PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 03 (2026) Page: 1015-1025
E-ISSN: 3109-8886

Analisis Segmentasi Produk Menggunakan Algoritma K-Means Clustering pada Dataset Tokopedia Product Reviews 2025

[1]Muhammad Balad Al-Amin, [2]Adit Pradika Yoga Putra, [3]Bintang Maldini

[123]Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia

[1]balad.work@gmail.com, [2]aditpradika05@gmail.com, [3]bintangmaldini1358@gmail.com

Abstract

The rapid growth of e-commerce in Indonesia, particularly on the Tokopedia platform, has generated a large volume of customer review data that can be utilized to support business decision-making. This study aims to develop product segmentation based on customer review characteristics using data mining techniques to support Business Intelligence in e-commerce marketplaces. The dataset used is the Tokopedia Product Reviews 2025 dataset from Kaggle, consisting of 5,521 unique products aggregated from the original review data. The study follows the CRISP-DM methodology, including Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Feature engineering was performed to generate analytical attributes, and the K-Means clustering algorithm was applied with the optimal number of clusters (k = 3), determined using the Elbow Method and Silhouette Score. The clustering results identified three product segments: High Quality (2,749 products), High Demand (one outlier product with exceptionally high sales), and High Volume (2,771 products). The resulting dataset was implemented as a Business Intelligence-ready dataset to support product performance monitoring and data-driven marketing strategy development.

Keywords: *K-Means Clustering, Data Mining, Product Segmentation, Business Intelligence, Tokopedia Product Reviews.*

Abstrak

Perkembangan e-commerce di Indonesia, khususnya pada platform Tokopedia, menghasilkan volume data ulasan pelanggan yang besar dan berpotensi dimanfaatkan sebagai dasar pengambilan keputusan bisnis. Penelitian ini bertujuan membangun segmentasi produk berdasarkan karakteristik ulasan pelanggan menggunakan teknik data mining untuk mendukung sistem Business Intelligence pada marketplace. Dataset yang digunakan adalah Tokopedia Product Reviews 2025 dari Kaggle yang terdiri atas 5.521 produk unik hasil agregasi data ulasan. Penelitian menerapkan metodologi CRISP-DM yang meliputi Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Feature engineering dilakukan untuk menghasilkan atribut analitis, kemudian algoritma K-Means Clustering diterapkan dengan jumlah klaster optimal sebanyak tiga berdasarkan Elbow Method dan Silhouette Score. Hasil clustering menghasilkan tiga segmen produk, yaitu High Quality (2.749 produk), High Demand (1 produk outlier dengan penjualan sangat tinggi), dan High Volume (2.771 produk). Hasil penelitian diimplementasikan dalam bentuk dataset siap pakai yang dapat dimanfaatkan pada dashboard Business Intelligence untuk mendukung pemantauan performa produk dan penyusunan strategi pemasaran berbasis data.

Kata Kunci: K-Means Clustering, Data Mining, Segmentasi Produk, Business Intelligence, Tokopedia Product Reviews.

A. PENDAHULUAN

Era digital telah mendorong pertumbuhan industri e-commerce secara signifikan di seluruh dunia, termasuk di Indonesia. Platform marketplace seperti Tokopedia menjadi salah satu ekosistem perdagangan digital terbesar yang menampung jutaan transaksi setiap harinya. Setiap transaksi disertai oleh data ulasan pelanggan yang memuat informasi berharga mengenai kepuasan konsumen, kualitas produk, sentimen pembelian, serta dinamika permintaan pasar.

Meskipun data ulasan tersebut tersedia dalam jumlah yang sangat besar, sebagian besar pelaku bisnis belum mampu mengolahnya secara sistematis untuk keperluan pengambilan keputusan strategis. Tantangan utama yang dihadapi meliputi volume data yang masif, keberagaman karakteristik produk, serta belum adanya mekanisme segmentasi yang mampu mengelompokkan produk berdasarkan pola kinerja dan permintaan. Kondisi tersebut mendorong perlunya pendekatan berbasis data mining yang mampu mengekstrak pola tersembunyi dari data ulasan pelanggan.

Teknik clustering merupakan salah satu pendekatan dalam unsupervised machine learning yang mampu mengelompokkan objek-objek dengan karakteristik serupa ke dalam segmen yang sama tanpa memerlukan label data

Diterima Redaksi: 10-06-2026 | Selesai Revisi: 27-06-2026 | Diterbitkan Online: 28-06-2026

sebelumnya. Salah satu algoritma yang banyak digunakan adalah K-Means Clustering karena mampu menghasilkan pengelompokan data secara efektif dan telah diterapkan pada berbagai penelitian segmentasi pelanggan maupun produk di bidang e-commerce. Berdasarkan permasalahan tersebut, penelitian ini bertujuan menerapkan algoritma K-Means Clustering pada data ulasan produk Tokopedia untuk menghasilkan segmentasi produk yang bermakna melalui pendekatan CRISP-DM.

Penelitian ini diawali dengan proses pembersihan dan persiapan data, kemudian dilanjutkan dengan feature engineering untuk membangun atribut yang merepresentasikan performa produk. Selanjutnya, algoritma K-Means Clustering diterapkan untuk menghasilkan segmentasi produk yang dapat diinterpretasikan dari sudut pandang bisnis. Hasil segmentasi tersebut kemudian diimplementasikan menjadi dataset siap pakai yang dapat dimanfaatkan dalam pengembangan dashboard Business Intelligence.

Melalui penelitian ini diharapkan dapat diperoleh segmentasi produk yang mampu mendukung pengambilan keputusan bisnis, khususnya dalam penyusunan strategi penetapan harga, promosi, dan pengelolaan inventaris. Selain itu, hasil penelitian juga diharapkan dapat menjadi referensi bagi pengembang sistem dalam membangun dashboard Business Intelligence berbasis kluster serta memberikan kontribusi akademis mengenai penerapan CRISP-DM dan K-Means Clustering pada domain e-commerce di Indonesia.

B. METODE

2.1 Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode data mining untuk melakukan segmentasi produk berdasarkan karakteristik ulasan pelanggan pada marketplace Tokopedia. Proses penelitian mengacu pada metodologi Cross Industry Standard Process for Data Mining (CRISP-DM) yang terdiri atas enam tahapan, yaitu Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Metodologi ini dipilih karena menyediakan alur kerja yang sistematis dalam proses pengolahan data hingga implementasi hasil analisis. Penelitian diawali dengan proses pengumpulan dan pemahaman data, dilanjutkan dengan persiapan data, pembentukan fitur (feature engineering), penerapan algoritma K-Means Clustering, evaluasi hasil clustering menggunakan Elbow Method dan Silhouette Score, hingga implementasi hasil dalam bentuk dataset yang digunakan pada dashboard Business Intelligence.

