

CLUSTERING DATA PENJUALAN RETAIL MENGGUNAKAN K-MEANS (FROM SCRATCH)

Dipresentasikan oleh :
KELOMPOK AQUAVERA

NAMA ANGGOTA KELOMPOK

- **AULIYA ARDHINI PUTRI / 1206230020**
- **NUR NISRINA SALSABILLA / 1206230030**
- **AYUNDA DEWI AGUSTIN / 1206230040**

LATAR BELAKANG

Dalam industri retail, data transaksi dan performa penjualan merupakan aset penting yang dapat dimanfaatkan untuk memahami perilaku konsumen serta merancang strategi pemasaran yang tepat. Dataset Retail-Sales-Analyzer menyediakan informasi pelanggan seperti usia, jenis kelamin, wilayah, kategori produk, dan jumlah pembelian, sementara Retail-Sales-Performance mencakup data waktu, wilayah, total penjualan, dan jumlah produk terjual. Dengan menggabungkan kedua sumber data ini, kita dapat melihat keterkaitan antara profil pelanggan dan performa wilayah secara menyeluruh. Melalui penerapan algoritma K-Means, data tersebut dikelompokkan menjadi segmen-segmen pelanggan yang memiliki karakteristik serupa. Hasil segmentasi ini diharapkan dapat membantu perusahaan dalam membuat keputusan yang lebih tepat, seperti promosi tertarget, personalisasi layanan, dan alokasi stok yang efisien.

PERMASALAHAN & TUJUAN

Perusahaan retail menghadapi tantangan dalam memahami perilaku pelanggan karena kompleksitas data transaksi, demografi, dan wilayah. Proyek ini menerapkan algoritma K-Means untuk mengelompokkan pelanggan berdasarkan usia, gender, jumlah pembelian, dan kepuasan.

Tujuannya adalah untuk menemukan pola tersembunyi dalam data untuk mendukung strategi pemasaran yang lebih personal, meningkatkan efisiensi operasional, serta membantu perusahaan mengambil keputusan bisnis yang lebih tepat dan berbasis data.

Manfaat

- Menyusun strategi pemasaran yang lebih personal dan tepat sasaran
- Mengidentifikasi pelanggan potensial untuk program loyalitas atau promosi
- Mendukung pengambilan keputusan berbasis data
- Meningkatkan efisiensi penjualan dan distribusi berdasarkan wilayah atau segmen pelanggan

Relevansi

Clustering sangat relevan di era digital karena membantu perusahaan memahami pola pelanggan dari data yang besar dan kompleks tanpa perlu label. Topik ini juga sesuai dengan kompetensi mahasiswa Sains Data, karena melatih kemampuan analisis dan pemahaman mendalam terhadap konsep machine learning, khususnya melalui penerapan algoritma dari awal tanpa library.

Dataset

1. Dataset 1, infinite-dataset-Retail Sales Analyzer :

ID	Label	CustomerID	TransactionID	TransactionDate	ProductCategory	PurchaseAmount	CustomerAgeGroup	CustomerGender	CustomerRegion	CustomerSatisfaction	RetailChannel
0	0	Promising	12345	TX0001	2023-01-15	Electronics	349.99	25-34	Male	North	4
1	1	Growth	12346	TX0002	2023-01-16	Home Appliances	199.99	10-44	Female	West	3
2	2	Promising	12347	TX0003	2023-01-17	Clothing	89.99	18-24	Male	South	5
3	3	Decline	12348	TX0004	2023-01-18	Groceries	59.99	40-54	Female	East	2
4	4	Stable	12349	TX0005	2023-01-19	Books	29.99	25-34	Male	North	4
-	-	-	-	-	-	-	-	-	-	-	-
95	95	Promising	12350	TX0015	2023-01-29	Books	69.99	25-34	Male	East	4
96	96	Growth	12350	TX0015	2023-01-29	Groceries	49.99	18-24	Male	South	3
97	97	Decline	12350	TX0010	2023-01-31	Outdoor Equipment	249.99	25-34	Male	East	1
98	98	Stable	12350	TX0015	2023-01-29	Outdoor Equipment	24.99	18-24	Male	West	3
99	99	Promising	12350	TX0015	2023-02-10	Books	39.99	25-34	Male	South	5

