



Segmentasi Pelanggan E-Commerce Berdasarkan Pola Pembatalan dan Pengembalian Pesanan Menggunakan K-Means

Yulia Natasya Farah Diba Arifin^{1*}, Anis Zubair²

^{1,2} Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Merdeka Malang, Indonesia

Info Artikel

Riwayat Artikel:

Diterima : **16 Desember 2025**

Direvisi : **17 Desember 2025**

Diterbitkan : **27 Desember 2025**

Kata Kunci:

Segmentasi Pelanggan,
K-Means,
Pembatalan & Pengembalian,
Perilaku Pelanggan,

Keywords:

Customer Segmentation,
K-Means,
Cancellation & Return,
Customer Behavior

ABSTRAK

Penelitian ini mengkaji segmentasi pelanggan *e-commerce* berdasarkan perilaku kritis pembatalan (*cancellation*) dan pengembalian (*return*) menggunakan algoritma *K-Means* sebagai *proof-of-concept*. Menggunakan Pakistan *E-Commerce* Dataset tahun 2017, penelitian melakukan pra-pemrosesan, rekayasa fitur perilaku (*Cancellation Rate*, *Return Rate*, *Average Order Value*, *Discount Sensitivity*, *Preferred Payment Method*, dan *Total Orders*), normalisasi Min-Max, dan pemodelan *K-Means*. Validasi jumlah kluster dilakukan dengan *Elbow Method*, *Silhouette Score*, serta verifikasi visual menggunakan PCA. Hasil menunjukkan K = 3 kluster yang stabil, *Price-Sensitive Customers* (69.48%), transaksi bernilai tinggi namun sensitif harga, *Loyal Customers* (13.55%), frekuensi dan konsistensi pembelian tinggi dengan CR/RR rendah, dan *High-Risk Customers* (16.97%), tingkat pengembalian tinggi dengan kontribusi nilai rendah. Temuan membuktikan bahwa *K-Means* efektif untuk mengidentifikasi pola perilaku terkait *cancellation* dan *return* dan menyediakan dasar bagi rekomendasi konseptual untuk manajemen risiko dan analisis lanjutan.

ABSTRACT

This study examines e-commerce customer segmentation based on cancellation and return behaviors using K-Means clustering as a proof-of-concept. Using the Pakistan E-Commerce Dataset (2017), we performed preprocessing, behavioral feature engineering (Cancellation Rate, Return Rate, Average Order Value, Discount Sensitivity, Preferred Payment Method, and Total Orders), Min-Max normalization, and K-Means modeling. Cluster number validation relied on the Elbow Method, Silhouette Score, and PCA visualization. Results indicate K = 3 stable clusters: Price-Sensitive Customers (69.48%) high per-order value but price-sensitive; Loyal Customers (13.55%), high frequency and low CR/RR; and High-Risk Customers (16.97%), high return rate with low value contribution. The findings demonstrate K-Means' effectiveness in identifying cancellation/return patterns and provide a conceptual basis for risk management and further analysis.

Penulis Korespondensi:

Yulia Natasya Farah Diba Arifin,
Program Studi Sistem Informasi,
Universitas Merdeka Malang
Email: tasyafrdba@gmail.com

This is an open access article under the [CC BY-SA](#) license



1. PENDAHULUAN

Perkembangan sektor *e-commerce* global terus menunjukkan laju yang pesat. Pendapatan *e-commerce Business-to-Consumer* (B2C) dunia diproyeksikan mencapai sekitar USD 5,5 triliun pada tahun 2027 [1].

Meskipun demikian, tingkat penetrasi *e-commerce* di pasar berkembang, seperti Pakistan, masih relatif kecil. Pakistan hanya menyumbang sekitar 0,01% dari total volume *e-commerce* dunia [2]. Sebagai gambaran, volume *e-commerce* Pakistan diperkirakan mencapai US\$5,2 miliar pada tahun 2023 [3] dan diprediksi tumbuh menjadi US\$7,7 miliar pada tahun 2024 [2]. Mayoritas transaksi dilakukan melalui perangkat seluler, dengan sekitar 75% aktivitas bisnis online dan 91% konsumen berbelanja menggunakan smartphone [2], [3].

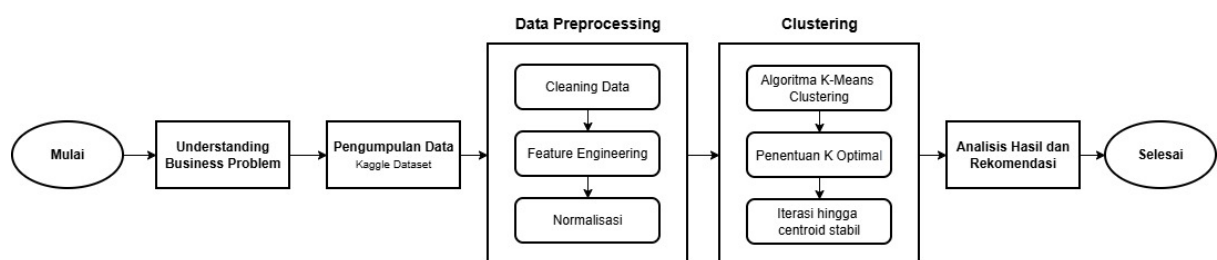
Isu krusial yang menonjol di pasar berkembang ini adalah dominasi metode pembayaran *Cash on Delivery* (COD), yang mencakup 75–95% volume transaksi *e-commerce* di Pakistan [2], [3]. Dominasi COD dipicu oleh rendahnya penetrasi perbankan (hanya sekitar 24% penduduk memiliki rekening bank) dan berkurangnya kepercayaan pada pembayaran digital [3]. Meskipun skema COD awalnya meningkatkan keyakinan pembeli, data empiris menunjukkan bahwa pesanan COD dua kali lebih mungkin dibatalkan atau dikembalikan dibandingkan pesanan prabayar [4]. Fenomena ini secara langsung terkait dengan disonansi kognitif dan *buyer's remorse* [5], [6], di mana pembeli memiliki waktu tunda antara keputusan pembelian dan pembayaran, yang meningkatkan potensi penyesalan dan, pada akhirnya, memicu pembatalan atau pengembalian produk. Perilaku konsumen Pakistan mencerminkan risiko operasional yang tinggi, misalnya, total nilai pembatalan pesanan mencapai lebih dari PKR 2 miliar, mayoritas berasal dari kategori fesyen dan elektronik [4].

Penelitian ini memfokuskan pada Pakistan *E-Commerce Dataset* tahun 2017 [7], khususnya pada kategori *Mobile & Tablets*, karena tercatat memiliki tingkat pembatalan dan pengembalian pesanan tertinggi, mencapai sekitar 50% dari total transaksi pada periode tersebut. Angka kegagalan yang signifikan ini menuntut identifikasi pola perilaku pelanggan yang berbeda. Kajian terdahulu mengenai segmentasi pelanggan umumnya berfokus pada perilaku pembelian dan nilai pelanggan seperti model *Recency, Frequency, Monetary* (RFM) untuk tujuan pemasaran dan retensi [8]. Meskipun terdapat penelitian tentang probabilitas pengembalian produk [9], segmentasi yang secara simultan menganalisis perilaku pembatalan (*cancellation*) dan pengembalian (*return*) masih relatif jarang ditemukan. Segmentasi yang berorientasi pada risiko operasional ini sangat penting untuk memahami pola kerugian finansial dan biaya logistik balik (*reverse logistics*) [10].