2.2 Data Mining

Data mining adalah proses komputasional untuk menemukan pola, anomali, dan korelasi dalam kumpulan data berukuran besar dengan tujuan mengekstrak pengetahuan yang berguna (Han, Pei, & Tong, 2022). Dalam konteks e-commerce, data mining diaplikasikan untuk memahami perilaku konsumen, mendeteksi kecurangan transaksi, serta membangun sistem rekomendasi produk. Penelitian terkini menunjukkan bahwa integrasi teknik data mining dengan platform marketplace mampu meningkatkan akurasi pengambilan keputusan bisnis secara signifikan (Sujatha et al., 2023).

2.3 Metodologi CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) merupakan kerangka kerja standar industri yang menggambarkan pendekatan umum yang digunakan oleh pakar data mining dalam memecahkan masalah (Schröer, Kruse, & Gómez, 2021). Metodologi ini terdiri atas enam fase yang bersifat iteratif: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Keunggulan CRISP-DM terletak pada strukturnya yang fleksibel dan berorientasi pada tujuan bisnis, sehingga banyak diadopsi dalam penelitian maupun industri.

2.4 K-Means Clustering

K-Means Clustering adalah algoritma unsupervised learning yang mempartisi dataset ke dalam k kluster dengan cara meminimalkan sum of squared distances antara setiap titik data dengan centroid klusternya (Ahmed, Seraj, & Islam, 2020). Algoritma ini bekerja secara iteratif: diawali dengan inisialisasi centroid secara acak, dilanjutkan dengan penugasan setiap titik data ke centroid terdekat, kemudian memperbarui posisi centroid berdasarkan rata-rata anggota kluster, dan proses ini berulang hingga konvergensi tercapai. K-Means banyak digunakan dalam segmentasi pelanggan karena kesederhanaannya dan skalabilitasnya terhadap dataset besar.

2.5 Elbow Method

Elbow Method adalah teknik heuristik untuk menentukan jumlah kluster optimal pada K-Means Clustering dengan cara memplot nilai inertia (within-cluster sum of squares) terhadap berbagai nilai k (Bholowalia & Kumar, 2014). Nilai k optimal dipilih pada titik di mana kurva menunjukkan perubahan kemiringan yang signifikan, membentuk 'siku' pada grafik. Penurunan inertia yang drastis sebelum titik tersebut dan relatif stagnan setelahnya menjadi indikator utama dalam pemilihan k.

2.6 Silhouette Score

Silhouette Score adalah metrik evaluasi kualitas clustering yang mengukur seberapa mirip suatu objek dengan klusternya sendiri dibandingkan dengan kluster lain (Rousseeuw, 1987). Nilai Silhouette Score berkisar antara -1 hingga 1, di mana nilai mendekati 1 mengindikasikan pemisahan kluster yang baik, nilai 0 menunjukkan objek berada di perbatasan dua kluster, dan nilai negatif mengindikasikan kesalahan pengelompokan. Kombinasi Elbow Method dan Silhouette Score memberikan dasar

evaluasi yang lebih komprehensif dalam penentuan k optimal.

2.7 Feature Engineering

Feature engineering adalah proses transformasi data mentah menjadi representasi fitur yang lebih informatif dan relevan untuk model machine learning (Zheng & Casari, 2018). Dalam konteks analisis ulasan produk, feature engineering mencakup agregasi statistik pada level produk, pembentukan indeks komposit seperti performance score dan popularity score, serta normalisasi fitur untuk memastikan skala yang seragam. Kualitas fitur yang dibangun secara langsung mempengaruhi kemampuan algoritma clustering dalam menghasilkan segmentasi yang bermakna.

2.8 Business Intelligence dan Dashboard

Business Intelligence (BI) merujuk pada seperangkat proses, teknologi, dan alat yang digunakan untuk mengubah data mentah menjadi informasi yang dapat ditindaklanjuti untuk pengambilan keputusan bisnis (Negash, 2004). Dashboard BI merupakan representasi visual dari indikator kinerja utama (KPI) yang memungkinkan pemangku kepentingan memantau kondisi bisnis secara real-time. Dalam konteks e-commerce, dashboard berbasis hasil clustering dapat membantu manajer produk dalam memantau pergeseran segmen dan merespons tren pasar secara proaktif.

2.9 Penelitian Terdahulu

Sejumlah penelitian telah mengkaji penerapan K-Means Clustering pada data e-commerce. Sujatha et al. (2023) menerapkan K-Means pada data ulasan Amazon dan berhasil mengidentifikasi tiga segmen pelanggan yang berbeda berdasarkan pola pembelian dan sentimen. Naeem et al. (2022) menggunakan CRISP-DM dalam analisis segmentasi produk fashion online dan menunjukkan bahwa feature engineering berbasis agregasi produk meningkatkan kualitas clustering secara signifikan. Di konteks Indonesia, Kurniawan & Santoso (2023) menerapkan K-Means pada data transaksi Tokopedia dan berhasil menghasilkan empat segmen pelanggan yang digunakan sebagai dasar strategi promosi. Penelitian ini melanjutkan dan mengembangkan temuan-temuan tersebut dengan berfokus pada segmentasi berbasis karakteristik produk, bukan pelanggan, serta mengintegrasikan hasil clustering ke dalam kerangka Business Intelligence.

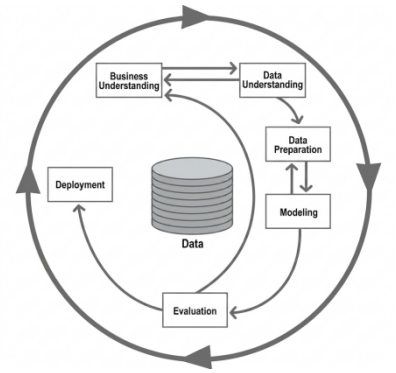
2.10 Sumber dan Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini adalah Tokopedia Product Reviews 2025 yang diperoleh dari platform Kaggle (https://www.kaggle.com/datasets/salmanabdu/tokopedia-product-reviews-2025). Dataset ini berisi rekaman ulasan produk dari platform Tokopedia yang mencakup enam kategori produk utama, yaitu Makanan & Minuman (998 ulasan), Olahraga (974 ulasan), Pertukangan (958 ulasan), Handphone & Tablet (911 ulasan), Kesehatan (908 ulasan), dan Elektronik (772 ulasan).

