

MODUL PRAKTIKUM

MATA KULIAH DATA MINING

PERTEMUAN 8

SEMESTER GENAP

TAHUN AJARAN 2024/2025



Disusun oleh:

Dwi Welly Sukma Nirad S.Kom, M.T

Aina Hubby Aziira M.Eng

Ghina Anfasha Nurhadi

Rifqi Asverian Putra

DEPARTEMEN SISTEM INFORMASI

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS ANDALAS

TAHUN 2025

IDENTITAS PRAKTIKUM

IDENTITAS MATA KULIAH

Kode mata kuliah	JSI62122
Nama mata kuliah	Data Mining
CPMK yang dibebankan pada praktikum	CPMK-3, CPMK-4 Mahasiswa mampu memahami teknik klasterisasi dalam data mining (CP-2)
Materi Praktikum Pertemuan 8	Konsep Dasar Hierarchical Clustering
	Metode Agglomerative
	Implementasi Hierarchical Clustering
	Analisis Dendrogram

IDENTITAS DOSEN DAN ASISTEN MAHASISWA

Nama Dosen Pengampu	1. Dwi Welly Sukma Nirad S.Kom, M.T 2. Aina Hubby Aziira M.Eng
Nama Asisten Mahasiswa (Kelas A)	1. 2211523034 - Muhammad Fariz 2. 2211521012 - Rizka Kurnia Illahi 3. 2211521010 - Dhiya Gustita Aqila 4. 2211522013 - Benni Putra Chaniago 5. 2211521017 - Ghina Anfasha Nurhadi 6. 2211523022 - Daffa Agustian Saadi 7. 2211521007 - Annisa Nurul Hakim 8. 2211522021 - Rifqi Asverian Putra 9. 2211521009 - Miftahul Khaira

	10. 2211521015- Nurul Afani 11. 2211523028 - M.Faiz Al-Dzikro
Nama Asisten Mahasiswa (Kelas B)	1. 2211523034 - Muhammad Fariz 2. 2211521012 - Rizka Kurnia Illahi 3. 2211521010 - Dhiya Gustita Aqila 4. 2211522013 - Benni Putra Chaniago 5. 2211521017 - Ghina Anfasha Nurhadi 6. 2211523022 - Daffa Agustian Saadi 7. 2211521007 - Annisa Nurul Hakim 8. 2211522021 - Rifqi Asverian Putra 9. 2211521009 - Miftahul Khaira 10. 2211521015- Nurul Afani 11. 2211523028 - M.Faiz Al-Dzikro

DAFTAR ISI

IDENTITAS PRAKTIKUM.....	2
IDENTITAS MATA KULIAH.....	2
IDENTITAS DOSEN DAN ASISTEN MAHASISWA.....	2
DAFTAR ISI.....	4
HIERARCHICAL CLUSTERING.....	5
A. KONSEP DASAR HIERARCHICAL CLUSTERING.....	5
B. METODE AGGLOMERATIVE.....	6
C. IMPLEMENTASI HIERARCHICAL CLUSTERING.....	7
D. ANALISIS DENDOGRAM.....	12
REFERENSI.....	14

HIERARCHICAL CLUSTERING

A. KONSEP DASAR HIERARCHICAL CLUSTERING

Hierarchical clustering merupakan salah satu metode analisis klaster *unsupervised learning* yang bertujuan untuk mengelompokkan data berdasarkan tingkat kemiripannya. Metode ini akan mengelompokkan data dengan membuat suatu bagan hirarki berupa dendrogram untuk menggambarkan hubungan antar klaster data dari yang paling mirip hingga paling berbeda. Setiap data yang mirip akan memiliki hubungan hirarki yang dekat dan membentuk klaster data. Bagan hirarki akan terus terbentuk hingga seluruh data terhubung dalam bagan hirarki tersebut. Klaster dapat dihasilkan dengan memotong bagan hirarki pada level tertentu.

Secara umum, *hierarchical clustering* dibagi menjadi dua jenis yaitu *agglomerative* dan *divisive*. Kedua metode ini dibedakan berdasarkan pendekatan dalam melakukan pengelompokkan data hingga membentuk dendrogram, dimana untuk *agglomerative* menggunakan *bottom-up manner* dan *divisive* menggunakan *top-down manner*. Metode *agglomerative* mengelompokkan data dimulai dari bawah dimana data sebagai satu klaster tersendiri dan secara iteratif menggabungkan data yang paling mirip sehingga menghasilkan sebuah klaster besar. Metode *divisive* mengelompokkan data dimulai dari atas dimana semua data sebagai satu klaster lalu secara iteratif membagi klaster hingga data menjadi satu klaster tersendiri atau sampai jumlah klaster yang diinginkan.