Oleh karena itu, penelitian ini bertujuan untuk mengisi kesenjangan literatur dengan mengintegrasikan dua dimensi perilaku risiko tersebut, yaitu pembatalan (*cancellation*) dan pengembalian (*return*), sebagai fitur utama segmentasi. Algoritma *K-Means Clustering* dipilih karena efektivitasnya dalam pengelompokan data perilaku berskala besar [11]. Hasil segmentasi diharapkan dapat memberikan pemahaman komprehensif mengenai pola risiko yang berbeda pada setiap segmen pelanggan, yang secara konseptual dapat mendukung perumusan strategi manajemen risiko transaksi pada platform *e-commerce*.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode data mining, khususnya pada teknik clustering eksploratif (*unsupervised learning*), untuk mengelompokkan pelanggan berdasarkan perilaku pembatalan dan pengembalian pesanan. Pendekatan ini dipilih karena mampu mengolah data berukuran besar secara objektif dan menemukan pola tersembunyi (*hidden patterns*) yang tidak dapat diidentifikasi melalui analisis deskriptif biasa [12]. Secara keseluruhan, penelitian ini dirancang sebagai *proof of concept* penerapan algoritma *K-Means* pada kasus perilaku negatif pelanggan *e-commerce*. Tahapan-tahapan dalam desain penelitian ini dapat digambarkan secara visual melalui Gambar 1. yang menggambarkan alur proses mulai dari pemahaman masalah, pengumpulan data, pra-pemrosesan, hingga penerapan algoritma dan analisis hasil.



Gambar 1. Diagram Alur Penelitian

2.1. Understanding Business Problem

Tahap pertama bertujuan memahami konteks permasalahan, yaitu tingginya tingkat pembatalan (*cancellation*) dan pengembalian (*return*) pada transaksi *e-commerce* di Pakistan. Fokus penelitian dipersempit pada transaksi tahun 2017 dalam kategori *Mobile & Tablets* karena kategori ini tercatat memiliki tingkat *cancel* atau *return* tertinggi, mencapai sekitar 50% dari total pesanan [7].

Karakteristik pasar Pakistan yang didominasi metode pembayaran tunai (COD), hambatan logistik, serta rendahnya kepercayaan terhadap pembayaran digital turut memengaruhi pola *cancel* atau *return* [3]. Oleh karena itu, penelitian ini diposisikan sebagai pembuktian konsep penerapan *K-Means* dalam memetakan pola risiko pelanggan berdasarkan data historis tersebut.

2.2. Pengumpulan Data (Data Gathering)

Data penelitian ini diperoleh dari dataset publik berjudul Pakistan *E-Commerce* Dataset yang dipublikasikan oleh [7] melalui platform Kaggle. Dataset ini memuat catatan transaksi *e-commerce* di Pakistan pada periode Maret 2016 hingga Agustus 2018, yang terdiri dari berbagai informasi seperti tanggal transaksi, kategori produk, harga, diskon, jumlah barang, metode pembayaran, serta status akhir pesanan. Seluruh data disediakan dalam format CSV sehingga memungkinkan untuk diolah lebih lanjut menggunakan perangkat lunak analisis data.

Tahap eksplorasi ini memungkinkan peneliti memahami struktur dataset, jumlah kolom, jenis variabel, serta kondisi awal data yang masih belum dibersihkan. Visualisasi tampilan dataset mentah ditunjukkan pada Gambar 2. Pada penelitian ini, fokus analisis dibatasi hanya pada transaksi tahun 2017, karena periode tersebut memiliki volume transaksi yang stabil serta mencerminkan pola pembatalan dan pengembalian yang relevan untuk dianalisis secara komprehensif.

item_id	status	created_at	sku	price	qty_ordered	grand_total	increment	category_name	sales_comm	discount_an	payment_m	Working Date	BI State
379040.0	complete	2017-01-01	Rajesh_Black Glue Gun	825.0	1.0	1185.0	100255234	Home & Living	0.0	0.0	cod	1/1/2017	Net
379041.0	complete	2017-01-01	RAJESH_RAJ223	360.0	1.0	1185.0	100255234	Home & Living	0.0	0.0	cod	1/1/2017	Net
379042.0	complete	2017-01-01	Adolphs_BF-4M	799.0	1.0	799.0	100255235	Men's Fashion	0.0	0.0	cod	1/1/2017	Net
379044.0	complete	2017-01-01	BO_nonRC-Jeep-Red	250.0	1.0	599.0	100255236	Kids & Baby	0.0	0.0	cod	1/1/2017	Net
379045.0	complete	2017-01-01	BP_PS03940-Green	349.0	1.0	599.0	100255236	School & Educat	0.0	0.0	cod	1/1/2017	Net
379046.0	complete	2017-01-01	Paramount_9789696371342	395.0	1.0	395.0	100255237	School & Educat	0.0	0.0	cod	1/1/2017	Net
379047.0	complete	2017-01-01	Infinitix-Hot-S-Grey	16899.0	1.0	15899.0	100255238	Mobiles & Tablet HDD	100640	1000.0	Payaxis	1/1/2017	Net
379048.0	complete	2017-01-01	Wild_MicrotouchMaxTrimmer	550.0	1.0	920.0	100255239	Beauty & Groom	0.0	0.0	cod	1/1/2017	Net
379049.0	complete	2017-01-01	Seven Star_SB-Cotton4	370.0	1.0	920.0	100255239	Home & Living	0.0	0.0	cod	1/1/2017	Net
379050.0	complete	2017-01-01	MYWALET_MW-060-BLACK	615.0	1.0	615.0	100255240	Men's Fashion	0.0	0.0	cod	1/1/2017	Net
379051.0	complete	2017-01-01	vitamin_Ultra Whey Protein	1500.0	1.0	1500.0	100255241	Health & Sports	0.0	0.0	cod	1/1/2017	Net
379052.0	order_refunded	2017-01-01	KureshiC_047-kemel	1099.0	1.0	1099.0	100255242	Beauty & Groom	0.0	0.0	cod	1/1/2017	Valid
379053.0	complete	2017-01-01	lucane_Plus Glass Protector for iF 645	645.0	1.0	645.0	100255243	Mobiles & Tablet	0.0	0.0	cod	1/1/2017	Net
379054.0	complete	2017-01-01	aybeez_ABZ-1047-L	299.0	1.0	299.0	100255244	Men's Fashion	0.0	0.0	cod	1/1/2017	Net
379056.0	complete	2017-01-01	Audionic_B-710-Black	1075.0	1.0	1075.0	100255245	Computing	0.0	0.0	cod	1/1/2017	Net
379058.0	complete	2017-01-01	AC_ac 178-green-Free size	799.0	1.0	48.0	100255246	Women's Fashion	0.0	0.0	cod	1/1/2017	Net
379060.0	complete	2017-01-01	RS_Coconut Bites	140.0	1.0	140.0	100255247	Soghaat	C-HDD-40844	0.0	cod	1/1/2017	Net
379061.0	complete	2017-01-01	RS_Coconut Bites	140.0	1.0	140.0	100255248	Soghaat	C-HDD-40844	0.0	cod	1/1/2017	Net
379062.0	canceled	2017-01-01	Turbo_100184	91.0	1.0	91.0	100255249	Home & Living	0.0	0.0	Payaxis	1/1/2017	Gross
379063.0	order_refunded	2017-01-01	ss_Cufflink Golf Kit	1000.0	1.0	2000.0	100255250	Men's Fashion	80211	0.0	cod	Friday, 20 December 2024	