Atribut utama yang tersedia pada dataset meliputi:

- product_id: identifikasi unik produk
- product_name: nama produk
- product_category: kategori produk
- product_price: harga produk dalam rupiah
- sold_count: jumlah unit yang terjual
- rating: penilaian produk pada skala 1–5
- review_id: identifikasi unik ulasan
- review_text: teks ulasan pelanggan
- review_date: tanggal ulasan
- sentiment_label: label sentimen (negative, neutral, positive)

2.11 Tahapan CRISP-DM



Gambar 1. Diagram Alur Metodologi CRISP-DM

a. Business Understanding

Tahap ini mendefinisikan tujuan bisnis dan rumusan masalah penelitian. Pertanyaan utama yang dijawab adalah: bagaimana segmentasi produk berdasarkan karakteristik ulasan dapat mendukung pengambilan keputusan bisnis di marketplace? Indikator keberhasilan ditetapkan berupa Silhouette Score yang layak (>0,3) dan profil kluster yang dapat diinterpretasikan secara bisnis.

b. Data Understanding

Pada tahap ini dilakukan eksplorasi awal terhadap struktur dataset, statistik deskriptif, distribusi variabel, serta identifikasi potensi masalah kualitas data. Eksplorasi meliputi analisis distribusi rating, distribusi sentimen, distribusi kategori produk, distribusi harga, dan distribusi jumlah penjualan.

c. Data Preparation

Tahap persiapan data mencakup serangkaian proses sebagai berikut:

- Data Cleaning: penghapusan duplikat, pemfilteran review_text kosong, pemfilteran harga tidak valid

(product_price ≤ 0), pemfilteran sold_count negatif, serta pemfilteran rating di luar rentang 1–5.

- Encoding Sentiment: transformasi nilai kategorikal sentiment_label menjadi nilai numerik ordinal (negative=0, neutral=1, positive=2) setelah normalisasi string (lowercase, strip whitespace).
- Product Aggregation: pengelompokan data pada level produk menggunakan group-by product_id dengan fungsi agregasi mean untuk rating, sentiment, product_price, dan sold_count, serta count untuk review_count.
- Feature Engineering: pembentukan empat fitur turunan baru yang dijelaskan pada subbab 3.4.
- Feature Validation: penanganan nilai infinite dan missing value menggunakan nilai median pada kolom numerik.
- Standardization: normalisasi seluruh fitur menggunakan StandardScaler dari scikit-learn untuk memastikan skala yang seragam sebelum proses clustering.

d. Feature Engineering

Empat fitur baru dibangun untuk merepresentasikan performa produk secara komprehensif:

- review_count: jumlah total ulasan per produk sebagai proksi popularitas dan keterlibatan pelanggan.
- performance_score: indeks komposit yang dihitung dengan formula: performance_score = (rating × 0,45) + (review_count × 0,05) + (sold_count × 0,0001). Bobot tertinggi diberikan pada rating karena merupakan indikator kualitas produk yang paling langsung.
- popularity_score: ukuran popularitas absolut produk yang dihitung sebagai: popularity_score = review_count × sold_count. Fitur ini mengukur kombinasi keterlibatan dan volume penjualan.
- review_density: rasio antara jumlah ulasan dan jumlah penjualan yang mencerminkan kecenderungan pembeli untuk meninggalkan ulasan: review_density = review_count / (sold_count + 1). Nilai nol diisikan jika sold_count sama dengan nol untuk menghindari pembagian dengan nol.

e. Modeling: K-Means Clustering

Algoritma K-Means Clustering diterapkan pada data yang telah dinormalisasi menggunakan parameter berikut: n_clusters=3 (ditentukan berdasarkan Elbow Method dan Silhouette Score), random_state=42 (untuk reproduktibilitas), dan n_init=10 (jumlah inisialisasi centroid). Delapan fitur digunakan sebagai input: rating, sentiment, product_price, sold_count, review_count, performance_score, popularity_score, dan review_density.

f. Evaluation

Evaluasi model clustering dilakukan menggunakan dua metode: (1) Elbow Method untuk mengidentifikasi titik infleksi pada kurva inertia terhadap berbagai nilai k (k=2 hingga k=10), dan (2) Silhouette Score untuk mengukur kualitas pemisahan antar kluster secara kuantitatif.

g. Deployment

Hasil clustering disimpan dalam file tokopedia_clustered.csv yang berisi seluruh atribut produk beserta label kluster. Dataset ini dirancang sebagai input untuk dashboard Business Intelligence yang dapat digunakan oleh pemangku kepentingan bisnis dalam memantau dan menganalisis segmen produk.

C. HASIL DAN PEMBAHASAN

3.1 Gambaran Dataset

Dataset Tokopedia Product Reviews 2025 yang digunakan dalam penelitian ini terdiri atas enam kategori produk utama. Setelah melalui proses agregasi pada level produk, dataset akhir (tokopedia_clustered.csv) menghasilkan 5.521 rekaman unik yang masing-masing merepresentasikan satu produk. Distribusi kategori produk menunjukkan sebaran yang relatif seimbang, dengan Makanan & Minuman sebagai kategori dengan jumlah produk terbanyak (998 produk), diikuti Olahraga (974 produk), Pertukangan (958 produk), Handphone & Tablet (911 produk), Kesehatan (908 produk), dan Elektronik (772 produk).

Tabel 1. Distribusi Produk per Kategori

Kategori Produk	Jumlah Produk	Persentase (%)	Kategori Produk
Makanan & Minuman	998	18,07	Makanan & Minuman
Olahraga	974	17,64	Olahraga
Pertukangan	958	17,35	Pertukangan
Handphone & Tablet	911	16,50	Handphone & Tablet
Kesehatan	908	16,45	Kesehatan
Elektronik	772	13,98	Elektronik
Total	**5.521**	**100,00**	**Total**

3.2 Hasil Data Cleaning

Proses data cleaning dilakukan terhadap dataset asli sebelum agregasi. Tahapan yang dilakukan meliputi penghapusan baris duplikat, pemfilteran review_text yang kosong (null), pemfilteran produk dengan harga tidak valid (product_price ≤ 0), pemfilteran sold_count bernilai negatif, serta pemfilteran rating yang berada di luar rentang valid 1–5. Setelah proses ini, dataset bersih siap untuk proses agregasi dan feature engineering lebih lanjut.

