Analisis Klasterisasi Penjualan Global Berdasarkan Kategori Produk dan Negara Tujuan Menggunakan Metode K-Means

Rafael Theo Santoso[1*], Aji Kia Ramadhani[2], Dyah Purwaningsih[3], Arcco Putra Azhariansyah[4]

[1]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa
[1*]230103074@mhs.udb.ac.id

[2]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa
[2]230103046@mhs.udb.ac.id

[3]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa
[3]230103190@mhs.udb.ac.id

[4]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa
[4]230103051@mhs.udb.ac.id

Abstrak— **Penelitian ini bertujuan untuk melakukan analisis klasterisasi terhadap data penjualan global guna memetakan karakteristik wilayah negara tujuan dan preferensi kategori produk secara bersamaan. Metodologi yang digunakan Adalah *Knowledge Discovery in Database* (KDD) yang mengimplementasikan algoritma *K-Means Clustering*. Dataset yang dinalisis mencakup 2.000 transaksi ritel dari 10 negara tujuan dan 8 kategori produk. Pada tahap pra=pemrosesan, penggunaan *StandardScaler* dan *Pricipal Component Analysis* (PCA) berhasil mereduksi dimensi dengan menangkap 55,6% varians data sekaligus mengingkatkan kualitas klasterisasi secara substansial. Berdasarkan pengujian metode *Elbow,* didapatkan jumlah klister optimal sebanyak tiga (K=3). Evaluasi hasil klasterisasi menunjukkan pemisahan yang sangat baik dengan *Silhouette Score* sebesar 0,5412 dan *Davies-Bouldin Index* 0,5309. Hasil analisis mengidentifikasi tiga profil pasar utama: Klaster 1 (*Pasar Mixed & Balanced)* yang responsive terhadap diversifikasi produk, Klaster 2 (*Pasar Beauty & Lifestyle)* dengan dominasi produk kecantikan dan olahraga, serta Klaster 3 (*Pasar Books & Education)* yang berorientasi pada buku dan kebutuhan harian. Selain itu, penelitian ini juga menghasilkan aplikasi web interaktif berbasis Streamlit Bernama GlobeCluster. Secara praktis, hasil riset ini memberikan rekomendasi strategis berbasis data bagi pemangku kebijakan bisnis internasional untuk mengoptimalkan alokasi inventaris logistik.**
Kata kunci— Data Mining, K-Means Clustering, PCA, Penjualan Global, Segmentasi Pasar

Abstract— **This study aims to conduct a cluster analysis on global sales data to simultaneously map the characteristics of destination countries and product category preferences. The methodology employs the Knowledge Discovery in Database (KDD) framework, implementing the K-Means Clustering algorithm. The analyzed dataset comprises 2,000 retail transactions across 10 destination countries and 8 product categories. In the pre-processing stage, utilizing StandardScaler and Principal Component Analysis (PCA) successfully reduced dimensions by capturing 55.5% of the data variance, substantially improving clustering quality. Based on the Elbow method evaluation, the optimal number of clusters is three (K=3). The clustering evaluation showed excellent separation with a Silhouette Score of 0.5412 and a Davies-Bouldin Index of 0.5309. The analysis identified three main market profiles: Cluster 1 (Mixed & Balanced Market), responsive to product diversification; Cluster 2 (Beauty & Lifestyle Market), dominated by beauty and sports products; and Cluster 3 (Books & Education Market), oriented towards books and daily groceries. Additionally, this research produced an interactive Streamlit-based web application named GlobeCluster. Practically, these findings provide data-driven strategic recommendations for international business stakeholders to optimize logistics inventory allocation.**
Keywords— Data Mining, K-Means Clustering, PCA, Global Sales, Market Segmentation.

I. PENDAHULUAN

Perkembangan teknologi informasi dan internet global telah mengubah lanskap perdagangan internasional dengan pesatnya pertumbuhan industri e-commerce[1]. Lonjakan volume transaksi digital ini menghasilkan banyak data penjualan ritel yang menyimpan potensi informasi berharga tentang dinamika pasar dunia[2]. Namun, banyaknya data transaksi ini sering kali menciptakan fenomena data melimpah, informasi kurang karena kegagalan dalam mengoptimalkan basis data akibat analisis yang terbatas[3].