Gambar 2. Pakistan E-Commerce Dataset 2017

2.3. Pra-pemrosesan Data (Data Preprocessing)

Tahap pra-pemrosesan data dilakukan untuk memastikan bahwa data yang akan dianalisis berada dalam kondisi bersih, konsisten, dan sesuai untuk kebutuhan algoritma *K-Means Clustering*. Proses ini meliputi tiga langkah utama, yaitu *data cleaning*, *feature engineering*, dan normalisasi data. Seluruh tahapan ini dilakukan untuk mengurangi *noise*, memperbaiki kualitas data, dan memastikan setiap variabel yang digunakan memiliki kontribusi yang proporsional terhadap pembentukan kluster.

a. Data Cleaning

Tahap data cleaning dilakukan untuk mengatasi berbagai ketidakonsistenan yang ditemukan pada dataset mentah [13]. Dataset awal transaksi tahun 2017 semula terdiri dari 290.920 baris data dengan 61.957 pelanggan unik. Namun, setelah implementasi kriteria *purposive sampling* yang ketat (fokus pada kategori *Mobile & Tablets*) dan penyaringan berbasis kelengkapan atribut, terjadi penyusutan signifikan menjadi 3.119 baris data dari 1.196 pelanggan unik yang memenuhi kriteria analisis. Penyusutan substansial ini merupakan

konsekuensi metodologis yang disengaja untuk memfokuskan analisis pada segmen pasar yang paling relevan dengan masalah penelitian. Proses pembersihan meliputi: (1) Eliminasi kolom non-esensial (misalnya *working_date*, *Customer Since*, *FY*, *BI Status*); (2) Penanganan inkonsistensi seperti baris duplikat, nilai kosong pada atribut kunci, dan koreksi nilai *grand_total* yang tidak sesuai dengan perhitungan *price × qty_ordered*; dan (3) Standardisasi status transaksi menjadi tiga kelompok utama (*complete*, *canceled*, *return*). Data *outlier* yang masih dapat dipertanggungjawabkan tetap dipertahankan. Proses ini menghasilkan dataset akhir yang memiliki kualitas tinggi, konsisten, dan representatif untuk pemodelan kluster berbasis perilaku.

b. Feature Engineering

Tahap *feature engineering* dilakukan dengan membentuk variabel-variabel baru yang relevan untuk kebutuhan analisis clustering. Variabel ini dirancang untuk menggambarkan perilaku pelanggan terkait pembatalan dan pengembalian pesanan. Terdapat lima fitur utama yang dibentuk, yaitu:

$$\text{Cancellation Rate (CR)} = \frac{\text{Jumlah Pesanan Dibatalkan}}{\text{Total Pesanan per Pelanggan}} \quad (1)$$

$$\text{Return Rate (RR)} = \frac{\text{Jumlah Pesanan Dikembalikan}}{\text{Total Pesanan per Pelanggan}} \quad (2)$$

$$\text{Average Order Value (AOV)} = \frac{\text{Total Nilai Transaksi}}{\text{Jumlah Pesanan per Pelanggan}} \quad (3)$$

$$\text{Discount Sensitivity (DS)} = \frac{\text{Total Diskon}}{\text{Total Nilai Transaksi}} \times 100\% \quad (4)$$

Fitur lain seperti *Total Orders* dan *preferred payment method* ditentukan berdasarkan metode pembayaran yang paling sering digunakan oleh pelanggan. Fitur-fitur tersebut dipilih karena secara langsung berhubungan dengan tujuan penelitian, yaitu mengidentifikasi karakteristik pelanggan berdasarkan tingkat pembatalan, pengembalian, dan perilaku konsumsi mereka [13].

Seluruh fitur dihitung menggunakan fungsi numerik pada *Python*, dan contoh hasil perhitungan ditampilkan pada Gambar 3. Hasil perhitungan ini digunakan sebagai dasar untuk *preprocessing* lanjutan dan sebagai input utama pada algoritma *K-Means*.



```
# cancellation rate (CR)
prep['CR'] = prep['is_cancel'] / prep['total_orders']

# return rate (RR)
prep['RR'] = prep['is_return'] / prep['total_orders']

# average order value (AOV)
prep['AOV'] = prep['grand_transaction'] / prep['total_orders']

# discount sensitivity (DS)
prep['DS'] = (prep['grand_discount'] / prep['grand_transaction']) * 100

# input NaN, Inf -> 0
prep['DS'] = prep['DS'].replace([np.inf, -np.inf], np.nan).fillna(0)

# preferred payment method (PPM)
PPM = df.groupby('Customer ID')['payment_method'].agg(lambda x: x.mode()[0]).reset_index(name='PPM')

# merging PPM
result = prep.merge(PPM, on='Customer ID', how='left')

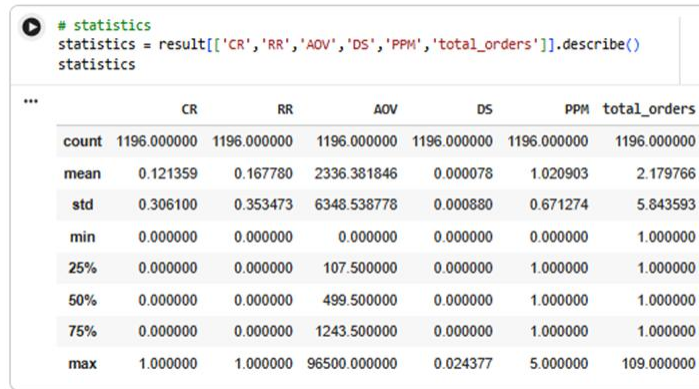
# encoding PPM
encoder = LabelEncoder()
result['PPM'] = encoder.fit_transform(result['PPM'])

# output
result = result[['Customer ID', 'CR', 'RR', 'AOV', 'DS', 'PPM', 'total_orders']]
result.head()
```

	Customer ID	CR	RR	AOV	DS	PPM	total_orders
0	4	0.333333	0.333333	181.111111	0.003630	1	9
1	16	0.000000	0.000000	100.000000	0.000000	1	1
2	20	0.166667	0.166667	1175.333333	0.001126	1	6
3	28	0.000000	0.000000	690.000000	0.000000	1	1
4	32	0.000000	0.000000	286.000000	0.000000	1	2

Gambar 3. Contoh Hasil Perhitungan Feature Engineering

Selanjutnya dilakukan analisis statistik deskriptif untuk seluruh fitur yang telah dibentuk guna memahami distribusi awal data sebelum proses normalisasi. Analisis ini mencakup nilai minimum, maksimum, rata-rata, dan standar deviasi. Hasil statistik ditunjukkan pada Gambar 4 dan memperlihatkan bahwa setiap fitur memiliki skala dan variasi yang berbeda. Misalnya, nilai AOV memiliki rentang yang sangat besar (min = 0, max = 96.500), sedangkan fitur DS dan CR berada pada skala yang jauh lebih kecil. Perbedaan rentang inilah yang mengonfirmasi dibutuhkan proses normalisasi agar seluruh variabel berada pada skala yang sebanding dan tidak menimbulkan bias dalam proses pembentukan kluster.



```
# statistics
statistics = result[['CR', 'RR', 'AOV', 'DS', 'PPM', 'total_orders']].describe()
statistics
```