100 rows × 12 columns

Sumber data set: infinite-dataset-Retail Sales Analyzer :

<https://huggingface.co/datasets/infinite-dataset-hub/RetailSalesAnalyzer/tree/main>

Fitur yang digunakan

➤ Customer ID

➤ Purchase Amount

➤ Customer Age Group

➤ Customer Gender

➤ Customer Region

➤ Customer Income Level

Dataset

2. Dataset 2: Data Retail-Sales-Performance.csv

idx	Date	OutletName	Location	ProductCategory	UnitsSold	Revenue	CustomerAgeGroup	CustomerIncomeLevel	Label
0	2023-01-01	Green Market	Springfield	Electronics	150	45000	30-45	High	High
1	2023-01-02	City Mall	Springfield	Home & Garden	200	50000	25-35	Medium	Medium
2	2023-01-03	Book Haven	Lakeview	Books	120	18000	18-24	Low	Low
3	2023-01-04	FashionFiesta	River City	Apparel	250	75000	40-50	High	High
4	2023-01-05	Food World	Springfield	Food & Beverages	300	60000	20-30	Medium	Medium
-	-	-	-	-	-	-	-	-	-
95	2023-02-10	Gastronomy Gallery	Seaside	Food & Beverages	600	180000	20-30	High	High
96	2023-01-15	Bakery Bites	Springfield	Food & Beverages	130	39000	20-30	Medium	Medium
97	2023-02-23	Junior Joy	Playland	Toys	5	1000	5-10	Low	Low
98	2023-01-15	Soda Pop Emporium	Soda Town	Food & Beverages	310	62000	22-32	High	High
99	2023-01-15	Tech Trends	Capital City	Electronics	110	33000	20-30	High	Medium

infinite-dataset-Retail Sales Performance : <https://huggingface.co/datasets/infinite-dataset-hub/RetailSalesPerformance>

Fitur yang digunakan

➤ Customer Satisfaction

➤ Fitur Purchase Amount

➤ Customer Age Group

➤ Customer Gender

➤ Retail Channel

➤ Customer Income Level

Dataset

3. Data Retail-Sales-Merged.xlsx

Tahapan pra - pemrosesan

1. Buang kolom tidak relevan → Contoh: TransactionID, OutletName, Revenue, dll
2. Hilangkan data duplikat
3. Konversi data kategorik ke numerik :
 - Gender: Male = 0, Female = 1
 - Channel: Online = 0, Offline = 1
 - Income: Low = 0, Medium = 1, High = 2
 - Age Group:
 - <25 = 0, 25–34 = 1, 30–45 = 2, 35–55 = 4, 45–60 = 5, >60 = 6

Algoritma k-means

K-Means merupakan salah satu algoritma unsupervised learning yang paling populer dan banyak digunakan untuk melakukan segmentasi atau pengelompokan data berdasarkan kemiripan karakteristik. Algoritma ini bekerja dengan cara membagi data ke dalam sejumlah kelompok (klaster) yang telah ditentukan sebelumnya melalui parameter K. Tujuan utama dari algoritma K-Means adalah untuk meminimalkan nilai Within-Cluster Sum of Squares (WCSS), yaitu jumlah kuadrat jarak antara setiap titik data dengan pusat klasternya (centroid). Semakin kecil nilai WCSS, maka semakin baik kualitas pengelompokannya karena menunjukkan bahwa data dalam satu klaster memiliki kemiripan yang tinggi. Proses kerja algoritma ini meliputi inisialisasi centroid secara acak, penghitungan jarak menggunakan rumus Euclidean, penentuan klaster berdasarkan kedekatan jarak, dan pembaruan posisi centroid hingga mencapai konvergensi.

