Click one below to reach me 👇 😊

ilham-akbar-3301abc

ilhamakr3301@gmail.com

+62 895-2258-9852

akbarcicada

From Data to Insights:

Customer Segmentation FMD Based with Kmeans and DBSCAN

Studi Kasus: *Online Retail 2010 Dataset*

By: Ilham Akbar

Pendahuluan

Apa itu Segmentasi Pelanggan?

Segmentasi pelanggan adalah proses mengelompokkan pelanggan ke dalam segmen yang homogen berdasarkan karakteristik, perilaku, atau nilai yang dimiliki, sehingga setiap kelompok dapat dipahami secara lebih spesifik.

Mengapa perlu melakukan Segmentasi Pelanggan?

Segmentasi pelanggan adalah proses mengelompokkan pelanggan ke dalam segmen yang homogen berdasarkan karakteristik, perilaku, atau nilai yang dimiliki, sehingga setiap kelompok dapat dipahami secara lebih spesifik.

Apa itu FMD?

FMD (Frequency, Monetary, Diversity) Adalah dimensi yang digunakan dalam metode segmentasi pelanggan yang menilai:

1. **Frequency**, indikator seberapa sering pelanggan bertransaksi (misalnya jumlah invoice unik).
2. **Monetary**, indikator berapa total nilai belanja pelanggan (misalnya Σ(harga × kuantitas)).
3. **Diversity**, indikator keberagaman jenis produk yang dibeli (misalnya jumlah kode produk unik).

Mengapa menggunakan dimensi FMD?

Dimensi FMD penting karena memberi gambaran utuh perilaku pelanggan: Frequency menilai loyalitas, Monetary kontribusi nilai, dan Diversity variasi kebutuhan, sehingga bisnis dapat mengenali pelanggan bernilai tinggi dan menyusun strategi pemasaran lebih tepat.

Alur Analisis

1. Persiapan Analisis & preprocessing data

Dataset

Dataset berjudul “Online Retail 2010” yang berisikan 30 ribu baris data transaksi dan data penjualan, dengan kolom sebagai berikut:

1. Invoice = Nomor unik transaksi
2. CustomerID = ID pelanggan
3. InvoiceDate = Tanggal transaksi
4. Quantity = Jumlah produk yang dibeli
5. Price = Harga per produk
6. Country = Negara pelanggan
7. StockCode / Description = Informasi produk

Dataset ini digunakan untuk menghitung indikator Frequency, Monetary, dan Diversity (FMD) setiap pelanggan.

Load Dataset & Import Library

```
# Load Dataset
df = pd.read_csv("/content/online retail 2010 30k.csv")

# Import Library

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans, DBSCAN
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score, davies_bouldin_score
from sklearn.neighbors import NearestNeighbors
```

Preprocessing Data

Kode & Penjelasan

```
df = df[(df["Quantity"] > 0) & (df["Price"] > 0)]
df = df[~df["StockCode"].str.contains("TEST", na=False)]
```

- **Baris pertama:**
  hanya menyimpan data dengan Quantity > 0 dan Price > 0, sehingga transaksi tidak valid (retur atau harga nol) dihapus.
- **Baris kedua:**
  menghapus baris dengan StockCode mengandung "TEST", yaitu data uji/cacat yang tidak relevan dengan analisis.

Implementasi FMD dan Reduksi PCA

Kode & Penjelasan

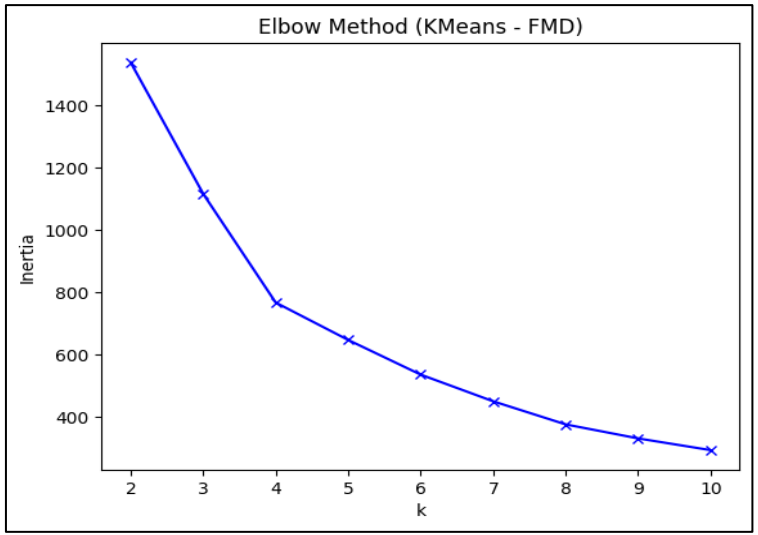
```
# Perhitungan FMD per customer
fmd = df.groupby("Customer ID").agg({
    "Invoice": "nunique",
    "Quantity": lambda x: np.sum
     (x * df.loc[x.index, "Price"]),
    "StockCode": "nunique"
}).reset_index().rename(columns={
    "Invoice": "Frequency",
    "Quantity": "Monetary",
    "StockCode": "Diversity" })

# Reduksi PCA1 dan PCA2
X_scaled = StandardScaler().fit_transform(fmd
[["Frequency","Monetary","Diversity"]])
fmd[["PCA1","PCA2"]] =
PCA(n_components=2).fit_transform(X_scaled)
```