	CR	RR	AOV	DS	PPM	total_orders
count	1196.000000	1196.000000	1196.000000	1196.000000	1196.000000	1196.000000
mean	0.121359	0.167780	2336.381846	0.000078	1.020903	2.179766
std	0.306100	0.353473	6348.538778	0.000880	0.671274	5.843593
min	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	0.000000	0.000000	107.500000	0.000000	1.000000	1.000000
50%	0.000000	0.000000	499.500000	0.000000	1.000000	1.000000
75%	0.000000	0.000000	1243.500000	0.000000	1.000000	1.000000
max	1.000000	1.000000	96500.000000	0.024377	5.000000	109.000000

Gambar 4. Statistik Deskriptif Hasil *Feature Engineering*

c. Normalisasi Data

Tahap normalisasi data dilakukan untuk menyeragamkan skala variabel, mengingat algoritma *K-Means* sensitif terhadap perbedaan skala antar-fitur. Normalisasi dilakukan menggunakan metode *Min-Max Scaling*, yang mengubah setiap nilai fitur ke dalam rentang 0–1. Metode ini dipilih karena sederhana, efektif, dan mempertahankan distribusi asli data. Rumus normalisasi:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (5)$$

Sebelum normalisasi, masih terlihat perbedaan yang signifikan pada skala setiap variabel, sebagaimana ditampilkan pada Tabel 1. Setelah dilakukan normalisasi, seluruh variabel telah berada dalam rentang 0–1 dan siap digunakan dalam proses *clustering*. Hasil normalisasi ditampilkan pada Tabel 2.

Tabel 1. Data Sebelum Normalisasi

Cust ID	CR	RR	AOV	DS	PPM	total_orders
4	0.3333	0.3333	181.11	0.00363	1	9
16	0.0000	0.0000	100.00	0.00000	1	1
20	0.1667	0.1667	1175.33	0.00126	1	6
28	0.0000	0.0000	690.00	0.00000	1	1
32	0.0000	0.0000	286.00	0.00000	1	2

Tabel 2. Data Sesudah Normalisasi

Cust ID	CR	RR	AOV	DS	PPM	total_orders
4	0.3333	0.3333	0.001877	0.14896	0.2	0.074074
16	0.0000	0.0000	0.001036	0.00000	0.2	0.000000
20	0.1667	0.1667	0.012810	0.046171	0.2	0.046296
28	0.0000	0.0000	0.007150	0.00000	0.2	0.000000
32	0.0000	0.0000	0.002964	0.00000	0.2	0.009259

2.4. Penerapan Algoritma *Clustering*

Setelah seluruh data selesai melalui proses *cleaning*, *feature engineering*, dan normalisasi, langkah berikutnya adalah menerapkan algoritma *K-Means Clustering* untuk mengelompokkan pelanggan berdasarkan kesamaan pola perilaku transaksi mereka. Algoritma *K-Means* merupakan salah satu teknik *unsupervised learning* yang paling banyak digunakan untuk segmentasi karena kemampuannya membentuk kluster secara efisien dengan meminimalkan jarak antar data dalam kluster yang sama [14].

Secara prinsip, algoritma *K-Means* bekerja dengan menentukan terlebih dahulu jumlah kluster *K*. Selanjutnya, algoritma memilih titik pusat awal (*initial centroids*) secara acak, lalu menghitung jarak setiap data terhadap centroid menggunakan *Euclidean Distance*. Setiap data ditempatkan pada kluster dengan jarak terdekat, kemudian centroid diperbarui berdasarkan rata-rata posisi anggota kluster. Proses ini berlangsung secara iteratif hingga posisi centroid stabil atau perubahan yang terjadi sangat kecil. Secara matematis, fungsi objektif *K-Means* dinyatakan sebagai berikut:

$$J = \sum_{i=1}^K \sum_{x_j \in C_i} \|x_j - \mu_i\|^2 \quad (6)$$

dengan keterangan:

- J = total *within-cluster sum of squares* (WCSS), yaitu total jarak kuadrat antara setiap data x_j dan centroid klaster-nya.
- K = jumlah klaster.
- C_i = himpunan data yang termasuk dalam klaster ke- i .
- μ_i = centroid (titik tengah) dari klaster ke- i .

Tujuan utama algoritma ini adalah meminimalkan variansi internal klaster (*intra-cluster variance*) sehingga anggota dalam satu klaster memiliki tingkat kemiripan yang tinggi, sedangkan perbedaan antar-klaster meningkat secara signifikan [15]. Penentuan jumlah klaster optimal dilakukan menggunakan *Elbow Method* dan *Silhouette Analysis*, kemudian diverifikasi melalui visualisasi *Principal Component Analysis* (PCA). *Elbow Method* mengidentifikasi jumlah klaster berdasarkan titik "siku" pada grafik WCSS, yaitu titik ketika penurunan WCSS mulai melambat meskipun nilai K meningkat. *Silhouette Analysis* digunakan untuk mengevaluasi kualitas pemisahan klaster melalui skor *Silhouette* yang berkisar dari -1 hingga $+1$, nilai mendekati $+1$ mengindikasikan struktur klaster yang semakin baik [16].

Dalam beberapa kasus, kedua metode tersebut dapat memberikan rekomendasi yang berbeda. Oleh karena itu, penelitian ini menggunakan PCA untuk mereduksi enam fitur menjadi dua komponen utama sehingga pola pemisahan klaster dapat diamati secara visual. Visualisasi PCA membantu memastikan bahwa klaster yang terbentuk memiliki batas yang jelas dan dapat diinterpretasikan secara statistik maupun bisnis [17], [18]. Evaluasi akhir dilakukan menggunakan *Silhouette Score* untuk menilai kualitas pemisahan klaster dan *Within-Cluster Sum of Squares* (WCSS) untuk mengukur kekompakan internal klaster. Nilai *Silhouette* yang lebih tinggi dan WCSS yang lebih rendah menunjukkan struktur klaster yang semakin baik dan homogen [19].

2.5. Analisis Hasil Clustering

Tahap ini bertujuan memberikan interpretasi menyeluruh terhadap hasil klasterisasi yang diperoleh dari algoritma *K-Means*. Analisis dilakukan menggunakan statistik deskriptif seperti nilai rata-rata, median, dan standar deviasi pada variabel utama penelitian, yaitu *Cancellation Rate* (CR), *Return Rate* (RR), *Average Order Value* (AOV), *Discount Sensitivity* (DS), serta *Preferred Payment Method* (PPM). Melalui evaluasi terhadap pola distribusi setiap variabel di masing-masing klaster, karakteristik dominan dapat diidentifikasi sehingga perbedaan perilaku pelanggan antar-segmen terlihat jelas.

Hasil analisis menunjukkan bahwa segmen dengan nilai CR dan RR tertinggi membentuk kelompok *high-risk customers*, yaitu pelanggan dengan kecenderungan pembatalan dan pengembalian yang tinggi. Sebaliknya, klaster dengan nilai transaksi besar dan tingkat risiko rendah diinterpretasikan sebagai *loyal customers* yang berpotensi memberikan kontribusi jangka panjang. Sementara itu, klaster dengan sensitivitas diskon tinggi mengindikasikan *price-sensitive customers* yang lebih responsif terhadap strategi promosi.

Temuan dari ketiga klaster ini tidak diarahkan untuk penyusunan strategi operasional secara langsung, melainkan digunakan sebagai dasar rekomendasi konseptual. Rekomendasi tersebut mencakup pemahaman pola perilaku tiap segmen, identifikasi potensi risiko berdasarkan karakteristik klaster, serta arah pengembangan analisis lanjutan dalam konteks *e-commerce*. Dengan demikian, tahap analisis ini berfungsi memperkuat tujuan penelitian sebagai pembuktian konsep penerapan *K-Means* dalam memetakan pola *cancellation* dan *return* secara akademik.