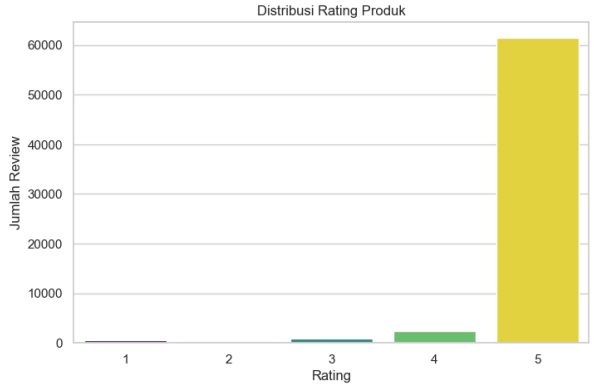
Validasi pasca-cleaning mengkonfirmasi tidak terdapat missing value maupun duplikasi pada kolom-kolom kritis.

Encoding sentimen dilakukan dengan mengubah nilai string pada kolom sentiment_label (setelah normalisasi lowercase dan strip whitespace) menjadi representasi numerik ordinal: negative=0, neutral=1, positive=2.

3.3 Exploratory Data Analysis

a. Distribusi Rating

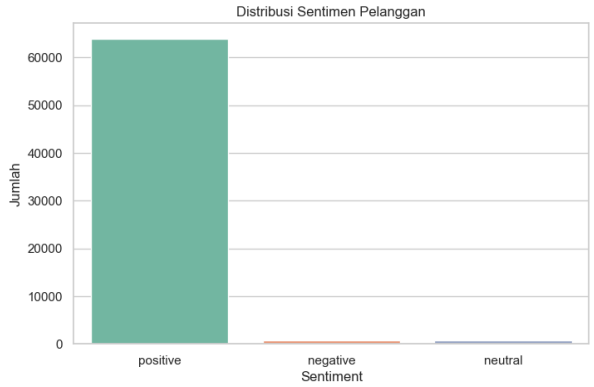
Analisis distribusi rating produk menunjukkan konsentrasi yang sangat tinggi pada nilai rating 5 bintang. Hal ini mencerminkan karakteristik dataset marketplace di mana produk yang memiliki banyak ulasan cenderung merupakan produk-produk berkualitas tinggi yang telah melewati seleksi pasar. Rata-rata rating keseluruhan dataset setelah agregasi adalah 4,92 dengan standar deviasi 0,19, menunjukkan distribusi yang sangat condong ke kanan (positively skewed).



Gambar 2. Distribusi Rating Produk

b. Distribusi Sentimen

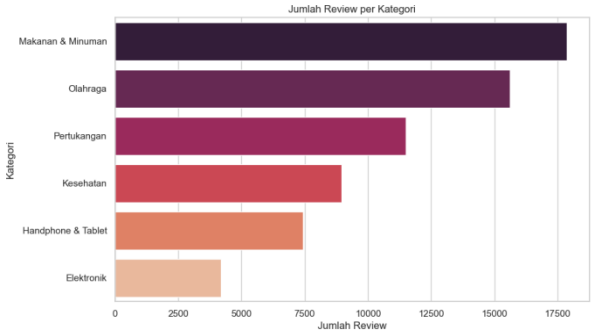
Distribusi sentimen pelanggan didominasi oleh sentimen positif, mencerminkan tingkat kepuasan pelanggan yang tinggi pada produk-produk yang terdapat dalam dataset. Setelah encoding numerik, nilai rata-rata sentimen pada dataset agregat adalah 1,97 (dari skala 0–2), mengkonfirmasi dominasi ulasan bersentimen positif.



Gambar 3. Distribusi Sentimen Pelanggan

c. Distribusi Kategori Produk

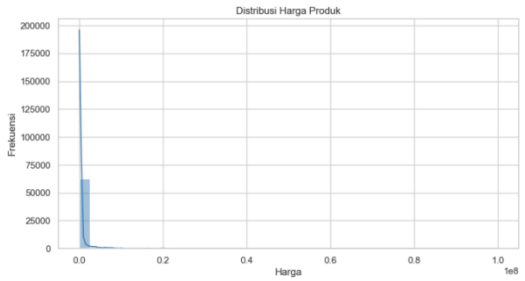
Sebaran produk antar kategori relatif seimbang dengan kisaran 13–18% per kategori. Distribusi ini mengindikasikan bahwa dataset bersifat representatif untuk berbagai segmen pasar Tokopedia.



Gambar 4. Distribusi Review per Kategori Produk

d. Distribusi Harga

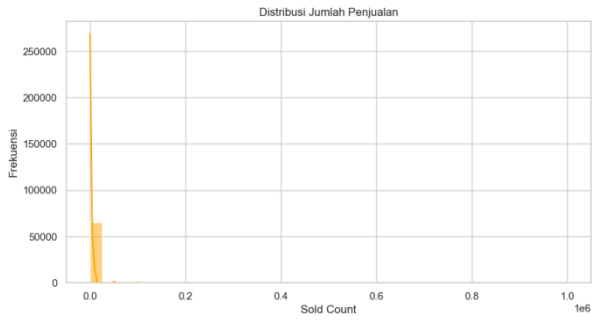
Distribusi harga produk sangat bervariasi dengan sebaran yang sangat lebar. Nilai minimum harga adalah Rp100, sementara nilai maksimum mencapai Rp99.999.000. Nilai rata-rata harga produk adalah sekitar Rp1.034.000, namun nilai median yang jauh lebih rendah (sekitar Rp95.000) mengindikasikan adanya distribusi yang sangat right-skewed akibat keberadaan produk-produk premium dengan harga sangat tinggi.