Dalam *hierarchical clustering*, metode untuk menentukan jarak antara dua klaster saat proses penggabungan berlangsung disebut *linkage*. Berikut ini adalah beberapa jenis *linkage* yang umum digunakan dalam *hierarchical clustering*:

1. *Single Linkage*, jarak antara dua klaster dihitung sebagai jarak minimum antara satu titik dalam klaster pertama dan satu titik dalam klaster kedua.
2. *Complete Linkage*, jarak antara dua klaster dihitung sebagai jarak maksimum antara satu titik dalam klaster pertama dan satu titik dalam klaster kedua.
3. *Average Linkage*, jarak antara dua klaster dihitung sebagai rata-rata semua jarak antar pasangan titik dari kedua klaster.

4. *Ward's Linkage*, jarak antara dua klaster dihitung berdasarkan peningkatan total variansi (spread) dalam klaster ketika dua klaster digabung.
5. *Centroid Linkage*, jarak antara dua klaster dihitung sebagai jarak Euclidean antara centroid (titik rata-rata) dari masing-masing klaster.

Kelebihan dari metode *hierarchical clustering*, yaitu :

1. Fleksibilitas terhadap berbagai jenis data
2. Tidak memerlukan penentuan jumlah klaster di awal
3. Kemampuan menangani data dengan berbagai ukuran dan bentuk
4. Bisa diterapkan untuk berbagai metrik jarak
5. Menghasilkan dendrogram yang memudahkan interpretasi

Kekurangan dari metode *hierarchical clustering*, yaitu :

1. Kompleksitas komputasi yang tinggi
2. Sensitivitas terhadap noise dan outlier
3. Kesulitan dalam menangani data berdimensi tinggi

B. METODE AGGLOMERATIVE

Agglomerative hierarchical clustering merupakan metode pengelompokan data yang bekerja dengan pendekatan *bottom-up*, artinya proses dimulai dari setiap data sebagai klaster tersendiri, kemudian secara bertahap menggabungkan klaster yang paling mirip hingga seluruh data tergabung dalam satu klaster besar. Langkah-langkah perhitungannya adalah sebagai berikut:

1. Hitung kemiripan (*similarity*) atau jarak antar klaster, dan buat *proximity matrix* (matriks kedekatan / jarak).
2. Anggap setiap titik data sebagai satu klaster tersendiri.
3. Gabungkan dua klaster yang paling mirip atau paling dekat.
4. Perbarui *proximity matrix* setelah penggabungan untuk merepresentasikan kedekatan diantara kelompok baru dan kelompok yang masih tersisa.
5. Ulangi langkah 3 dan 4 hingga seluruh data tergabung dalam satu klaster besar.

Rumus umum dalam membentuk matrik jarak, misal dengan Manhattan Distance :

$$D_{man}(x, y) = \sum_{j=1}^d |x_j - y_j|$$

atau menggunakan Euclidean Distance :

$$D(x_2, x_1) = \sqrt{\sum_{j=1}^d |x_{2j} - x_{1j}|^2}$$

Berikut beberapa metode pengelompokkan dalam *Agglomerative hierarchical clustering* :

- Single Linkage (Jarak Terdekat)

$$d_{uv} = \min\{d_{uv}\}, d_{uv} \in D$$

- Complete Linkage (Jarak Terjauh)

$$d_{uv} = \max\{d_{uv}\}, d_{uv} \in D$$

- Average Linkage (Jarak Rata-Rata)

$$d_{uv} = \text{average}\{d_{uv}\}, d_{uv} \in D$$

C. IMPLEMENTASI HIERARCHICAL CLUSTERING

Metropolitan Mall ingin meningkatkan efektivitas strategi pemasarannya dengan mengidentifikasi target pelanggan yang memiliki tingkat konversi tinggi. Untuk mencapai tujuan ini, pihak manajemen *mall* memutuskan untuk melakukan segmentasi pelanggan berdasarkan data demografis dan perilaku belanja yang telah dikumpulkan.