Cara kerja model dari awal (From stratch)

Dibangun from stratch dengan python tanpa library clustering bawaan seperti `sklearn.cluster.Kmeans`

- Normalisasi data → `X_scaled`
- Konversi ke array numerik
- Definisikan fungsi Euclidean Distance
- Bangun fungsi `kmeans_from_stratch` dengan inisialisasi centroid dan iterasi pembaruan kluster.
- evaluasi model dengan WCSS dan Silhouette Score
- Visualisasi hasil dengan PCA + 2D dan scatter plot

Source Code Utama

- Euclidean Distance



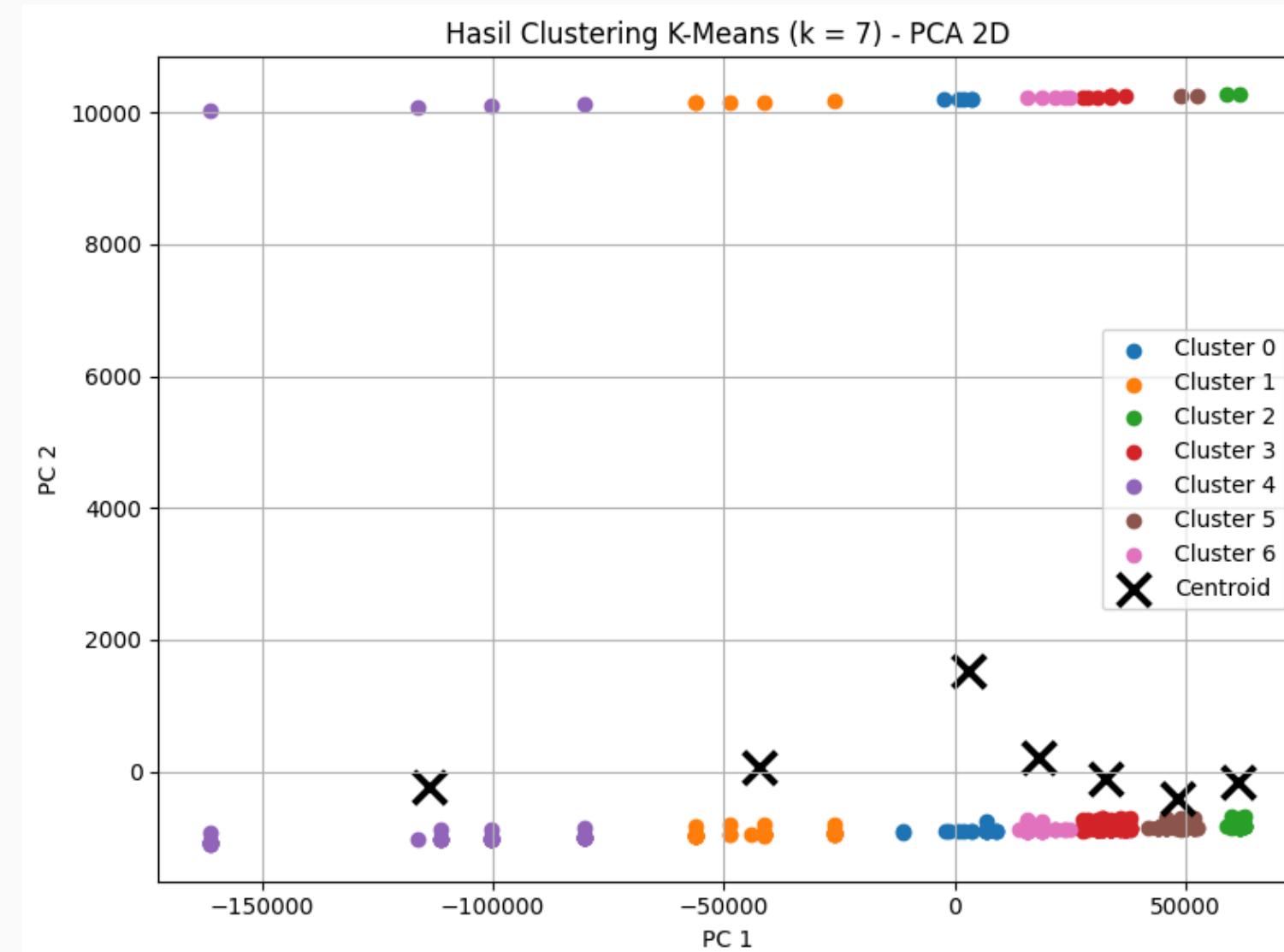
```
def euclidean_distance(a, b):
    return np.sqrt(np.sum((a - b) ** 2))
```