Dalam konteks pasar global, kompleksitas perilaku komsumen terlihat jelas dalam dataset

operasional yang mencatat riwayat 2000 transaksi ritel lintas batas negara secara global[4]. Ekosistem transaksional ini mencakup segmentasi geografisnke berbagai negara tujuan utama seperti Australia, Brasil, Kanada, Jerman, Prancis, India, Jepang, Meksiko, dan Amerika Serikat. Selain itu, pangkalan data ini mencakup berbagai jenis komoditas yang meliputi kategori produk Kecantikan, Buku, Pakaian, Elektronik, Barang Kebutuhan, Perabotan Rumah & Dapur, Olahraga, dan Mainan.

Untuk mengatasi hambatan dalam mengekstrak informasi dari tumpukan data yang sangat besar dan bervariasi, penerapan metodologi data mining yang berbasis kecerdasan buatan menjadi solusi yang penting[5]. Penelitian ini untuk memahami karakteristik pasar secara menyeluruh dapa dilakukan dengan baik melalui pendekatan segmentasi konsumen menggunakan model pembelajaran mesin (*machine learning*)[6]. Salah satu algoritma pembelajaran tanpa pengawasan (*unsupervised learning*) yang terbukti paling andal dan efisien dalam mengidentifikasi pola perilaku pembelian adalah algoritma K-Means clustering[7].

Dalam praktiknya, proses ekstraksi pengetahuan ini diterapkan dalam kerangka kerja yang terorganisir dan sistematis seperti metodologi KDD maupun CRIPS-DM. Dengan pendekatan terorganisir ini variabel transaksional yang bersifat multimedia bisa disederhanakan tanpa menghilangkan inti informasi penting yang terdapat di dalamnya[8]. Oleh sebab itu, penelitian ini bertujuan untuk mengimplementasikan metode data science dengan memanfaatkan teknik klasterisasi untuk menguraikan kompleksitas pola penjualan komersial.

Beberapa penelitian terdahulu telah membuktikan keandalan algoritma K-Means dalam menaganalis data transaksi penjualan. Penelitian yang dilakukan oleh [9] menunjukkan bahwa K-Means sangat efektif untuk mengelompokkan data transaksi penjualan guna menyusun strategi pemasaran yang lebih terfokus dan efisien. Keandalan metode ini juga dibuktikan oleh [10] dalam industri kuliner, di mana klasterisasi pola penjualan terbukti memberikan wawasan strategis untuk manajemen stok dan efisiensi operasional. Dalam lingkup perdangan internasional, penelitian [11] Sukses mengintergrasikan metode K-Means untuk mengklasterisasi produk berdasarkan negara tujuan ekspor, yang menghasilkan segmentasi pasar objektf untuk penentuan skala prioritas alokasi inventaris. Berdasarkan kajian literatur tersebut, penelitian ini akan mengadaptasi metode K-Means untuk memecahkan masalah segmentasi pasar bersakal gloa yang memadukan variabel kategori produk dan negara tujuan secara bersamaan.

Tujuan utama dari penelitian ini adalah untuk melakukan analisis klasterisasi yang mendalam terhadap data penjualan global dengan tujuan memetakan karakteristik wilayah negara tujuan dan preferensi komoditas produk secara bersamaan menggunakan metode K-Means. Melalui pengelompokan berdasarkan dataset historis ini, diharapkan bisa terungkap pola hubungan tersembunyi yang menghubungkan volume transaksi wilayah tertentu dengan kategori produk spesifik secara akurat. Secara praktis, hasil dari riset ini ditargetkan dapat memberikan rekomendasi keputusan yang berbasis data kepada pemangku kebijakan bisnis internasoinal dalam rangka mengoptimalkan alokasi inventaris barang digudang kawasan serta meningkatkan efisiensi biaya logistik.

II. METODOLOGI PENELITIAN

Penelitian ini dilaksanakan dengan menerapkan kerangka kerja *Knowledge Discovery in Database* (KDD). KDD merupakan pendekatan metodologis yang terstruktur untuk mengidentifikasi dan mengekstraksi pola yang valid, baru, dan bermanfaat dari sekumpulan data berskala besar [12].Tahapan KDD dalam penelitian ini diadaptasi ke dalam tiga fase utama, yaitu pra-pemrosesan data, data mining, dan evaluasi [12].

1. Pra-Pemrosesan Data

Data mentah transaksi e-commerce sering kali mengandung noise atau nilai anomali yang dapat mengurangi akurasi pemodelan. Oleh karena itu, tahap pra-pemrosesan dilakukan untuk membersihkan data dan menyeleksi atribut yang relevan. Data kemudian ditransformasikan dari format awalnya menjadi matriks numerik kontinu. Transformasi ini bertujuan agar mesin dan algoritma dapat menghitung jarak matematis antar data secara optimal tanpa adanya ambiguitas [13].