Gambar 5. Distribusi Harga Produk

e. Distribusi Sold Count

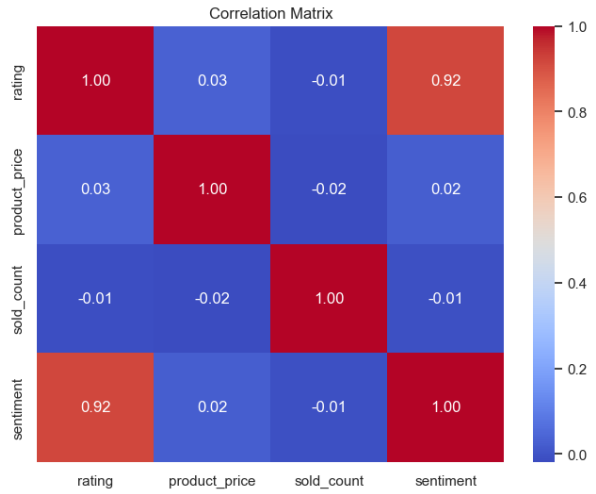
Distribusi jumlah penjualan (sold_count) juga menunjukkan pola right-skewed yang ekstrem. Sebagian besar produk memiliki jumlah penjualan rendah hingga sedang, sementara sebagian kecil produk mencapai jumlah penjualan yang sangat tinggi. Kondisi ini mencerminkan fenomena distribusi power-law yang umum dijumpai pada pasar e-commerce, di mana sejumlah kecil produk mendominasi volume penjualan keseluruhan.



Gambar 6. Distribusi Jumlah Penjualan

f. Heatmap Korelasi

Analisis korelasi antara fitur-fitur utama (rating, product_price, sold_count, sentiment) menunjukkan pola hubungan yang informatif. Rating dan sentimen memiliki korelasi positif sedang, mengkonfirmasi bahwa produk dengan rating tinggi cenderung mendapatkan ulasan bersentimen positif. Sebaliknya, korelasi antara harga dan jumlah penjualan bersifat negatif sedang, menunjukkan bahwa produk dengan harga lebih terjangkau cenderung memiliki volume penjualan yang lebih tinggi.



Gambar 7. Heatmap Korelasi Fitur Utama

3.4 Feature Engineering

Proses feature engineering menghasilkan empat fitur turunan yang dirancang untuk merepresentasikan dimensi performa produk secara lebih komprehensif dari sekadar fitur aslinya.

Tabel 2. Fitur Hasil Feature Engineering

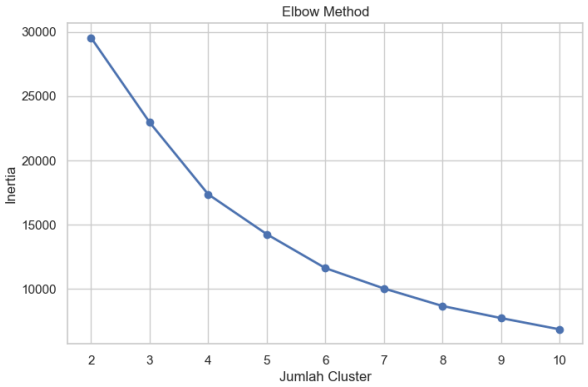
Fitur Baru	Formula	Tujuan
review_count	COUNT(review_id) per product_id	Mengukur keterlibatan pelanggan
performance_score	rating×0.45 + review_count×0.05 + sold_count×0.0001	Indeks kinerja produk komposit
popularity_score	review_count × sold_count	Ukuran popularitas absolut
review_density	review_count / (sold_count + 1)	Rasio ulasan terhadap penjualan

Setelah pembentukan fitur, dilakukan validasi untuk menangani nilai infinite dan missing value. Nilai infinite digantikan dengan NaN terlebih dahulu, kemudian seluruh missing value pada kolom numerik diisi menggunakan nilai median masing-masing kolom. Langkah ini memastikan keandalan dataset sebelum proses standardisasi.

3.5 Hasil Clustering

a. Elbow Method

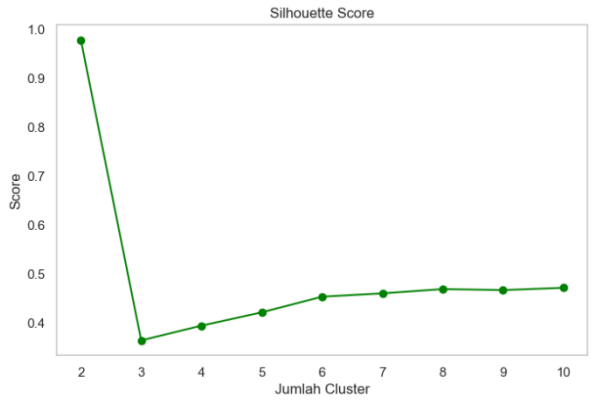
Elbow Method diterapkan dengan mengevaluasi nilai inertia pada rentang k=2 hingga k=10. Grafik inertia menunjukkan penurunan yang tajam dari k=2 ke k=3, kemudian penurunan menjadi lebih landai setelah k=3. Pola 'siku' (elbow) yang terbentuk pada k=3 mengindikasikan bahwa tiga kluster merupakan jumlah partisi yang paling efisien.



Gambar 8. Penentuan Jumlah Kluster Optimal

b. Silhouette Score

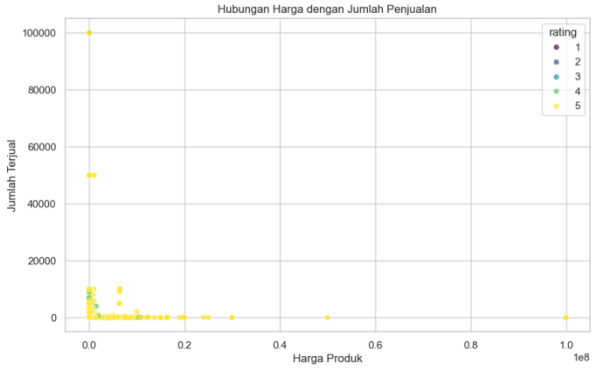
Evaluasi Silhouette Score dilakukan untuk seluruh nilai k pada rentang yang sama. Nilai Silhouette Score dihitung menggunakan metrik Euclidean distance pada data yang telah dinormalisasi. Nilai k=3 dipilih berdasarkan pertimbangan gabungan antara Elbow Method dan nilai Silhouette Score yang menunjukkan kualitas pemisahan kluster yang memadai.



Gambar 9. Grafik Silhouette Score untuk Nilai K

c. Distribusi Kluster

Hasil K-Means Clustering dengan k=3 menghasilkan distribusi kluster sebagai berikut: Kluster 0 memuat 2.749 produk (49,8% dari total), Kluster 1 memuat 1 produk (0,02%), dan Kluster 2 memuat 2.771 produk (50,2%). Distribusi ini menunjukkan bahwa dua kluster utama memiliki ukuran yang hampir seimbang, sementara Kluster 1 merupakan outlier ekstrem yang terpisah karena karakteristiknya yang jauh berbeda dari mayoritas produk.