- **FMD per Customer**
  Data dikelompokkan berdasarkan Customer ID, Frequency adalah jumlah invoice unik per pelanggan, Monetary adalah total nilai belanja (Quantity × Price), dan Diversity adalah jumlah produk unik (StockCode).
- **Reduksi Dimensi dengan PCA**
  Fitur Frequency, Monetary, Diversity distandarisasi agar setara skala. Lalu diubah menjadi PCA1 dan PCA2, yang menyederhanakan data untuk memudahkan visualisasi dan analisis cluster.

2. Eksperimen & Pemilihan Parameter KMeans

Visualisasi Elbow Plot

Output & Penjelasan



1. Sumbu X (k) menunjukkan jumlah cluster yang dicoba (2 sampai 10).
2. Sumbu Y (Inertia) menunjukkan ukuran seberapa rapat anggota dalam tiap cluster (semakin kecil semakin baik).
3. Titik tekukan (elbow point) terlihat jelas di k=3, di mana penurunan inertia mulai melambat.
4. Artinya, menambah cluster setelah k=3 tidak memberikan peningkatan signifikan dalam kualitas pemisahan cluster.
5. Dengan demikian, jumlah **cluster optimal untuk KMeans pada data ini adalah 3**.

Eksperimen dan Fit K model KMeans

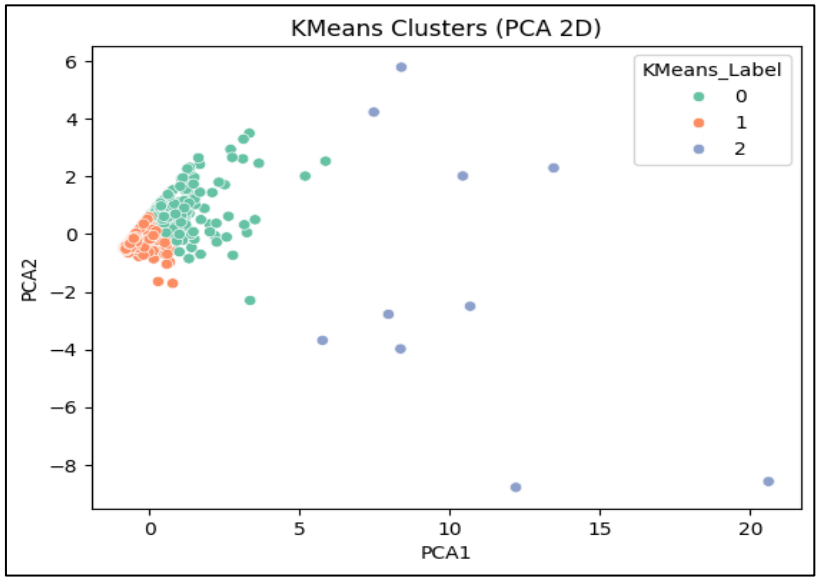
Output & Penjelasan

```
k=2 → Silhouette=0.8939, DBI=0.7429
k=3 → Silhouette=0.5781, DBI=0.8797
k=4 → Silhouette=0.5783, DBI=0.8250
k=5 → Silhouette=0.5773, DBI=0.9095
k=6 → Silhouette=0.5294, DBI=0.7841
k=7 → Silhouette=0.5350, DBI=0.9026
k=8 → Silhouette=0.5351, DBI=0.7798
k=9 → Silhouette=0.5361, DBI=0.7312
k=10 → Silhouette=0.4031, DBI=0.7700
```

1. Kode mengevaluasi K-Means (k=2–10) dengan tiga metrik: inertia (SSE), Silhouette Score (semakin tinggi semakin baik), dan Davies Bouldin Index (semakin rendah semakin baik).
2. Hasil menunjukkan k=2 punya Silhouette tertinggi (0.8939) dan DBI terendah (0.7429), artinya cluster sangat terpisah jelas.
3. Untuk k=3–5, Silhouette stabil di ~0.57 dan DBI sedikit naik (0.82–0.91), masih wajar untuk clustering yang cukup baik.
4. Berdasarkan eksperimen, **k=3 dipilih karena kualitas cluster tetap baik dan hasil lebih detail** dan mudah diinterpretasikan dibanding k=2.