2. Data Mining

Data mining merupakan proses inti dalam kerangka kerja KDD yang bertujuan mengekstraksi pola dan pengetahuan dari data yang telah melalui tahap pra-pemrosesan. Pada penelitian ini, data mining diimplementasikan melalui dua tahapan utama, yaitu penentuan jumlah klaster optimal dan pelaksanaan algoritma K-Means Clustering.

a. Penentuan Jumlah Klaster Optimal

Sebelum proses algoritma K-Means dijalankan, jumlah klaster yang paling optimal (K) harus ditentukan terlebih dahulu. Pendekatan yang digunakan dalam penelitian ini adalah Metode Elbow. Metode ini mengevaluasi kualitas klasterisasi dengan menghitung nilai *Sum of Squared Errors* (SSE) pada berbagai pengujian nilai K [14]. Titik optimal yang dipilih sebagai jumlah klaster adalah ketika grafik penurunan nilai SSE mulai membentuk sudut patahan yang melandai secara signifikan, menyerupai bentuk siku. Pengujian dilakukan pada ruang PCA 2D dengan rentang K = 1 hingga 9, yang kemudian dikonfirmasi dengan metrik *Silhouette Score* sebagai validasi tambahan.

b. Algoritma K-Means Clustering

K-Means adalah salah satu algoritma unsupervised learning berjenis partisional yang membagi data ke dalam sejumlah kelompok berdasarkan tingkat kemiripan karakteristiknya [15]. Proses komputasi pada algoritma K-Means dijalankan pada data yang telah direduksi dimensinya (X_pca) melalui langkah-langkah sistematis berikut:

a. Inisialisasi: Penentuan nilai K dan inisialisasi titik pusat klaster (centroid) secara acak pada ruang dimensi PCA (2D).
b. Perhitungan Jarak: Perhitungan jarak kedekatan antara setiap titik data terhadap titik centroid menggunakan formula Euclidean Distance pada ruang PCA:

$$d(x, c) = \sqrt{\sum_{i=1}^{n} (x_i - c_i)^2}$$

c. Alokasi Klaster: Pengalokasian setiap data ke dalam klaster yang memiliki nilai jarak Euclidean paling minimum.
d. Pembaruan Centroid: Pembaruan letak centroid dengan cara menghitung nilai rata-rata (mean) dari seluruh titik data yang tergabung dalam klaster yang sama.
e. Konvergensi: Pengulangan iterasi perhitungan jarak dan pembaruan letak centroid hingga kondisi konvergen tercapai, yang ditandai dengan tidak adanya data yang berpindah kelompok atau total pergeseran centroid mendekati nol.

Untuk interpretasi hasil bisnis, centroid pada ruang PCA diinverse-transform kembali ke ruang fitur asli melalui inverse PCA dan inverse *StandardScaler*, sehingga nilai centroid dapat dibaca dalam satuan USD asli.

3. Evaluasi

Tahap akhir dari proses *Knowledge Discovery in Database* (KDD) adalah melakukan evaluasi dan validasi terhadap pola atau model yang telah terbentuk. Pada penelitian ini, kualitas dari algoritma K-Means diukur menggunakan dua metrik evaluasi standar, yaitu:

a. *Silhouette Score*: Metrik ini digunakan untuk mengukur seberapa baik sebuah objek ditempatkan di dalam klasternya sendiri (kohesi) dibandingkan dengan klaster lain (separasi). Nilai *Silhouette Score* berada pada rentang -1 hingga 1. Nilai yang positif dan mendekati 1 menunjukkan bahwa objek telah terklasterisasi dengan tepat dan memiliki jarak yang jauh dari klaster tetangganya.
b. *Davies-Bouldin Index* (DBI): Metrik ini mengevaluasi kualitas klaster dengan menghitung rasio antara penyebaran data di dalam klaster dengan jarak antarpusat klaster (centroid). Semakin kecil nilai DBI (mendekati angka 0), maka model klasterisasi dianggap semakin baik karena mengindikasikan bahwa klaster-klaster yang terbentuk sangat padat dan terpisah dengan jelas antara satu sama lain.