Gambar 10. Scatter Plot Visualisasi Hasil Clustering (Harga vs. Sold Count)

3.6 Profil Tiap Kluster

Tabel berikut menyajikan profil rata-rata setiap kluster berdasarkan delapan fitur yang digunakan dalam proses clustering.

Tabel 3. Profil Rata-rata Setiap Kluster

Fitur	**Kluster 0 (High Quality)**	**Kluster 1 (High Demand)**	**Kluster 2 (High Volume)**
Jumlah Produk	2.749	1	2.771
Rating (rata-rata)	4,99	4,75	4,85
Sentiment (rata-rata)	2,00	1,90	1,95
Harga Rata-rata (Rp)	1.844.025	87.000	282.254
Sold Count (rata-rata)	178,70	1.000.000	1.948,22
Review Count (rata-rata)	4,98	20,00	18,71
Performance Score	2,51	103,14	3,31
Popularity Score	1.041,39	20.000.000	37.989,46
Review Density	0,09	0,00	0,05

a. Kluster 0 — High Quality

Kluster 0 mengelompokkan 2.749 produk (49,8% dari total) yang dicirikan oleh rating rata-rata tertinggi (4,99 dari skala 5,00) dan sentimen rata-rata hampir sempurna (2,00). Harga rata-rata kluster ini adalah Rp1.844.025, jauh lebih tinggi dibandingkan kluster lain, dengan harga median sekitar Rp125.000 dan harga tertinggi mencapai Rp99.999.000. Jumlah penjualan rata-rata relatif rendah (178,70 unit) dengan review_count rata-rata 4,98 ulasan per produk.

Karakteristik paling menonjol dari kluster ini adalah tingkat kepuasan pelanggan yang luar biasa tinggi tercermin dari nilai rating dan sentimen, meskipun volume penjualan dan jumlah ulasan per produk relatif rendah. Kluster ini didominasi oleh produk-produk dengan harga yang sangat bervariasi namun memiliki persepsi kualitas yang sangat baik di mata pelanggan. Nilai review_density yang lebih tinggi (0,09) dibandingkan kluster lain mengindikasikan keterlibatan aktif pelanggan dalam memberikan ulasan meskipun volume penjualan tidak terlalu besar.

b. Kluster 1 — High Demand

Kluster 1 adalah kluster outlier yang hanya terdiri atas satu produk. Produk ini memiliki karakteristik yang jauh melampaui produk-produk lain, dengan jumlah penjualan mencapai 1.000.000 unit — menjadikannya produk dengan permintaan terbesar dalam dataset. Harga produk ini adalah Rp87.000, berada pada segmen terjangkau. Rating sebesar 4,75 dan sentimen 1,90 masih menunjukkan kualitas yang baik, meskipun sedikit di bawah rata-rata kluster lain. Performance score sebesar 103,14 jauh melampaui kluster lain (2,51 dan 3,31) karena bobot sold_count dalam formulanya. Popularity score sebesar 20.000.000 — 542 kali lebih besar dari Kluster 2 — mengkonfirmasi posisi produk ini sebagai outlier ekstrem dalam hal popularitas. Pemisahan produk ini ke dalam kluster tersendiri oleh algoritma K-Means merupakan hasil yang valid dan dapat diinterpretasikan dengan baik.

c. Kluster 2 — High Volume

Kluster 2 mengelompokkan 2.771 produk (50,2% dari total) yang dicirikan oleh volume penjualan dan jumlah ulasan yang jauh lebih tinggi dibandingkan Kluster 0. Rata-rata sold_count kluster ini adalah 1.948,22 unit, hampir 11 kali lebih tinggi dari Kluster 0. Review_count rata-rata sebesar 18,71 ulasan per produk juga jauh melebihi Kluster 0 (4,98 ulasan). Rating rata-rata 4,85 dan sentimen 1,95 menunjukkan tingkat kepuasan pelanggan yang tinggi, meskipun sedikit lebih rendah dibandingkan Kluster 0. Harga rata-rata sebesar Rp282.254 berada di bawah Kluster 0 namun masih jauh di atas harga produk Kluster 1. Kluster ini merepresentasikan produk-produk mainstream yang telah berhasil membangun basis pelanggan loyal dengan volume transaksi yang konsisten tinggi.

3.7 Interpretasi Bisnis

Ketiga segmen produk yang dihasilkan memberikan implikasi bisnis yang berbeda dan dapat ditindaklanjuti secara langsung oleh pemangku kepentingan marketplace.

Produk-produk pada Kluster 0 (High Quality) merupakan kandidat utama untuk program promosi berbasis kualitas, seperti label 'Top Rated' atau 'Best Quality'. Meskipun volume penjualannya relatif rendah, tingkat kepuasan pelanggan yang sangat tinggi menjadikannya aset penting untuk membangun reputasi platform. Strategi yang direkomendasikan meliputi peningkatan visibilitas melalui featured placement dan optimisasi harga untuk mendorong konversi.

Produk pada Kluster 1 (High Demand) merupakan produk strategis yang membutuhkan perhatian khusus dalam manajemen stok dan rantai pasok. Dengan volume penjualan ekstrem, ketersediaan stok yang memadai dan kelancaran logistik menjadi prioritas utama. Produk ini juga merupakan kandidat ideal untuk program bundling dan cross-selling.

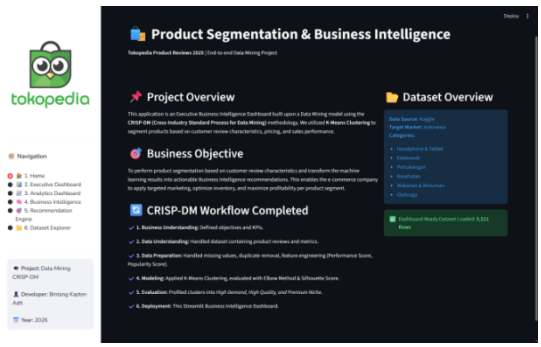
Produk-produk pada Kluster 2 (High Volume) merupakan tulang punggung penjualan platform dengan kombinasi volume penjualan tinggi dan kepuasan pelanggan yang baik. Strategi yang tepat untuk segmen ini adalah pemeliharaan loyalitas pelanggan melalui program review incentive, optimisasi biaya logistik melalui efisiensi skala, dan pengembangan program seller partnership untuk menjaga kualitas.