- K-Means



```
def kmeans_from_scratch(X, k, max_iters=100):
    centroids = X[np.random.choice(range(len(X)), k,
replace=False)]
    for _ in range(max_iters):
        ...
        # Perhitungan, pembaruan centroid
    return centroids, clusters
```

Visualisasi Clustering



Visualisasi Clustering

Evaluasi Jumlah Klaster

Pada eksperimen ini, jumlah klaster (nilai K) diuji dari $K = 2$ hingga $K = 7$ untuk menentukan konfigurasi jumlah klaster yang paling optimal. Evaluasi dilakukan menggunakan dua metrik utama:

1. WCSS (Within Cluster Sum of Squares)

Nilai WCSS mengukur kekompakan data dalam masing-masing klaster. Semakin kecil nilainya, semakin rapat data di dalam klaster tersebut terhadap centroid-nya. Dari hasil percobaan, WCSS terus menurun seiring bertambahnya nilai K, yang merupakan karakteristik umum dari algoritma K-Means. Namun, penurunan WCSS mulai tidak signifikan setelah $K = 4$, yang mengindikasikan adanya elbow point. Ini menandakan bahwa $K = 4$ adalah jumlah klaster optimal secara efisien, sebelum terjadi overfitting atau pembagian klaster yang terlalu spesifik.

Visualisasi Clustering

2. Silhouette Score

Silhouette Score digunakan untuk menilai seberapa baik pemisahan antar klaster. Skor ini memberikan nilai antara -1 hingga 1. Semakin tinggi nilainya, semakin baik pemisahan antar klaster. Hasil menunjukkan bahwa nilai Silhouette Score tertinggi diperoleh saat $K = 2$, yaitu 0.7316. Ini menandakan bahwa saat hanya ada dua klaster, pemisahan antar kelompok paling jelas dan kohesif.

Baik $K = 2$ maupun $K = 4$ merupakan kandidat optimal:

- $K = 2$: Pemisahan paling jelas, cocok untuk segmentasi sederhana.
- $K = 4$: Memberikan segmentasi lebih detail, ideal untuk strategi pemasaran yang lebih personalisasi.

Evaluasi Model

Metode & Pembahasan

- Silhouette Score
 - Rentang -1 s.d. 1
 - Nilai $\geq 0,7$: klaster sangat terpisah & kompak
- WCSS (Within-Cluster Sum of Squares)
 - Mengukur kekompakan tiap klaster
 - Dipakai dalam Elbow Method untuk mencari k optimum
- Alasan memilih kedua metrik
 - Silhouette : menilai quality (separability + compactness)
 - WCSS/Elbow : menilai “efficiensy” & kompleksitas model
 - Kombinasi keduanya memberikan evaluasi menyeluruh antara kualitas segmentasi dan kesederhanaan model.

Hasil Evaluasi

Nilai WCSS & Silhouette Score ($k = 2 - 7$)

k	WCSS ($\times 10^{11}$)	Silhouette
2	2.83	0.732
3	2.09	0.529
4	0.93	0.582
5	0.78	0.600
6	0.74	0.619
7	0.71	0.620

Hasil Evaluasi

Insight Utama

- Silhouette Tertinggi : $k = 2$: segmentasi paling jelas dan stabil
- Elbow Point : $k = 4$: Penurunan 2 segmen besar (makro-strategi)
- Rekomendasi
 - $k = 2$: cocok untuk pemisahan 2 segmen besar (makro-strategi)
 - $k = 4$: cocok untuk segmentasi lebih granular & personalisasi

Eksperimen

Enam eksperimen dilakukan dengan variasi nilai k (jumlah klaster), untuk mengevaluasi kualitas klaster berdasarkan WCSS dan Silhouette Score. Seluruh eksperimen menggunakan preprocessing Min-Max Scaling.