III. HASIL DAN PEMBAHASAN

O. Pra-Pemrosesan Data

Dataset yang digunakan dalam penelitian ini adalah Global Retail E-Commerce Dataset yang bersumber

dari platform Kaggle https://www.kaggle.com/datasets/smayanj/e-commerce-transactions-dataset. Dataset ini diunggah oleh AJ pada tahun 2025 dan terdiri dari 2.000 transaksi ritel lintas 10 negara (Australia, Brazil, Canada, France, Germany, India, Japan, Mexico, UK, USA) dan 8 kategori produk (Beauty, Books, Clothing, Electronics, Grocery, Home & Kitchen, Sports, Toys). Data mentah diagregasi menjadi matriks 10×8 yang merepresentasikan total penjualan (USD) per negara per kategori. Karena rentang nilai antarkategori berbeda, dilakukan standardisasi menggunakan *StandardScaler* (Z-score) agar distribusi data setara dan memenuhi prasyarat PCA. Reduksi dimensi dengan 2 komponen PCA menangkap 55,5% varians data (PC1: 35,0%, PC2: 20,5%), yang cukup merepresentasikan pola dominan penjualan global.

Tabel 1. Nilai Proyeksi Data pada Ruang PCA (2D)

Negara	PC1	PC2
Australia	2.8665	1.3407
Brazil	-1.3748	1.7484
Canada	2.6978	0.7912
France	0.1724	-1.4992
Germany	-1.6042	0.5964
India	1.2076	-2.38
Japan	-0.8061	1.1332
Mexico	-0.5522	-1.0689
UK	-2.3524	-0.3195
USA	-0.2547	-0.3422

Proyeksi PCA menunjukkan Australia dan Canada memiliki nilai PC1 positif tinggi, sementara Brazil dan Germany memiliki nilai negatif signifikan, mengindikasikan perbedaan pola konsumsi yang kontras antarwilayah.

P. Data Mining

Data mining merupakan proses inti dalam kerangka kerja KDD yang bertujuan mengekstraksi pola dan pengetahuan dari data yang telah melalui tahap pra-pemrosesan. Pada penelitian ini, data mining diimplementasikan melalui tiga tahapan utama, yaitu penentuan jumlah klaster optimal, pelaksanaan algoritma K-Means Clustering, serta perancangan dan implementasi sistem aplikasi GlobeCluster.

a. Penentuan Jumlah Klaster Optimal

Sebelum menjalankan algoritma K-Means, ditentukan terlebih dahulu jumlah klaster optimal (K) menggunakan Metode Elbow. Metode ini menghitung nilai *Sum of Squared Errors* (SSE) untuk berbagai nilai K dan mengidentifikasi titik "siku" di mana penurunan SSE mulai melandai secara signifikan.

Tabel 2. Nilai SSE untuk Berbagai Jumlah Klaster (K)

K	SSE	Penurunan SSE
1	44.39	-
2	22.20	22.1918
3	7.65	14.5523
4	4.93	2.7186
5	2.12	2.8153
6	1.48	0.6410
7	0.82	0.6514
8	0.47	0.3510
9	0.17	0.3083

Berdasarkan Tabel 2, penurunan SSE paling signifikan terjadi dari K=1 ke K=2 (22.1918) dan dari K=2 ke K=3 (14.5523). Setelah K=3, penurunan SSE mulai melandai (2.7186 untuk K=3 ke K=4). Oleh karena itu, K=3 dipilih sebagai jumlah klaster optimal karena menyeimbangkan kompleksitas model dan kualitas klasterisasi. Grafik *Silhouette Score* pada ruang PCA 2D juga menunjukkan nilai tertinggi pada K=3 (0.5412), yang mengkonfirmasi pemilihan K=3 sebagai optimal.

b. Algoritma K-Means Clustering dan Hasil Klasterisasi

Berdasarkan nilai K optimal yang telah ditentukan, algoritma K-Means dijalankan pada data yang telah direduksi dimensinya (X_pca) dengan langkah-langkah sistematis sebagai berikut: (a) inisialisasi tiga titik centroid secara acak pada ruang PCA 2D; (b) perhitungan jarak Euclidean antara setiap titik data terhadap centroid; (c) alokasi data ke klaster dengan jarak minimum; (d) pembaruan posisi centroid dengan nilai rata-rata anggota klaster; dan (e) iterasi berulang hingga kondisi konvergensi tercapai.

Algoritma K-Means pada data X_pca mencapai konvergensi dalam 3 iterasi dengan rincian pada Tabel 3.