Visualisasi Scatterplot PCA KMeans

Output & Penjelasan

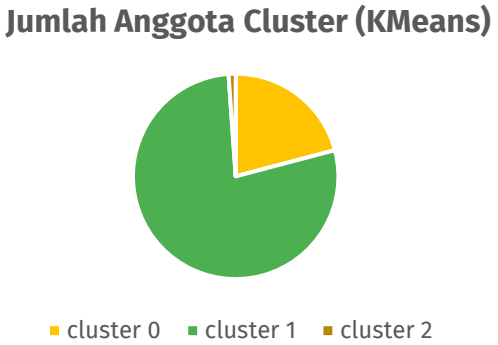


1. Data Visualisasi menunjukkan hasil K-Means dengan k=3 dalam 2 dimensi PCA untuk melihat pola distribusi cluster.
2. Warna titik merepresentasikan label cluster: hijau (0), oranye (1), dan biru (2).
3. Cluster 0 dan 1 tampak berdekatan serta sebagian tumpang tindih, mengindikasikan kemiripan karakteristik, sedangkan cluster 2 terpisah jauh, menandakan kelompok dengan ciri sangat berbeda.
4. Pola ini memperlihatkan struktur 3 cluster yang wajar, dua cluster utama yang mirip (0 & 1) serta satu cluster outlier (2) yang mewakili data dengan karakteristik unik/ekstrem.

3. Interpretasi Output KMeans

Ringkasan KMeans:

KMeans_Label	Frequency	Monetary	Diversity	Customer ID
0	2.303191	1172.305644	64.861702	188
1	1.143875	392.682070	17.582621	702
2	12.200000	20031.119000	166.100000	10

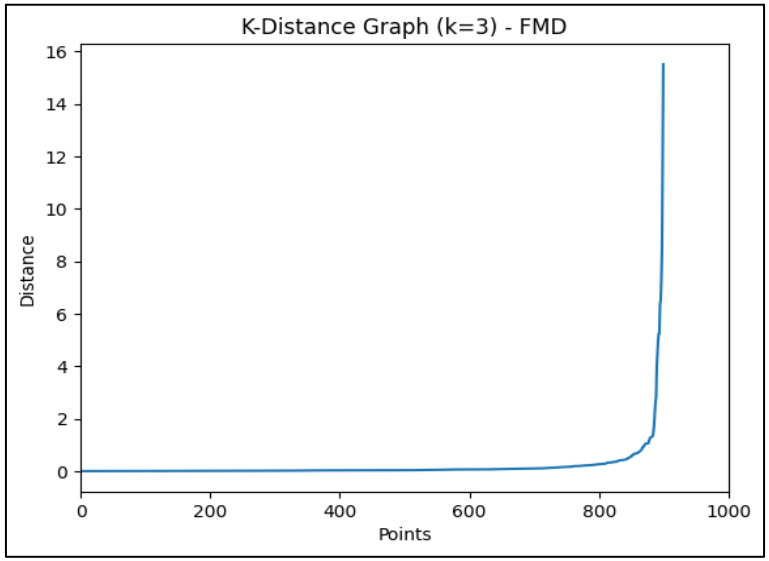


1. Cluster 0 (188 pelanggan): Frekuensi rendah (~2), monetary sedang (~1172), diversity sedang (~65). Pelanggan menengah, jarang belanja tapi sekali belanja cukup besar dan beragam, kemungkinan korporat kecil/komunitas.
2. Cluster 1 (702 pelanggan): Frekuensi sangat rendah (~1), monetary rendah (~393), diversity rendah (~18). Kelompok terbesar namun low-value, belanja sekali-sekali dengan kontribusi revenue minim, kemungkinan pelanggan individu berkebutuhan spesifik.
3. Cluster 2 (10 pelanggan): Frekuensi tinggi (~12), monetary sangat tinggi (~20031), diversity sangat tinggi (~166). Kelompok VIP, kecil tapi sangat aktif, loyal, dan berkontribusi besar, kemungkinan grosir/reseller.
4. KMeans berhasil memperlihatkan struktur 3 cluster yang jelas: dua cluster utama yang mirip (0 & 1) serta satu cluster outlier (2) yang mewakili data dengan karakteristik unik/ekstrem.