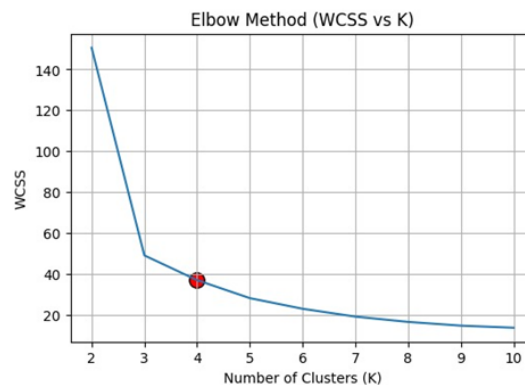
3. HASIL DAN ANALISIS

Bagian ini menyajikan hasil proses klasterisasi yang diperoleh dari penerapan algoritma *K-Means* pada data pelanggan *e-commerce* tahun 2017. Seluruh analisis dilakukan berdasarkan serangkaian tahapan mulai dari penentuan jumlah klaster optimal, evaluasi kualitas segmentasi, hingga interpretasi karakteristik tiap

klaster dengan tujuan memberikan gambaran konseptual mengenai pola *cancellation* dan *return* pada pelanggan.

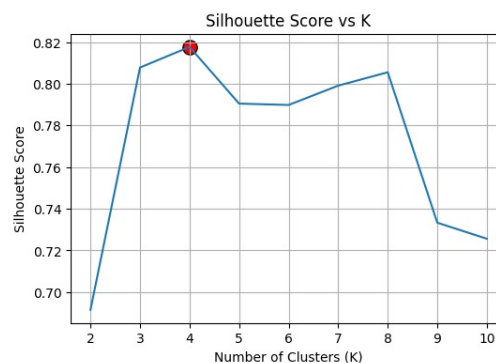
3.1. Penentuan Jumlah Klaster (K)

Penentuan jumlah klaster merupakan langkah krusial dalam proses *K-Means* karena nilai K menentukan struktur segmentasi yang dihasilkan. Analisis awal dilakukan menggunakan *Elbow Method* dan *Silhouette Score* sebagai acuan kuantitatif. Pada Gambar 5 kurva WCSS menunjukkan bahwa penurunan nilai WCSS dari K=2 ke K=3 cukup signifikan, sedangkan setelah K=4 penurunannya mulai melandai. Titik siku terlihat pada K=4, yang secara numerik menunjukkan adanya peningkatan kompaksi klaster meskipun manfaat tambahan setelah titik ini relatif kecil (*diminishing returns*).



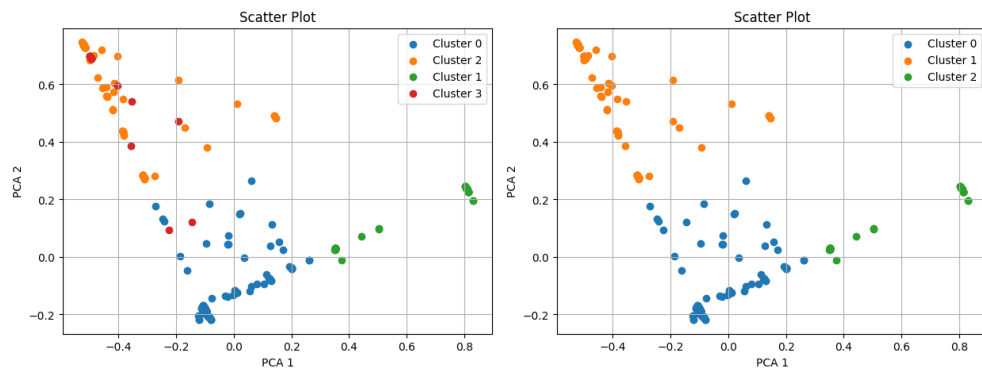
Gambar 5. Grafik *Elbow Method* (WCSS vs K)

Selanjutnya, Gambar 6 menampilkan nilai *Silhouette* untuk beberapa nilai K. Nilai tertinggi berada di sekitar K=4 dengan skor sekitar 0,82, menunjukkan pemisahan klaster yang baik menurut metrik internal ini. Namun, kenaikan skor *Silhouette* dari K=3 ke K=4 tidak terlalu jauh, dan setelah K=4 skor mulai menurun, sehingga K=4 hanya muncul sebagai kandidat, bukan keputusan final.



Gambar 6. Grafik *Silhouette Score* (nilai rata-rata skor vs K)

Meskipun demikian, literatur menegaskan bahwa baik *Elbow* maupun *Silhouette* bersifat *heuristic* dan tidak selalu mencerminkan struktur alami data [20], [21]. Oleh karena itu, analisis ini dilengkapi dengan validasi visual menggunakan *Principal Component Analysis* (PCA). Hasil visualisasi pada Gambar 7 menunjukkan bahwa pada K=4, klaster tambahan tidak membentuk kelompok yang jelas dan cenderung tumpang tindih dengan klaster lain, mengindikasikan terjadinya *forced clustering*, yaitu kondisi ketika algoritma "dipaksa" membentuk klaster yang sebenarnya tidak memiliki identitas struktural [22].

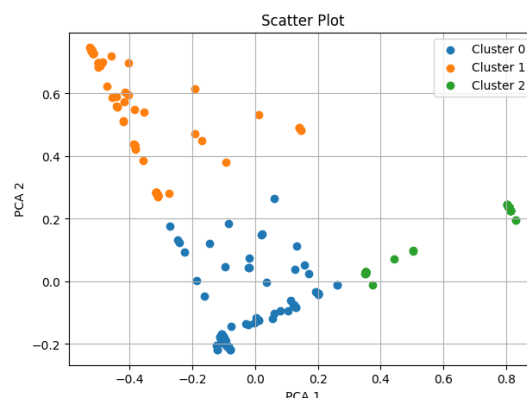


Gambar 7. Perbandingan Implementasi K=4 dan K=3

Sebaliknya, visualisasi PCA untuk K=3 memperlihatkan tiga kelompok yang lebih kompak, terpisah, dan stabil secara spasial. Ketiga klaster menempati area yang konsisten, sehingga lebih representatif terhadap pola alami data. Dengan mempertimbangkan kombinasi metrik kuantitatif dan hasil validasi visual tersebut, K=3 dipilih sebagai jumlah klaster final karena memberikan struktur segmentasi yang paling stabil serta dapat diinterpretasikan secara statistik maupun bisnis.

3.2. Hasil Clustering (Segmentasi Pelanggan)

Setelah jumlah klaster ditetapkan sebanyak tiga, analisis dilanjutkan untuk melihat pola pembentukan segmen pelanggan. Visualisasi PCA pada Gambar 8 menunjukkan pemisahan klaster yang cukup jelas, di mana Klaster 0 tersebar luas di area tengah, Klaster 1 lebih kompak di sisi kiri-atas, dan Klaster 2 berada di sisi kanan dengan distribusi yang lebih sempit. Pola ini mengindikasikan bahwa masing-masing klaster memiliki karakteristik yang berbeda dan dapat dibedakan secara visual.



Gambar 8. Hasil Clustering menggunakan K=3

Distribusi jumlah pelanggan pada masing-masing klaster dapat dilihat pada Tabel 3. Klaster 0 merupakan kelompok terbesar dengan 69,48% pelanggan, diikuti Klaster 2 sebesar 16,97%, sedangkan Klaster 1 merupakan kelompok terkecil dengan 13,55%. Proporsi ini menunjukkan struktur klaster yang stabil dan tidak mengalami ketimpangan ekstrem.