Tabel 3. Ringkasan Proses Iterasi K-Means

Iterasi	Pergeseran Centroid	Jumlah C1	Jumlah C2	Jumlah C3	Status

1	2.2003	0	8	2	Berlanjut
2	3.9435	4	4	2	Berlanjut
3	0.00	4	4	2	Konvergensi

Hasil akhir penugasan klaster per negara disajikan pada Tabel 4.

Tabel 4. Hasil Akhir Penugasan Klaster per Negara

Negara	Klaster
Australia	K3
Brazil	K2
Canada	K3
France	K1
Germany	K2
India	K1
Japan	K2
Mexico	K1
UK	K2
USA	K1

Untuk interpretasi hasil bisnis, centroid pada ruang PCA di-inverse-transform kembali ke ruang fitur asli melalui inverse PCA dan inverse *StandardScaler*, sehingga nilai centroid dapat dibaca dalam satuan USD asli. Nilai centroid final yang merepresentasikan karakteristik rata-rata penjualan tiap klaster disajikan pada Tabel 5.

Tabel 5. Nilai Centroid Final (Skala Asli / USD)

	Beauty	Books	Clothing	Electronics	Grocery	Home & Kitchen	Sports	Toys
Cluster 1	1043 3.37	1369 7.89	1091 1.21	13517 .98	1384 4.32	1066 1.12	1313 3.13	1467 6.58
Cluster 2	1454 3.07	1218 6.05	1150 9.64	9349. 52	8563 .75	1130 0.63	1364 6.65	1306 7.96
Cluster 3	1182 7.67	2148 0.12	1484 7.78	13583 .37	1557 2.80	1410 6.44	9053 .47	9263 .48

Tabel 6. Metrik Evaluasi K-Means Clustering

Metrik Evaluasi	Nilai	Interpretasi
Silhouette Score	0,5412	Nilai positif mendekati 1 menunjukkan objek terklasterisasi dengan tepat dan memiliki jarak yang jauh dari klaster tetangganya
Davies-Bouldin Index (DBI)	0,5309	Nilai di bawah 1 mengindikasikan klaster terpisah dengan sangat baik dan memiliki tingkat kepadatan data yang tinggi
Jumlah Klaster (K)	3	Ditentukan optimal berdasarkan Metode Elbow dan validasi Silhouette Score
Jumlah Negara	10	Australia, Brazil, Canada, France, Germany, India, Japan, Mexico, UK, USA

Berdasarkan hasil pada Tabel 6, nilai *Silhouette Score* sebesar 0,5412 menunjukkan bahwa kualitas klaster yang terbentuk sudah baik karena berada di atas 0,5. Hal ini berarti data pada umumnya jauh lebih dekat dengan titik centroid klasternya sendiri jika dibandingkan dengan klaster lainnya. Sementara itu, nilai *Davies-Bouldin Index* sebesar 0,5309 yang rendah (berada di bawah 1) mengindikasikan bahwa klaster-klaster yang terbentuk terpisah dengan sangat baik dan memiliki tingkat kepadatan data yang tinggi.

Hasil evaluasi ini juga menunjukkan peningkatan performa yang sangat signifikan jika dibandingkan dengan metode sebelumnya tanpa penerapan *StandardScaler* dan PCA. Pada metode sebelumnya, nilai *Silhouette Score* hanya mencapai 0,1034 dan *Davies-Bouldin Index* berada di angka 1,4407. Peningkatan metrik validasi yang drastis ini membuktikan bahwa kombinasi penggunaan *StandardScaler* dan *Principal Component Analysis* (PCA) pada tahap pra-pemrosesan data berhasil meningkatkan kualitas klasterisasi secara substansial.

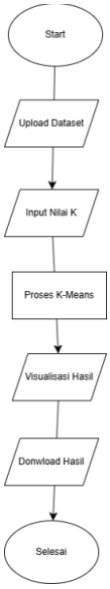
Berdasarkan hasil klasterisasi dan evaluasi metrik tersebut, diidentifikasi tiga profil pasar utama:

1. K1 — Pasar Mixed & Balanced (France, India, Mexico, USA): responsif terhadap diversifikasi produk dengan distribusi penjualan yang relatif merata.
2. K2 — Pasar Beauty & Lifestyle (Brazil, Germany, Japan, UK): dominasi produk kecantikan, olahraga, dan mainan.
3. K3 — Pasar Books & Education (Australia, Canada): berorientasi pada buku dan kebutuhan harian (grocery).

c. Desain dan Implementasi Sistem

Sebelum proses klasterisasi, dirancang desain sistem yang menggambarkan alur kerja dan interaksi pengguna. Desain sistem terdiri dari flowchart yang merepresentasikan alur proses klasterisasi dari awal

hingga akhir, dimulai dari pengunggahan dataset, pemrosesan K-Means, visualisasi hasil, hingga ekspor data. Berdasarkan desain tersebut, dikembangkan aplikasi web interaktif GlobeCluster berbasis Streamlit dengan antarmuka intuitif agar pengguna dapat melakukan analisis K-Means tanpa keahlian pemrograman.