3.8 Implementasi Dashboard Business Intelligence

Dataset hasil clustering (tokopedia_clustered.csv) dirancang sebagai sumber data utama untuk dashboard Business Intelligence yang memungkinkan pemantauan real-time dan analisis segmen produk secara interaktif. Dashboard direkomendasikan untuk memuat tiga modul utama.

a. Halaman Overview Dashboard

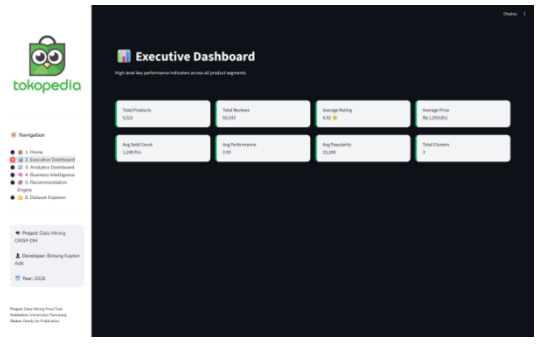
Halaman Overview Dashboard berfungsi sebagai tampilan awal yang menyajikan ringkasan hasil analisis data secara menyeluruh. Dashboard ini menampilkan informasi mengenai distribusi produk pada setiap kluster, gambaran jumlah produk berdasarkan segmen, serta beberapa indikator utama seperti rata-rata rating, harga produk, dan jumlah penjualan (sold_count). Informasi tersebut memberikan gambaran umum mengenai kondisi dataset sehingga pengguna dapat memahami hasil segmentasi dengan cepat sebelum melakukan analisis yang lebih mendala.



Gambar 11. Tampilan Beranda Dashboard Utama

b. Halaman Product Analytics Dashboard

Halaman Product Analytics Dashboard digunakan untuk mengeksplorasi hasil segmentasi secara lebih rinci melalui berbagai visualisasi interaktif. Dashboard ini menyediakan analisis berdasarkan kluster maupun kategori produk, serta menampilkan hubungan antara harga dan jumlah penjualan dalam bentuk scatter plot. Selain itu, pengguna dapat membandingkan nilai performance score dan popularity score pada setiap segmen sehingga karakteristik masing-masing kluster dapat dianalisis secara lebih komprehensif.

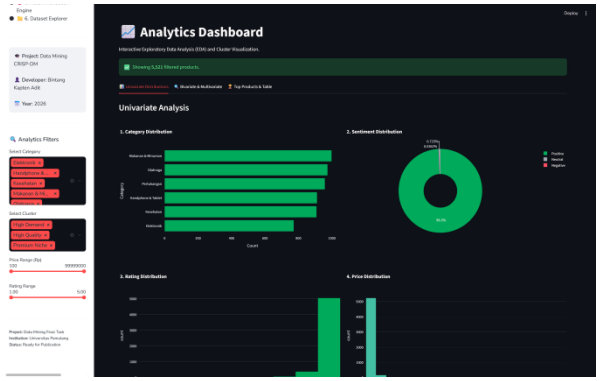


Gambar 12. Tampilan Executive Dashboard

c. Halaman Business Recommendation Dashboard

Halaman Business Recommendation Dashboard menyajikan hasil analisis dalam bentuk rekomendasi yang dapat mendukung pengambilan keputusan bisnis.

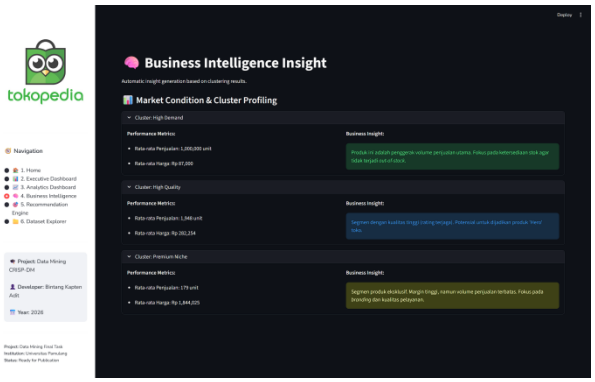
Dashboard ini menampilkan prioritas produk berdasarkan hasil clustering, rekomendasi strategi untuk setiap segmen, serta pemantauan perubahan kluster dari waktu ke waktu. Dengan adanya informasi tersebut, pengguna dapat menentukan tindakan yang sesuai terhadap masing-masing kelompok produk sehingga strategi bisnis yang diterapkan menjadi lebih efektif dan berbasis data.



Gambar 13. Tampilan Modul Analitik Dashboard

d. Halaman Business Intelligence Insight

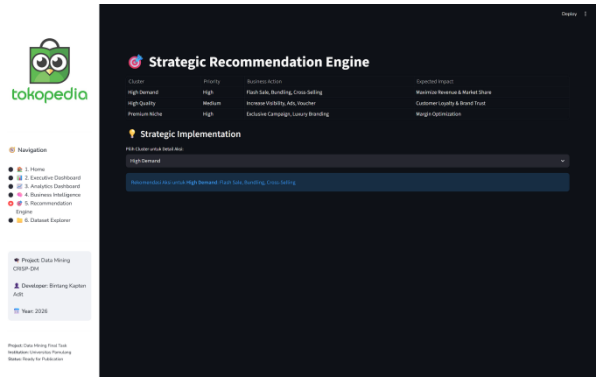
Halaman Business Intelligence Insight menyajikan hasil interpretasi dari proses clustering dalam bentuk informasi yang mudah dipahami oleh pengguna. Setiap kluster dijelaskan berdasarkan karakteristik utamanya, seperti performa produk, tingkat penjualan, dan kisaran harga, kemudian dilengkapi dengan business insight yang dapat dijadikan dasar pengambilan keputusan. Melalui tampilan ini, pengguna tidak hanya mengetahui hasil segmentasi, tetapi juga memahami kondisi setiap kelompok produk beserta peluang strategi yang dapat diterapkan untuk meningkatkan kinerja bisnis.