Tabel 3. Proporsi Pelanggan per Klaster

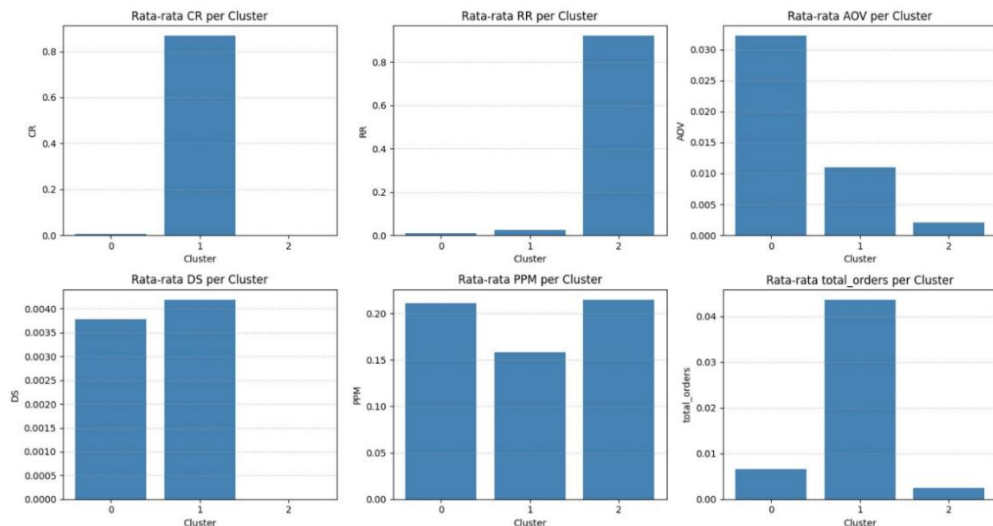
Cluster	Jumlah Customer	Presentase (%)
0	831	69.48
1	162	13.55
2	203	16.97

Analisis rata-rata variabel pada Gambar 8 dan Tabel 4 menunjukkan bahwa setiap klaster memiliki perilaku yang berbeda secara konsisten. Klaster 0 memiliki AOV tertinggi dengan tingkat CR dan RR yang rendah, sehingga menggambarkan pelanggan bernilai besar dengan risiko minimal. Klaster 1 menunjukkan frekuensi transaksi (total_orders) tertinggi dan CR serta RR yang rendah, menandakan pelanggan paling aktif dan stabil. Sementara itu, Klaster 2 ditandai oleh RR tertinggi dengan nilai AOV, DS, CR, dan total_orders yang rendah,

sehingga menjadi kelompok pelanggan dengan risiko pengembalian terbesar namun kontribusi ekonomi paling kecil.

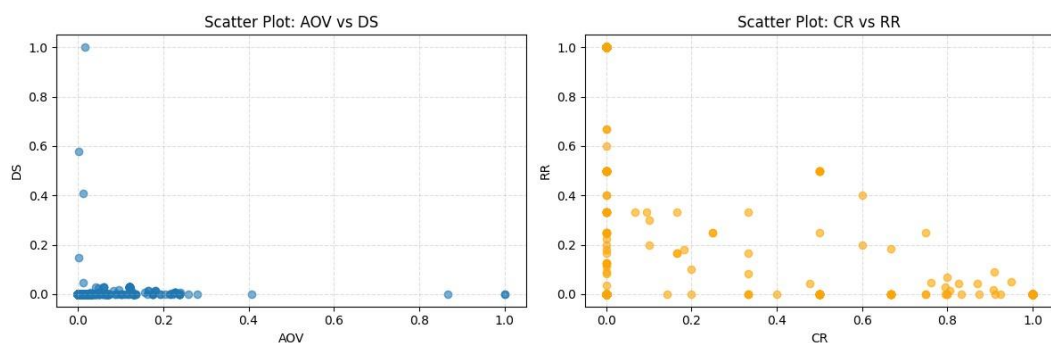
Tabel 4. Nilai Rata-Rata Variabel Tiap Klaster

Cluster	CR	RR	AOV	DS	PPM	total_orders
0	0.005399	0.011431	0.032207	0.003781	0.210590	0.006607
1	0.868265	0.026126	0.010948	0.004195	0.158025	0.043724
2	0.000000	0.920854	0.002085	0.000000	0.214778	0.002417



Gambar 9. Sebaran Rata-Rata Variabel per Cluster

Validasi tambahan dilakukan melalui scatter plot pada Gambar 9. Scatter plot AOV vs DS menunjukkan bahwa sebagian besar pelanggan memiliki AOV dan DS rendah, mengindikasikan transaksi bernilai kecil dan tidak bergantung pada diskon. Scatter plot CR vs RR memperlihatkan bahwa mayoritas pelanggan berada pada zona risiko rendah (CR rendah, RR rendah), tetapi terdapat dua pola risiko spesifik, pelanggan dengan CR rendah namun RR tinggi, serta pelanggan dengan CR tinggi namun RR rendah, yang menunjukkan heterogenitas perilaku pembatalan dan pengembalian.



Gambar 10. Scatter Plot Perbandingan Variabel

3.3. Karakteristik dan Profil Segmen Pelanggan

Hasil klasterisasi K-Means dengan K=3 menghasilkan tiga segmen pelanggan dengan karakteristik perilaku yang berbeda dan konsisten. Secara proporsi, Klaster 0 mendominasi dengan 69,48% pelanggan, diikuti Klaster 2 sebesar 16,97%, serta Klaster 1 sebagai kelompok terkecil dengan 13,55%. Pembagian ini merepresentasikan variasi perilaku pembelian, tingkat risiko, dan sensitivitas diskon yang muncul secara alami pada data.

a. Klaster 0 – Price-Sensitive Customers

Klaster ini merupakan kelompok terbesar dan dicirikan oleh nilai transaksi yang tinggi namun tidak konsisten. Pelanggan pada segmen ini memiliki AOV dan *Discount Sensitivity* tertinggi, sehingga keputusan

pembelian sangat dipengaruhi oleh insentif promosi. Meskipun demikian, CR dan RR relatif rendah, mengindikasikan bahwa ketika bertransaksi, pelanggan umumnya menyelesaikan pembelian tanpa pembatalan atau pengembalian. Secara bisnis, segmen ini tergolong *transactional customers* yang memberikan pendapatan besar per transaksi, tetapi memiliki ketergantungan tinggi pada promo dan tidak menunjukkan loyalitas jangka panjang.

b. Klaster 1 – *Loyal Customers*

Segmen ini memiliki profil paling stabil dan paling bernilai bagi perusahaan. Pelanggan dalam klaster ini menunjukkan frekuensi pembelian tertinggi, disertai CR dan RR yang rendah, serta sensitivitas diskon yang minimal. Nilai AOV berada pada tingkat sedang, tetapi kontribusi jangka panjangnya sangat signifikan. Klaster ini merupakan kelompok dengan *Customer Lifetime Value* tertinggi karena perilaku pembelian konsisten dan minim risiko, sehingga layak menjadi prioritas dalam strategi retensi.

c. Klaster 2 – *High-Risk Customers*

Klaster ini mencerminkan pelanggan berisiko tinggi karena memiliki Return Rate tertinggi dibanding klaster lain, namun AOV, total orders, dan CR berada pada tingkat rendah. Pola ini menunjukkan rendahnya kontribusi ekonomi tetapi tingginya potensi biaya operasional akibat proses pengembalian. Segmen ini membutuhkan perhatian khusus terkait penyebab ketidakpuasan atau *mismatch* produk, karena tingginya RR tidak diimbangi nilai transaksi yang memadai.