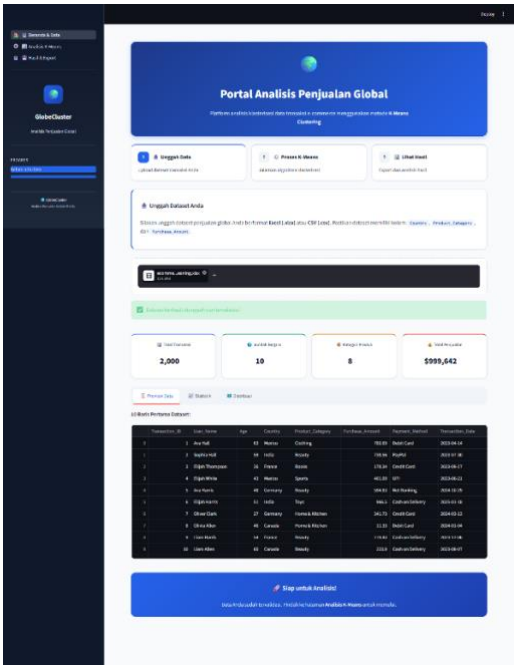


Gambar 1. Flowchart Sistem Klasterisasi

Sistem terdiri dari tiga halaman utama yang diakses melalui sidebar navigasi:

1. Halaman Beranda & Data

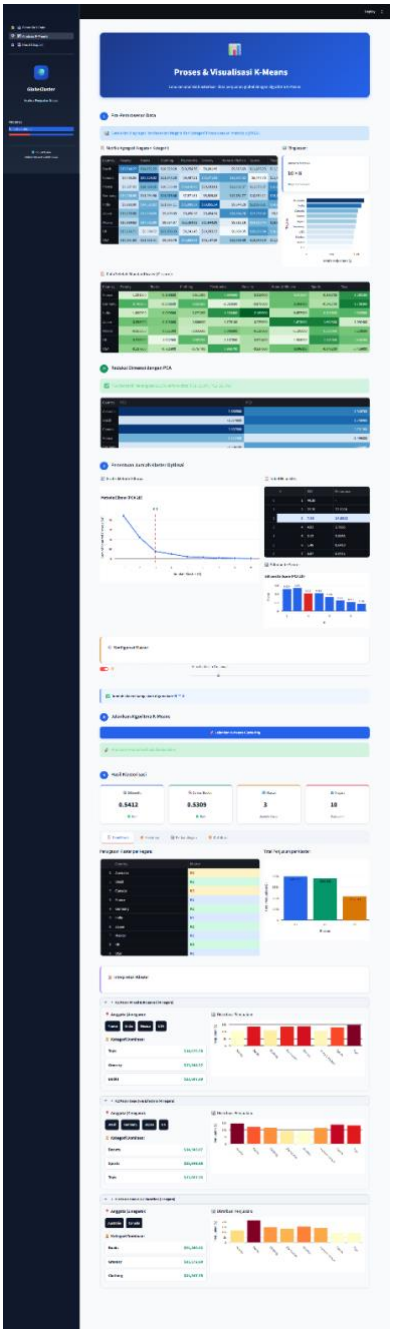
Halaman ini merupakan titik masuk pengguna untuk mengunggah dataset (Excel/CSV) dengan validasi otomatis kolom wajib (Country, Product_Category, Purchase_Amount). Setelah berhasil diunggah, sistem menampilkan ringkasan statistik (Total Transaksi, Jumlah Negara, Kategori Produk, Total Penjualan) dalam format kartu metrik dan tiga tab eksplorasi: Preview Data, Statistik, serta Distribusi.



