lebih spesifik dan tepat sasaran. Adapun interpretasi dari masing-masing klaster adalah sebagai berikut:

- **Klaster 0** merupakan kelompok pelanggan dengan volume pembelian yang sangat tinggi, frekuensi transaksi yang intens, serta riwayat pembelian yang panjang. Rata-rata usia yang relatif matang menunjukkan kemungkinan kelompok ini memiliki daya beli yang tinggi. Oleh karena itu, klaster ini dapat dikategorikan sebagai segmen pelanggan premium atau loyal, yang memiliki kontribusi signifikan terhadap pendapatan perusahaan. Strategi pemasaran yang relevan untuk kelompok ini meliputi program loyalitas, penawaran eksklusif, dan prioritas layanan untuk mempertahankan dan meningkatkan retensi pelanggan.
- **Klaster 1 dan 2** menunjukkan karakteristik pembelian dalam kategori menengah, baik dari segi nominal transaksi maupun intensitas pembelian. Kelompok ini merepresentasikan pelanggan yang relatif stabil dan menunjukkan potensi untuk dikembangkan lebih lanjut. Intervensi strategis seperti promosi bundling, penawaran diskon khusus, serta program upselling dapat diterapkan untuk meningkatkan nilai transaksi dan keterlibatan mereka secara bertahap.
- **Klaster 3** terdiri dari pelanggan dengan frekuensi dan volume pembelian yang rendah namun konsisten. Meskipun kontribusinya tidak sebesar klaster-klaster sebelumnya, pola pembelian yang tetap menunjukkan adanya potensi loyalitas yang dapat dibangun. Strategi yang dapat diterapkan mencakup edukasi produk,

pengiriman informasi berkala, serta kampanye pemasaran yang lebih personal untuk mendorong keterlibatan jangka panjang.

- **Klaster 4** dicirikan oleh volume dan frekuensi pembelian yang paling rendah di antara seluruh klaster, serta riwayat transaksi yang minim. Kelompok ini dapat dikategorikan sebagai pelanggan pasif atau sporadis. Dalam konteks pemasaran, klaster ini merupakan target utama untuk upaya reaktivasi, seperti pemberian diskon besar, kampanye retargeting, atau komunikasi promosi yang bersifat personal guna memicu ketertarikan kembali terhadap produk yang ditawarkan.

Secara keseluruhan, hasil interpretasi ini menegaskan bahwa pendekatan segmentasi berbasis K-Means mampu menghasilkan kelompok pelanggan yang berbeda secara karakteristik dan relevan secara bisnis, sehingga dapat dijadikan dasar bagi pengambilan keputusan dalam merancang strategi pemasaran yang lebih adaptif dan efektif.

4.6. Perbandingan dengan Penelitian Sebelumnya

Penelitian ini menunjukkan bahwa segmentasi pelanggan menggunakan algoritma K-Means dengan nilai K=5 memberikan hasil optimal, sebagaimana ditunjukkan oleh Elbow Method dan nilai tertinggi Silhouette Score. Temuan ini sejalan dengan studi sebelumnya seperti A. Alamsyah [11] dan S.Sharyanto [12] yang juga menemukan nilai K optimal pada kisaran 4–6. Perbedaan nilai K antar penelitian dipengaruhi oleh struktur data dan konteks domain, namun pendekatan metodologis yang digunakan konsisten dan mengikuti praktik terbaik data mining. Visualisasi dua dimensi menggunakan Principal Component Analysis (PCA) menunjukkan kelima klaster terdistribusi dengan jelas dan tanpa tumpang tindih signifikan. Komponen utama PC1 dan PC2 menjelaskan seluruh variansi data (62% dan 38%), memperkuat validitas segmentasi. Selain itu, distribusi proporsi pelanggan per klaster antara 12,5%–25% menunjukkan segmentasi yang seimbang tanpa dominasi klaster tertentu. Keunggulan utama penelitian ini terletak pada kombinasi evaluasi klasterisasi yang komprehensif dan visualisasi mendalam menggunakan PCA serta diagram pie, yang bersama-sama meningkatkan keandalan, validitas, dan relevansi hasil untuk strategi pemasaran aktual.

KESIMPULAN

Penelitian ini berhasil menerapkan algoritma k-means untuk melakukan segmentasi pelanggan berdasarkan volume pembelian, dengan tujuan meningkatkan efektivitas pemasaran e-commerce. Menggunakan data transaksi dari kaggle, analisis menghasilkan lima klaster pelanggan yang berbeda, dengan klaster 0 sebagai kelompok dominan. Penentuan jumlah klaster optimal dilakukan dengan dua pendekatan evaluatif: elbow method dan silhouette score. Keduanya menyimpulkan bahwa jumlah klaster terbaik adalah lima (k=5), karena pada titik tersebut nilai wcss mulai stabil dan nilai silhouette mencapai angka tertinggi sebesar 0.81. Visualisasi hasil klasterisasi dengan pca menunjukkan pemisahan yang relatif jelas antar klaster, walaupun ada sedikit tumpang tindih. Meskipun hasil menunjukkan keberhasilan dalam segmentasi, penelitian ini menghadapi keterbatasan seperti distribusi klaster yang kurang merata dan overlap antara klaster. Oleh karena itu, studi lanjutan disarankan untuk memperkaya atribut data serta membandingkan kinerja k-means dengan algoritma lain seperti dbscan. Secara umum, segmentasi berbasis k-means memberikan manfaat besar dalam memahami karakteristik pelanggan, yang pada akhirnya dapat digunakan sebagai dasar pengambilan keputusan pemasaran yang lebih tepat sasaran dan berbasis data mixed-methods untuk mengintegrasikan wawasan kualitatif guna memperkaya interpretasi segmen pelanggan.