4. Eksperimen & Pemilihan Parameter DBSCAN

Visualisasi K Distance Graph

Output & Penjelasan



1. K Distance Graph digunakan dalam metode DBSCAN untuk menentukan parameter epsilon (ε), dengan melihat titik “elbow” (siku) pada kurva jarak.
2. Sumbu-X adalah titik data yang diurutkan berdasarkan jarak ke tetangga ke-3 terdekat.
3. Sumbu-Y adalah nilai jarak ke tetangga ke-3, semakin tinggi berarti semakin jauh dari titik-titik sekitarnya.
4. Kurva **relatif datar hingga sekitar titik ke-800, lalu naik tajam**, bagian siku ini menunjukkan kandidat nilai ε, yaitu jarak maksimum agar DBSCAN dapat memisahkan cluster dengan baik dan mendeteksi outlier.

Eksperimen dan Fit Eps & Min_Sample model DBSCAN

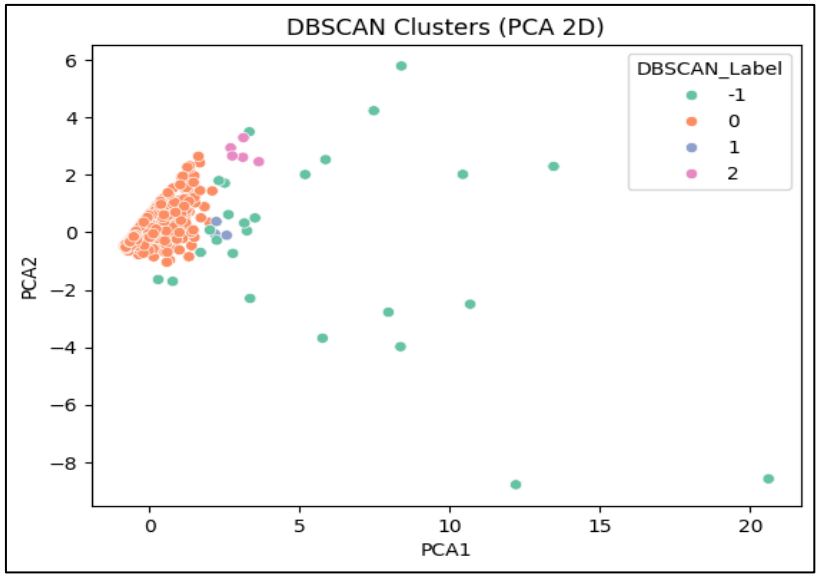
Output & Penjelasan

eps	min_samples	clusters	noise_points	silhouette	DBI
0.3	3	6	59	0.339065	1.428896
0.3	5	5	65	0.335962	1.525738
0.3	10	3	96	0.325391	1.552727
0.5	3	5	40	0.323833	1.507849
0.5	5	4	44	0.326192	1.539860
0.5	10	3	62	0.324406	1.551480
0.7	3	3	26	0.653062	1.363855
0.7	5	2	29	0.729673	1.241915

1. Dilakukan evaluasi parameter DBSCAN (eps dan min_samples) dengan menilai jumlah cluster, noise, serta kualitas melalui Silhouette dan DBI.
2. Hasil eps 0.3–0.5: Silhouette rendah (~0.32) dan DBI tinggi (~1.5), dengan noise cukup besar (40–96), menandakan pemisahan cluster tak optimal.
3. Hasil eps 0.7: kualitas meningkat signifikan, Silhouette jauh lebih tinggi (0.65–0.73), DBI lebih rendah (~1.24–1.36), dan noise berkurang (26–29), sehingga cluster lebih jelas.
4. Maka dari itu eps **0.7 dengan min_samples = 3 dipilih karena menghasilkan 3 cluster dengan keseimbangan terbaik** antara kualitas, jumlah cluster yang masuk akal, serta kemudahan interpretasi dibanding parameter lain.