3.4. Rekomendasi Berdasarkan Pola Karakteristik Pelanggan

Rekomendasi yang disusun pada tahap ini berfungsi untuk menerjemahkan temuan klasterisasi ke dalam implikasi manajerial yang dapat diuji dan diterapkan secara langsung, tanpa mengubah ruang lingkup penelitian yang bersifat konseptual. Segmentasi K-Means mengungkap tiga kelompok pelanggan dengan pola perilaku yang berbeda, yang memberikan arah intervensi taktis yang terdiferensiasi untuk manajemen risiko dan optimalisasi nilai pelanggan.

a. Klaster 0 – *Price-Sensitive Customers*

Pelanggan dalam klaster ini dicirikan oleh nilai transaksi yang tinggi serta sensitivitas diskon yang sangat besar (AOV tinggi, DS tinggi), dengan risiko utama *margin erosion* akibat ketergantungan pada insentif promosi. Implikasi Taktis: Strategi harus diarahkan pada optimalisasi profitabilitas alih-alih sekadar volume penjualan. Hal ini melibatkan pengembangan model untuk menentukan ambang batas diskon yang optimal (*discount capping model*) yang masih efektif memicu pembelian besar tanpa menggerus margin keuntungan secara berlebihan. Secara konseptual, intervensi ini bertujuan mengalihkan nilai insentif dari potongan harga eksplisit ke *value-added services* (misalnya *free premium shipping*).

b. Klaster 1 – *Loyal Customers*

Segmen ini adalah aset strategis, ditunjukkan dengan frekuensi pembelian tertinggi, konsistensi perilaku, dan tingkat risiko operasional yang minimal (CR/RR rendah). Kelompok ini sensitif terhadap gangguan *customer experience*. Implikasi Taktis: Fokus harus pada strategi retensi berbasis nilai jangka panjang (*high-value retention*). Hal ini mencakup implementasi jalur layanan prioritas eksklusif (*priority support*) dan peningkatan jaminan kualitas secara otomatis (misalnya perpanjangan garansi atau kemudahan pengembalian yang disederhanakan). Secara konseptual, intervensi ini bertujuan mempertahankan utilitas pelanggan tertinggi, menjadikan mereka sebagai *benchmark* stabilitas perilaku dalam data analitik.

c. Klaster 2 – *High-Risk Customers*

Klaster ini merupakan liabilitas operasional, ditandai dengan Return Rate (RR) tertinggi dan kontribusi nilai transaksi yang minimal (AOV/TO terendah). Kelompok ini meningkatkan biaya reverse logistics secara signifikan. Implikasi Taktis: Implikasi langsung harus diarahkan pada pengujian mekanisme mitigasi risiko pembayaran untuk menekan biaya logistik yang timbul. Secara konseptual, hal ini dapat diuji coba melalui: (1) Pengujian Kebijakan *Prepaid-Only*: Membatasi pelanggan dalam segmen ini untuk menggunakan metode pembayaran prabayar (*prepaid-only*) pada transaksi berikutnya jika RR mereka melampaui ambang batas kerugian, sehingga memitigasi risiko *Return to Origin* (RTO) yang tinggi akibat COD. (2) Analisis Model Deposit: Menguji penerapan sistem deposit atau biaya penanganan pengembalian (*return processing fee*) minimal

sebagai prasyarat transaksi untuk menekan perilaku *impulsive return* (pengembalian impulsif). Intervensi ini secara akademik berakar pada prinsip pengalihan risiko operasional [23] dan secara langsung ditujukan untuk meningkatkan profitabilitas bersih dari segmen yang tidak menguntungkan.

Secara keseluruhan, hasil segmentasi menunjukkan bahwa algoritma *K-Means* mampu memetakan pola pembelian, pembatalan, dan pengembalian secara konsisten, serta mengidentifikasi pelanggan bernilai tinggi, kelompok sensitif harga, dan pelanggan dengan risiko operasional besar. Dengan demikian, segmentasi ini menyediakan dasar empirik yang kuat untuk merancang arah kebijakan retur dan pendekatan CRM yang berbeda antar segmen, yang dapat dijadikan hipotesis untuk pengujian strategi operasional di studi implementatif selanjutnya.

4. KESIMPULAN

Penelitian ini bertujuan mengidentifikasi pola pembatalan (*cancellation*) dan pengembalian (*return*) pesanan pada pelanggan *e-commerce* melalui pendekatan unsupervised learning menggunakan algoritma *K-Means*. Seluruh tahapan analisis telah menunjukkan bahwa tujuan penelitian yang dirumuskan pada bagian pendahuluan dapat dicapai dengan baik. Enam indikator perilaku pelanggan *Cancellation Rate*, *Return Rate*, *Average Order Value*, *Discount Sensitivity*, *Preferred Payment Method*, dan *Total Orders*, terbukti mampu membedakan karakteristik pelanggan secara signifikan, sehingga menjadi dasar yang kuat untuk proses klusterisasi.

Berdasarkan kombinasi evaluasi numerik (*Elbow Method* dan *Silhouette Score*) serta validasi visual PCA, jumlah kluster optimal ditetapkan sebanyak tiga kluster. Struktur $K=3$ memberikan pemisahan paling stabil dan representatif terhadap variasi perilaku pelanggan. Kluster hasil segmentasi menunjukkan tiga pola perilaku yang konsisten, *Price-Sensitive Customers* (nilai transaksi tinggi namun bergantung pada diskon), *Loyal Customers* (frekuensi pembelian tinggi dan sangat stabil), dan *High-Risk Customers* (tingkat pengembalian tinggi dengan kontribusi transaksi rendah). Hasil ini mengonfirmasi bahwa *K-Means* efektif digunakan untuk memetakan pola pembatalan dan pengembalian pesanan secara konseptual, sesuai ruang lingkup penelitian. Temuan ini menegaskan bahwa segmentasi berbasis perilaku risiko dapat menjadi dasar yang kuat untuk merancang arah intervensi taktis, seperti pengujian kebijakan pembayaran prabayar (*prepaid-only*) pada segmen berisiko tinggi untuk mitigasi biaya logistik balik.

Secara keseluruhan, penelitian ini membuktikan bahwa analisis kluster dapat menjadi alat pendukung pemahaman perilaku pelanggan *e-commerce*, khususnya terkait dinamika risiko dan potensi nilai tiap segmen. Meskipun signifikan, penelitian ini memiliki limitasi metodologis karena menggunakan data historis (Pakistan *E-Commerce Dataset* tahun 2017) yang membatasi generalisasi temuan ke kondisi pasar *e-commerce* yang lebih dinamis saat ini. Oleh karena itu, penelitian selanjutnya disarankan untuk fokus pada integrasi data transaksi terbaru (misalnya periode 2022–2025) di konteks geografis yang lebih relevan, uji komparasi performa *K-Means* dengan algoritma berbasis densitas (DBSCAN atau HDBSCAN) atau algoritma lain yang tidak mengasumsikan bentuk kluster sferis, serta pengembangan model klasifikasi prediktif yang menggunakan label kluster *High-Risk* sebagai *ground truth* untuk identifikasi risiko pelanggan secara *real-time*.