Gambar 2. Tampilan Halaman Beranda & Data GlobeCluster

2. Halaman Analisis K-Means

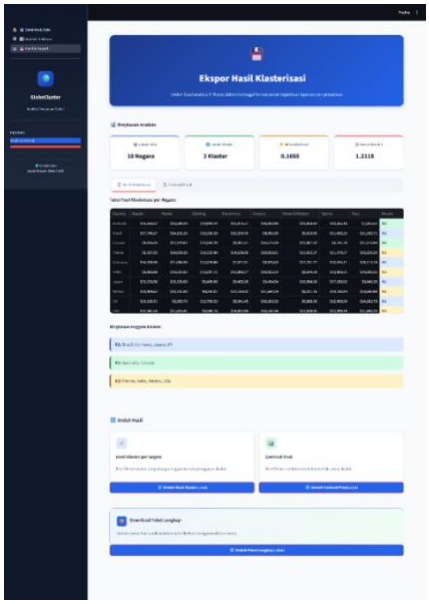
Halaman inti sistem yang mengimplementasikan tahapan KDD secara visual melalui lima bagian: (1) Pra-Pemrosesan Data — matriks agregasi 10×8, standardisasi *StandardScaler* (Z-score), dan heatmap; (2) Reduksi Dimensi PCA — proyeksi data ke ruang 2D dengan informasi varians; (3) Penentuan K Optimal — grafik Elbow dan *Silhouette Score* pada ruang PCA dengan slider interaktif; (4) Jalankan K-Means — eksekusi algoritma pada data X_pca; serta (5) Hasil Klasterisasi — empat kartu metrik (Silhouette 0,5412; DBI 0,5309; K=3; 10 negara), tabel penugasan klaster, dan empat tab visualisasi (Tabel, Heatmap, Perbandingan, Distribusi). Bagian bawah menyajikan interpretasi tiga profil pasar: K1 Mixed & Balanced (France, India, Mexico, USA), K2 Beauty & Lifestyle (Brazil, Germany, Japan, UK), dan K3 Books & Education (Australia, Canada).



Gambar 3. Tampilan Halaman Analisis K-Means GlobeCluster

3. Halaman Hasil & Export

Halaman repositori dokumentasi dengan ringkasan metrik, tabel lengkap penugasan klaster, centroid final, ringkasan anggota berbasis card, serta tiga opsi ekspor: CSV data, CSV centroid, dan Excel multi-sheet.



Gambar 4. Tampilan Halaman Hasil & Export GlobeCluster

1. Evaluasi

Untuk mengukur kualitas klaster yang terbentuk dari proses pemodelan, penelitian ini menggunakan dua metrik evaluasi standar, yaitu *Silhouette Score* dan *Davies-Bouldin Index*. Hasil pengukuran dan interpretasi lengkap telah disajikan pada Tabel 6 di subbab B.2. Sebagai ringkasan, model menghasilkan nilai Silhouette Score sebesar 0,5412 dan *Davies-Bouldin Index* sebesar 0,5309, yang menunjukkan pemisahan klaster yang sangat baik.

Perbandingan dengan metode tanpa pra-pemrosesan (*Silhouette Score* 0,1034; DBI 1,4407) membuktikan bahwa penerapan *StandardScaler* dan PCA berhasil meningkatkan kualitas klasterisasi secara signifikan. Validasi tambahan melalui aplikasi GlobeCluster menunjukkan bahwa hasil klasterisasi konsisten antara komputasi backend dan visualisasi frontend, dengan konvergensi algoritma tercapai dalam 3 iterasi sebagaimana tercatat pada Tabel 3.

IV. KESIMPULAN

Penelitian ini mengimplementasikan K-Means clustering dengan *StandardScaler* dan PCA untuk menganalisis pola penjualan global dari 2.000 transaksi e-commerce lintas 10 negara dan 8 kategori produk. *StandardScaler* memenuhi prasyarat PCA yang mereduksi dimensi menjadi 2 komponen

dengan menangkap 55,5% varians data (PC1: 35,0%, PC2: 20,5%). Metode Elbow pada ruang PCA menghasilkan K=3 optimal, yang dikonfirmasi *Silhouette Score* 0,5412 dan *Davies-Bouldin Index* 0,5309 — peningkatan signifikan dibandingkan metode tanpa PCA (0,1034 dan 1,4407). Hasil analisis mengidentifikasi tiga profil pasar: K1 Pasar Mixed & Balanced (France, India, Mexico, USA); K2 Pasar Beauty & Lifestyle (Brazil, Germany, Japan, UK) dengan dominasi Beauty, Sports, dan Toys; serta K3 Pasar Books & Education (Australia, Canada) yang berorientasi pada kebutuhan edukasional. Penelitian ini juga menghasilkan aplikasi web interaktif GlobeCluster berbasis Streamlit untuk visualisasi dan ekspor hasil analisis.