Visualisasi Scatterplot PCA DBSCAN

Output & Penjelasan

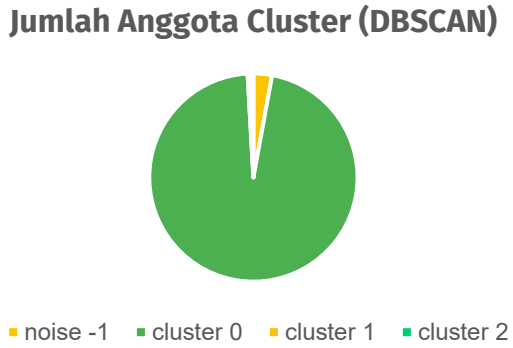


1. Data hasil clustering DBSCAN ditampilkan dalam 2 dimensi PCA agar distribusi antar-cluster mudah diamati.
2. Warna titik oranye (cluster 0), biru (cluster 1), ungu (cluster 2), dan hijau (label –1/outlier).
3. Pola Cluster 0 (oranye) mendominasi sebagai kelompok utama dan padat, sedangkan cluster 1 dan 2 relatif kecil dan terletak di sekitar pinggir cluster utama. Titik hijau tersebar jauh dari kelompok utama, menandakan outlier dengan karakteristik berbeda.
4. DBSCAN mampu mengidentifikasi mayoritas pelanggan sebagai satu cluster utama, sekaligus memisahkan kelompok kecil bernilai menengah (1 & 2) serta outlier penting (–1) yang memiliki ciri khas ekstrem dibanding mayoritas data.

5. Interpretasi Output DBSCAN

Ringkasan DBSCAN:

DBSCAN_Label	Frequency	Monetary	Diversity	Customer ID
-1	7.615385	10082.894231	113.115385	26
0	1.300231	486.490490	25.781755	866
1	6.000000	1553.416667	49.666667	3
2	3.200000	1650.240000	156.200000	5



1. Cluster 0 (866 pelanggan): Frekuensi rendah (~1,3), monetary rendah (~486), diversity rendah (~26). Kelompok umum low-value, mayoritas hanya belanja sekali-sekali dengan kontribusi revenue kecil.
2. Cluster 1 (3 pelanggan): Frekuensi cukup tinggi (~6), monetary menengah (~1553), diversity sedang (~50). Kelompok kecil dengan aktivitas belanja lebih rutin dan kontribusi menengah.
3. Cluster 2 (5 pelanggan): Frekuensi sedang (~3,2), monetary menengah (~1650), namun diversity sangat tinggi (~156). Pelanggan langka yang membeli produk sangat beragam.
4. Label -1 (26 pelanggan/outlier): Frekuensi sedang-tinggi (~7,6), monetary sangat tinggi (~10083), diversity tinggi (~113). Pelanggan premium bernilai tinggi yang tidak masuk cluster manapun.
5. DBSCAN berhasil mengidentifikasi satu cluster utama berisi mayoritas pelanggan low-value, dua cluster kecil dengan nilai menengah, serta outlier premium yang bertransaksi lebih sering, bernilai tinggi, dan beragam sehingga penting sebagai target khusus.

6. Kesimpulan & Saran

Kesimpulan

1. Total pelanggan unik = 900, terbagi dalam beberapa segmen dengan karakteristik berbeda.
2. Mayoritas pelanggan termasuk segmen bernilai rendah dengan frekuensi belanja sekali, nilai kecil, dan variasi produk sedikit.
3. Pelanggan bernilai menengah membeli dalam jumlah sedang, kadang bulk, dengan variasi produk cukup tinggi.
4. Pelanggan premium punya pola pembelian besar, sering, dan variasi sangat beragam. Segmen eksklusif yang menyumbang revenue terbesar.
5. KMeans lebih stabil, menghasilkan 3 segmen yang jelas (low, medium, premium), sedangkan DBSCAN lebih detail.

Saran

1. **Low Value:**
   pelanggan umum, perlu ditarget dengan edukasi, **promosi ringan, serta upselling/cross-selling** untuk menaikkan frekuensi dan nilai belanja dengan biaya efisien.
2. **Medium Value:**
   pembeli borongan, perlu strategi yang mendorong traksaksi rutin melalui program seperti **bundling/paket hemat** guna memperkuat pola bulk purchase.
3. **Premium/VIP:**
   segmen kecil namun penyumbang revenue terbesar, perlu layanan eksklusif dan personal seperti **diskon khusus, akses early product, Account manager, dan bonus volum**e

Thanx!

for your attention

I am fully open to suggestions, feedback, and collaboration.
Reach me through the contacts below 👇 😊

 ***ilham-akbar-3301abc***

 ***+62 895-2258-9852***

 ***ilhamakr3301@gmail.com***