Gambar 14. Tampilan Business Intelligence Dashboard

e. Halaman Strategic Recommendation Engine

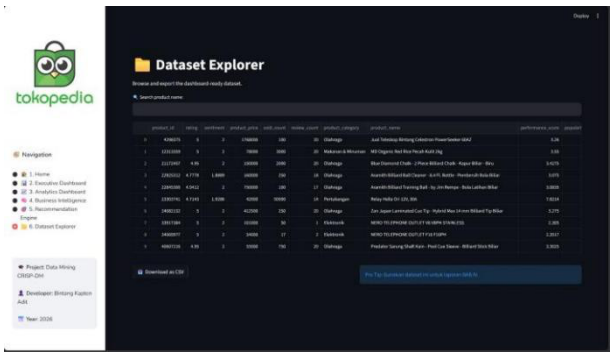
Dashboard Strategic Recommendation Engine menampilkan rekomendasi strategi bisnis berdasarkan hasil segmentasi produk yang telah diperoleh. Setiap kluster disertai tingkat prioritas, usulan tindakan yang dapat dilakukan, serta dampak yang diharapkan terhadap performa bisnis. Penyajian informasi tersebut membantu pihak manajemen menentukan langkah yang paling sesuai bagi setiap segmen produk, sehingga keputusan yang diambil menjadi lebih terarah, efektif, dan didukung oleh hasil analisis data.



Gambar 15. Tampilan Modul Rekomendasi Dashboard

f. Halaman Dataset Explorer

halaman Dataset Explorer yang berfungsi sebagai media untuk melihat dan mengelola dataset hasil proses segmentasi produk. Halaman ini menyediakan fitur pencarian berdasarkan nama produk sehingga pengguna dapat menemukan data tertentu dengan lebih cepat. Selain itu, seluruh atribut hasil pengolahan, seperti rating, sentiment, product_price, sold_count, review_count, performance_score, dan atribut lainnya, ditampilkan dalam bentuk tabel yang mudah dibaca. Dashboard juga menyediakan fitur Download as CSV sehingga dataset dapat diekspor dan dimanfaatkan untuk analisis lanjutan maupun penyusunan laporan. Dengan adanya fitur ini, hasil pengolahan data menjadi lebih mudah diakses, dikelola, dan digunakan sebagai pendukung pengambilan keputusan berbasis data.



Gambar 16. Tampilan Dataset Explorer

D. PENUTUP

4.1 Kesimpulan

Penelitian ini berhasil menerapkan metodologi CRISP-DM secara komprehensif untuk membangun sistem segmentasi produk berbasis K-Means Clustering pada dataset

Tokopedia Product Reviews 2025. Beberapa temuan utama yang dapat disimpulkan adalah sebagai berikut.

Pertama, proses feature engineering yang menghasilkan empat fitur turunan (review_count, performance_score, popularity_score, dan review_density) terbukti krusial dalam memperkaya representasi karakteristik produk melampaui fitur-fitur asli dataset. Fitur-fitur ini memungkinkan pemisahan kluster yang lebih bermakna secara bisnis.

Kedua, algoritma K-Means Clustering dengan k=3 yang ditetapkan berdasarkan Elbow Method dan Silhouette Score menghasilkan tiga segmen produk yang dapat diinterpretasikan dengan jelas: Kluster 0 (High Quality) dengan 2.749 produk berrating sangat tinggi, Kluster 1 (High Demand) yang merupakan outlier dengan satu produk berpenjualan 1.000.000 unit, dan Kluster 2 (High Volume) dengan 2.771 produk yang memiliki volume transaksi dan ulasan tinggi.

Ketiga, hasil segmentasi memberikan basis yang kuat untuk perumusan strategi bisnis yang terdifferensiasi. Setiap kluster memiliki karakteristik yang berbeda dan memerlukan pendekatan bisnis yang berbeda pula dalam hal penetapan harga, promosi, manajemen stok, dan pengembangan program loyalitas.

Keempat, dataset hasil clustering (tokopedia_clustered.csv) yang dihasilkan siap digunakan sebagai sumber data untuk pengembangan dashboard Business Intelligence yang mendukung pengambilan keputusan berbasis data pada platform marketplace.

4.2 Implikasi terhadap Business Intelligence dan Sistem Informasi

Penelitian ini menunjukkan bahwa integrasi teknik data mining dengan kerangka Business Intelligence dapat menciptakan nilai tambah yang signifikan bagi platform e-commerce. Segmentasi produk berbasis clustering tidak hanya menyederhanakan kompleksitas katalog produk yang masif, tetapi juga menyediakan dimensi analitik baru yang tidak tersedia pada sistem pelaporan konvensional. Dalam perspektif sistem informasi, arsitektur pipeline yang dibangun—dari data mentah, pembersihan, feature engineering, clustering, hingga dataset siap pakai—merupakan prototipe yang dapat diadopsi dalam pengembangan sistem data warehouse dan ETL (Extract, Transform, Load) untuk platform marketplace.

4.3 Saran

Beberapa saran untuk penelitian selanjutnya:

- Penggunaan algoritma clustering alternatif seperti DBSCAN atau Gaussian Mixture Models untuk membandingkan kualitas segmentasi, terutama dalam menangani outlier seperti Kluster 1.
- Integrasi analisis teks (Natural Language Processing) pada review_text untuk memperkaya fitur sentimen melampaui label sentimen biner atau terner.
- Penerapan analisis temporal untuk memantau perubahan kluster produk dari waktu ke waktu sebagai basis sistem early warning.
- Pengembangan dashboard Business Intelligence secara aktual menggunakan platform seperti Tableau, Power BI, atau Looker Studio dengan koneksi langsung ke dataset clustered.
- Perluasan dataset dengan menambahkan data dari platform marketplace lain untuk analisis komparatif lintas platform.

E. DAFTAR PUSTAKA

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), 1295. https://doi.org/10.3390/electronics9081295

Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, 105(9), 17–24.

Han, J., Pei, J., & Tong, H. (2022). Data mining: Concepts and techniques (4th ed.). Morgan Kaufmann.

Kurniawan, R., & Santoso, H. B. (2023). Customer segmentation on Indonesian e-commerce using K-Means clustering: A case study of Tokopedia. Journal of Information Systems Engineering and Business Intelligence, 9(1), 45–56. https://doi.org/10.20473/jisebi.9.1.45-56

Naeem, M., Ozuem, W., Howell, K., & Ranfagni, S. (2022). A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. International Journal of Qualitative Methods, 21. https://doi.org/10.1177/16094069221111440

Negash, S. (2004). Business intelligence. Communications of the Association for Information Systems, 13(1), 177–195. https://doi.org/10.17705/1CAIS.01315

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Sujatha, R., Aravinth, J., Chatterjee, J. M., & Morales-Menendez, R. (2023). Customer behaviour analysis based on big data analytics on e-commerce platform. Intelligent

Systems with Applications, 17, 200182. https://doi.org/10.1016/j.iswa.2023.200182

Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: Principles and techniques for data scientists. O'Reilly Media