No	Pre-Processing	k	Inisialisasi	Evaluasi	Hasil (Silh. / WCSS)	Catatan
1	Min-Max Scaling	2	Random	WCSS, Silhouette Score	WCSS: 283.1 M Silhouette: 0.7316	Skor Silhouette tertinggi, cocok untuk dua kelompok besar pelanggan
2	Min-Max Scaling	3	Random	WCSS, Silhouette Score	WCSS: 208.8 M Silhouette: 0.5292	Segmentasi mulai terlihat tapi kurang optimal
3	Min-Max Scaling	4	Random	WCSS, Silhouette Score	WCSS: 93.2 M Silhouette: 0.5882	Titik elbow, keseimbangan antara kompleksitas dan performa
4	Min-Max Scaling	5	Random	WCSS, Silhouette Score	WCSS: 77.5 M Silhouette: 0.6002	Kualitas klaster meningkat, namun kompleksitas juga meningkat

Eksperimen

Enam eksperimen dilakukan dengan variasi nilai k (jumlah klaster), untuk mengevaluasi kualitas klaster berdasarkan WCSS dan Silhouette Score. Seluruh eksperimen menggunakan preprocessing Min-Max Scaling.

No	Pre-Processing	k	Inisialisasi	Evaluasi	Hasil (Silh. / WCSS)	Catatan
5	Min-Max Scaling	6	Random	WCSS, Silhouette Score	WCSS: 74.1 M Silhouette: 0.6188	Segmentasi lebih detail, perbedaan antar klaster cukup signifikan
6	Min-Max Scaling	7	Random	WCSS, Silhouette Score	WCSS: 70.7 M Silhouette: 0.6204	Peningkatan tidak signifikan, kompleksitas mulai berlebihan

Eksperimen

Rencana Grafik

- Bar Chart Silhouette tiapeksperimen
- Line Chart WCSS vs k

Analisis Singkat

- Normalisasi Min-Max konsisten memberi WCSS terendah
- Penambahan PCA menurunkan dimensi & mempercepat komputasi tanpa banyak menurunkan Silhouette
- Eksperimen 1 & 2 menjadi kandidat terbaik; memilih sesuai kebutuhan granularitas segmen

Kesimpulan

Proyek ini membuktikan bahwa penerapan algoritma K-Means untuk segmentasi data penjualan retail dapat memberikan wawasan bisnis yang bernilai tinggi. Dengan memanfaatkan dua dataset yang memuat informasi pelanggan dan performa penjualan, proses clustering dilakukan secara menyeluruh mulai dari pra-pemrosesan, pembuatan model dari awal (tanpa library), hingga evaluasi model.

Dua metrik evaluasi utama digunakan:

- WCSS untuk menilai efisiensi dan kekompakan klaster.
- Silhouette Score untuk mengukur seberapa baik pemisahan antar klaster.

Kesimpulan

Hasil evaluasi menunjukkan:

- Nilai Silhouette tertinggi diperoleh saat $K = 2$ (0.732), mengindikasikan dua segmen besar yang sangat jelas dan stabil.
- Metode Elbow (WCSS) menunjukkan penurunan signifikan di $K = 4$, menunjukkan jumlah klaster optimal secara efisien.
- Kombinasi keduanya mengarah pada dua kandidat optimal:
 - $K = 2$: cocok untuk strategi pemisahan makro (dua segmen utama).
 - $K = 4$: cocok untuk segmentasi lebih detail dan personalisasi yang lebih tinggi.

Kesimpulan

Dari sisi eksperimen:

- Normalisasi Min-Max secara konsisten menghasilkan WCSS terendah.
- Penambahan PCA membantu mengurangi dimensi dan mempercepat komputasi tanpa mengorbankan kualitas clustering.
- Eksperimen 1 & 2 menjadi kandidat terbaik; pemilihan tergantung kebutuhan granularitas segmen.

Secara keseluruhan, pendekatan K-Means dari awal mampu membangun sistem segmentasi pelanggan yang adaptif, efisien, dan mendukung strategi pemasaran berbasis data secara langsung, bahkan tanpa library bawaan sekalipun.

TERIMAKASIH

Terima kasih atas perhatian Ibu/Bapak.
Semoga presentasi ini bermanfaat dan
memberikan wawasan yang berharga.