Secara praktis, K1 direkomendasikan strategi diversifikasi produk, K2 memprioritaskan alokasi inventaris pada kategori Beauty, Sports, dan Toys, serta K3 fokus pada Books dan Grocery dengan bundling produk edukasi. Keterbatasan terletak pada ukuran dataset yang relatif kecil (10 negara), sehingga penelitian selanjutnya dapat memperluas cakupan dataset, menambahkan dimensi temporal, membandingkan algoritma clustering lainnya, dan mengeksplorasi jumlah komponen PCA yang lebih optimal.

REFERENSI

[1] I. Q. A, L. Anggraini, and G. D. Asmara, "Analysis of the Development of E-Commerce Transactions in the 6 Highest Transaction Countries in Southeast Asia," vol. 8, no. 2, 2025, doi: 10.18196/jerss.v8i2.22033.

[2] R. I. Manarung, E. Widodo, and A. M. Rifai, "Sales Data Clustering Using the K-Means Algorithm to Determine Retail Product Needs," vol. 5, no. April, pp. 226–234, 2025.

[3] M. Almaripat, A. Faqih, and A. R. Rinaldy, "Sales Data Classterization Analysis Using K-Means Method for Marketing Strategy Development," vol. 4, no. 2, 2025.

[4] Y. Qu, "Research on Purchasing Behavior Pattern of E-commerce Platform Consumers Based on Big Data Analysis," vol. 0, pp. 187–191, 2025, doi: 10.54254/2754-1169/177/2025.22481.

[5] Y. Putri, D. Aldo, W. Ilham, U. T. Padang, and C. I. Cendekia, "Retail Marketing Strategy Optimization : Customer Segmentation with Artificial Intelligence Integration and K-Means Clustering," vol. 8, no. 4, pp. 2155–2163, 2024.

[6] H. Priscianus Mikael Kia Mado, "Implementasi Algoritma Clustering K-Means untuk," vol. 6, no. 3, pp. 1680–1686, 2025.

[7] H. Safitri, S. Putri, L. Geni, F. Merry, and M. Wati, "Penerapan K-Means Clustering untuk Segmentasi Konsumen E-Commerce Berdasarkan Pola Pembelian," vol. 7, pp. 89–99, 2025.

[8] M. Evelin, "Implementation of Customer Segmentation Model using K-Means and DBSCAN for Fashion Industry Product Transaction," vol. 9, no. November, pp. 2559–2568, 2025.

[9] M. Fajar Maulana 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, 2024, doi: 10.36040/jati.v9i1.12433.

[10] D. Rusvinasari and L. H. Annisa, "Klasterisasi Pola Penjualan Menu Makanan pada Rumah Makan menggunakan Metode K-Means Clustering," *J. Inform. J. Pengemb. IT*, vol. 10, no. 2, pp. 398–409, 2025, doi: 10.30591/jpit.v10i2.8511.

[11] R. A. Nabila *et al.*, "Implementasi Teknik RFM dan K-Means untuk Klasterisasi Produk Berdasarkan Negara Tujuan," vol. 02, no. 01, pp. 14–24, 2026.

[12] R. A. Pramudya and V. B. Adiguna, "K-Means Clustering Analysis For Identifying Product Purchase Patterns Based On Country On E-Commerce Platforms," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 4, pp. 83–88, 2025, doi: 10.31004/riggs.v4i4.3296.

[13] A. R. F. Falih, R. Kurniawan, Y. Arie Wijaya, and S. Anwar, "Algoritma K-Mean Untuk Optimalisasi Model Clustering Data Penjualan Toko Online Di Tiktok Shop Dalam Strategi Pemasaran," *J. Sist. Inf. Kaputama*, vol. 9, no. 1, pp. 1–11, 2025, doi: 10.59697/jsik.v9i1.929.

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

[15] R. G. Praditya, G. Sembodo, and J. Heikal, "Market segmentation analysis to find out products and services that suit customer needs using the python KMEANS clustering method (Case study: Superindo Tambun Area, Bekasi)," *J. Tek. Ind. Terintegrasi*, vol. 7, no. 4, pp. 2072–2081, 2024, doi: 10.31004/jutin.v7i4.35889.