DAFTAR PUSTAKA

[1] A. R. Sinaga and G. D. Pranata, “Clustering Data Penjualan Produk pada Toko Yudha dengan Algoritma K-Means,” *J. Univ. Inform. dan Bisnis Indones.*, 2021, doi: 10.37278/sisinfo.v3i2.638.

[2] B. Apriyanto and S. L. M. Sitio, “Penerapan K-Means dalam Menganalisis Pola Pembelian Pelanggan Pada Data Transaksi E-Commerce,” *bit-Tech*, vol. 7, no. 3, pp. 790–797, Apr. 2025, doi: 10.32877/bt.v7i3.2195.

[3] M. Kasmi and J. Wayan, “PRINSIP-PRINSIP PEMASARAN Pangkep and the Islands State Polytechnic for Agriculture,” 2023. [Online]. Available: https://www.researchgate.net/publication/372717306

[4] I. Syafrinal and E. L. Febrianti, “PENERAPAN ALGORITMA K-MEANS PADA APLIKASI DATA MINING UNTUK MENENTUKAN POLA PENJUALAN (STUDI KASUS: ZAHRA MART),” *J. Ilm. Digit. Inf. Technol.*, vol. 13, no. 1, pp. 31–40, 2023, doi: https://doi.org/10.51920/jd.v13i1.320.

[5] M. Fajar, M. Adji, and G. Dwilestari, “ANALISIS DATA TRANSAKSI PENJUALAN BARANG MENGGUNAKAN TEKNIK K-MEANS CLUSTERING,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 1, pp. 619–625, 2025, doi: https://doi.org/10.36040/jati.v9i1.12433.

[6] E. Febrianty, L. Awalina, and W. I. Rahayu, “Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail Optimizing Marketing Strategies with Customer Segmentation Using K-Means Clustering on Online Retail Transactions,” *J. Teknol. dan Inf.*, vol. 13, 2023, doi: 10.34010/jati.v13i2.

[7] T. M. Dista and F. F. Abdulloh, “Clustering Pengunjung Mall Menggunakan Metode K-Means dan Particle Swarm Optimization,” *J. MEDIA Inform. BUDIDARMA*, vol. 6, no. 3, p. 1339, Jul. 2022, doi: 10.30865/mib.v6i3.4172.

[8] A. Yani, Z. Azmi, D. Suherdi, S. Informasi, and S. Triguna Dharma, “Implementasi Data Mining Menganalisa Data Penjualan Menggunakan Algoritma K-Means Clustering,” *Maret*, vol. 2, no. 2, pp. 315–323, 2023, doi: 10.53513/jursi.v2i2.6357.

[9] K. Pola *et al.*, “Clustering Sales Patterns of Best Selling and Less Selling Products at El Jhon Bengkulu Stores Using the K-Medoid Method,” *J. Kom.*, vol. 2, no. 2, pp. 637–642, 2022, doi: 10.53697/jkomitek.v2i2.

[10] U. Arfan and N. Paraga, “Perbandingan Algoritma K-Means, Naïve Bayes dan Decision Tree Dalam Memprediksi Penjualan Bahan Bakar Minyak,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 4, pp. 1379–1389, Jul. 2024, doi: 10.57152/malcom.v4i4.1566.

[11] A. Alamsyah *et al.*, “Customer Segmentation Using the Integration of the Recency Frequency Monetary Model and the K-Means Cluster Algorithm,” *Sci. J. Informatics*, vol. 9, no. 2, pp. 189–196, 2022, doi: 10.15294/sji.v9i2.39437.

[12] S. Sharyanto and D. Lestari, “Penerapan Data Mining Untuk Menentukan Segmentasi Pelanggan Dengan Menggunakan Algoritma K-Means dan Model RFM Pada E-Commerce,” *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 4, p. 866, 2022, doi: 10.30865/jurikom.v9i4.4525.

[13] A. Ristyawan, A. Nugroho, and T. K. Amarya, “Optimasi Preprocessing Model Random Forest Untuk Prediksi Stroke,” vol. 12, no. 1, pp. 29–44, 2025.

[14] H. H. Q. Hayqal, Oni Soesanto, and Yuana Sukmawaty, “K-Means Clustering dan Principal Component Analysis (PCA) Dalam Radial Basis Function Neural Network (RBFNN) Untuk Klasifikasi Data Multivariat,” *J. Math. Theory Appl.*, vol. 4, no. 1, pp. 1–7, 2022, doi: 10.31605/jomta.v4i1.1757.

[15] M. Hilmy Naufan, R. Kurniawan, and T. Suprapti, “OPTIMASI NILAI DAVIES BOULDIN INDEX PADA PROGRAM PENDAFTARAN TANAH SISTEMATIS LENGKAP (PTSL) MENGGUNAKAN ALGORITMA K-MEANS DAN PCA,” vol. 20, pp. 17–28, 2025, doi: 10.30587/e-link.v20i1.9063.