REFERENSI

- [1] ITA, "Pakistan - eCommerce," International Trade Administration. Accessed: Nov. 23, 2025. [Online]. Available: <https://www.trade.gov/country-commercial-guides/pakistan>
ecommerce#:~:text=grow%20steadily%20as%20there%20are,2%20billion%20in%202023
- [2] PCMI, "The Rise of E-Commerce in Pakistan 2024-2027," Payment & Commerce Market Intelligence. Accessed: Nov. 23, 2025. [Online]. Available: <https://paymentscmi.com/insights/pakistan-ecommerce-market-data/#:~:text=Challenges%20in%20the%20Market%3A>
- [3] ITA, "eCommerce Sales & Size Forecast," International Trade Administration. Accessed: Nov. 23, 2025. [Online]. Available: <https://www.trade.gov/ecommerce-sales-size-forecast#:~:text=Global%20B2C%20ecommerce%20revenue%20is,entertainment%2C%20beverages%2C%20and%20food>
- [4] Ecommerce, "Why So Many Fashion Orders Get Canceled or Returned in Pakistan's eCommerce Space." Accessed: Nov. 23, 2025. [Online]. Available: <https://ecommerce.com.pk/why-so-many-fashion-orders-get-canceled-or-returned-in-pakistans-ecommerce-space/#:~:text=In%202024%2C%20fashion%20returns%20in,that%20clothing%20brands%20handle%20daily>

- [5] I. N. Dewanti and I. Irwansyah, "Disonansi Kognitif Dalam Perilaku Konsumen Masyarakat Indonesia Terhadap Pembelian Produk Tanpa Logo Halal," *JURNAL LENSEA MUTIARA KOMUNIKASI*, vol. 5, no. 1, pp. 99–109, Jun. 2021, doi: 10.51544/jlmk.v5i1.1657.
- [6] A. Lochab, Dr. P. Khurana, Dr. M. Arora, S. Narula, and R. Kapoor, "Buy Now, Regret Later"? A Review Of Instant Gratification And Post Purchase Dissonance In E-Commerce," *SSRN Electronic Journal*, 2025, doi: 10.2139/ssrn.5399613.
- [7] Usmani and Sana Rasheed, "Pakistan's Largest E-Commerce Dataset," Kaggle. Accessed: Oct. 14, 2025. [Online]. Available: <https://www.kaggle.com/datasets/zusmani/pakistans-largest-ecommerce-dataset>
- [8] S. Kumar, R. Rani, S. K. Pippal, and R. Agrawal, "Customer segmentation in e-commerce: K-means vs hierarchical clustering," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 23, no. 1, pp. 119–128, Feb. 2025, doi: 10.12928/TELKOMNIKA.v23i1.26384.
- [9] J. P. Das, "Customer Segmentation Based on Probability of Product Return in E-Commerce Platform," in *Business Analytics and Intelligence: Driving Strategy with Data*, QTanalytics India, 2025, pp. 53–74. doi: 10.48001/978-81-980647-7-6-4.
- [10] A. B. Stevenson and J. Rieck, "Investigating Returns Management across E-Commerce Sectors and Countries: Trends, Perspectives, and Future Research," *Logistics*, vol. 8, no. 3, p. 82, Aug. 2024, doi: 10.3390/logistics8030082.
- [11] K. Auliasari and M. Kertaningtyas, "Penerapan Algoritma K-Means untuk Segmentasi Konsumen Menggunakan R," *Jurnal Teknologi dan Manajemen Informatika*, vol. 5, no. 2, Dec. 2019, doi: 10.26905/jtmi.v5i2.3644.
- [12] M. F. Fadhilah, A. Lovely, A. Suyoso, and I. Puspitasari, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Customer Segmentation with Clustering Algorithm Based on Recency, Frequency, and Monetary (RFM) Attributes Segmentasi Pelanggan dengan Algoritma Clustering Berdasarkan Atribut Recency, Frequency dan Monetary (RFM)," vol. 5, no. 1, pp. 48–56, 2025, doi: 10.57152/malcom.v5v1.1491.
- [13] R. Siagian, P. Sirait, and A. Halim, "Penerapan Algoritma K-Means dan K-Medoids untuk Segmentasi Pelanggan pada Data Transaksi E-Commerce," *SISTEMASI: Jurnal Sistem Informasi*, vol. 11, no. 2, pp. 2540–9719, May 2022, [Online]. Available: <http://sistemasi.ftik.unisi.ac.id>
- [14] C. Tacharri, A. Rohmani, and A. Fahmi, "Strategic Clustering of Poverty Areas in Central Java Using K-Means and Silhouette Evaluation," *Sinkron*, vol. 9, no. 2, pp. 895–904, May 2025, doi: 10.33395/sinkron.v9i2.14734.
- [15] M. F. Akbar and L. Zahrotun, "K-Means Centroid Optimization with Genetic Algorithm for Clustering Micro, Small, Medium Enterprises in Yogyakarta," *JUITA: Jurnal Informatika*, pp. 87–97, Aug. 2025, doi: 10.30595/juita.v13i2.25480.
- [16] A. N. B. Prasetyo, M. Maimunah, and P. Sukmasetya, "K-Means Clustering Method for Determining Waste Transportation Routes to Landfill," *Jurnal Riset Informatika*, vol. 5, no. 3, pp. 277–284, Jun. 2023, doi: 10.34288/jri.v5i3.219.
- [17] W. R. Nurputra, W. Witanti, and A. Komarudin, "Principal Component Analysis (PCA) Untuk Meningkatkan Hasil Klasterisasi Penjualan Video Game Menggunakan Algoritma K-Means," *Jurnal Locus Penelitian dan Pengabdian*, vol. 4, no. 8, Aug. 2025, doi: 10.58344/locus.v4i8.4151.
- [18] D. Romantika Marpaung, E. Gunawan, F. Rio Fa, and A. Christianto, "Klasterisasi Negara Dunia Berdasarkan Data Sosioekonomi dan Demografi Tahun 2023 dengan PCA dan K-Means," vol. 14, no. 1, pp. 2715–7849, 2025, doi: 10.34010/komputa.v14i1.
- [19] R. Hidayati, A. Zubair, A. Hidayat Pratama, and L. Indana, "Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering Silhouette Coefficient Analysis in 6 Measuring Distances of K-Means Clustering," *Techno.COM: Jurnal Teknologi Informasi dan Komunikasi*, vol. 20, no. 2, pp. 186–197, 2021.
- [20] P. Vania and B. Nurina Sari, "Perbandingan Metode Elbow dan Silhouette untuk Penentuan Jumlah Klaster yang Optimal pada Clustering Produksi Padi menggunakan Algoritma K-Means," *Jurnal Ilmiah Wahana Pendidikan*, vol. 9, no. 21, pp. 547–558, 2023, doi: 10.5281/zenodo.10081332.
- [21] M. Dwi Ananda, K. Nurmelita Malik, and A. Fitri Nur Masruriyah, "Studi Komparatif Algoritma K-Means dan K-Medoids untuk Segmentasi Informasi Kesehatan," 2025. [Online]. Available: <http://jurnal.bsi.ac.id/index.php/co-science>
- [22] Faris Helmi, "Mengenal t-SNE: Teknik Visualisasi untuk Menggali Pola Tersembunyi dalam Data," Purwadhika. Accessed: Dec. 01, 2025. [Online]. Available: <https://purwadhika.com/blog/mengenal-t-sne-teknik-visualisasi-untuk-menggali-pola-tersembunyi-dalam-data>
- [23] J. Ye, "Analysis on E-commerce Order Cancellations Using Market Segmentation Approach," in *2021 The 4th International Conference on Computers in Management and Business*, New York, NY, USA: ACM, Jan. 2021, pp. 33–40. doi: 10.1145/3450588.3450596.