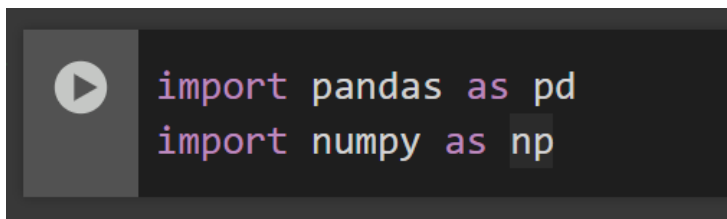
Dengan memahami profil pelanggan yang lebih mungkin berkonversi, tim pemasaran dapat merancang kampanye yang lebih tepat sasaran dan mengalokasikan sumber daya secara efisien. *Mall* menggunakan dataset pelanggan yang berisi informasi tentang CustomerID,

Gender, Age, Annual Income (dalam ribu dolar), dan Spending Score (skala 1-100) yang menggambarkan perilaku belanja pelanggan.

Dalam melakukan segmentasi ini, Metropolitan Mall menerapkan algoritma Hierarchical Clustering untuk mengelompokkan pelanggan ke dalam beberapa segmen dengan karakteristik serupa. Metode hierarki ini dipilih karena kemampuannya membentuk struktur pengelompokan bertingkat yang memungkinkan mall untuk memahami hubungan antar segmen pelanggan dengan lebih mendalam. Pendekatan ini juga memungkinkan identifikasi kelompok pelanggan yang memiliki potensi konversi tinggi berdasarkan pendapatan tahunan dan pola pengeluaran mereka.

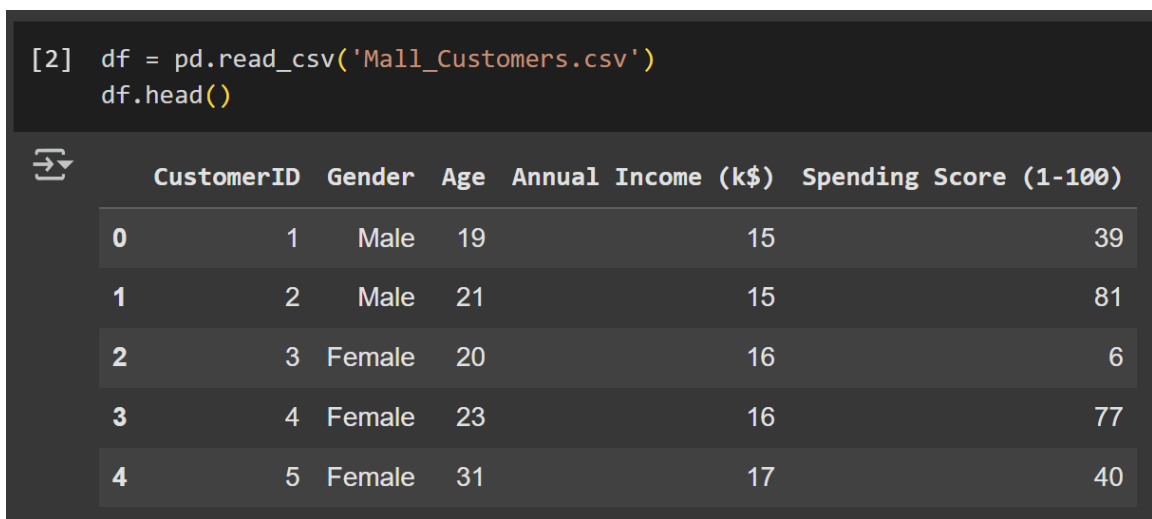
Langkah-langkah analisis segmentasi pelanggan adalah sebagai berikut:

1. Import Library yang diperlukan



```
import pandas as pd
import numpy as np
```

2. Import dataset



```
[2] df = pd.read_csv('Mall_Customers.csv')
df.head()
```

	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

3. Melihat keadaan dataset

```
[3] df.shape
```

```
(200, 5)
```

```
[4] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 200 entries, 0 to 199  
Data columns (total 5 columns):  
#   Column                Non-Null Count  Dtype    
---  ---                  
0   CustomerID            200 non-null   int64    
1   Gender                200 non-null   object    
2   Age                  200 non-null   int64    
3   Annual Income (k$)    200 non-null   int64    
4   Spending Score (1-100) 200 non-null   int64    
dtypes: int64(4), object(1)  
memory usage: 7.9+ KB
```

```
[5] df.describe()
```

```
<table>  
<thead>  
<tr>  
<th></th>  
<th>CustomerID</th>  
<th>Age</th>  
<th>Annual Income (k$)</th>  
<th>Spending Score (1-100)</th>  
</tr>  
<tbody>  
<tr>  
<td>count</td>  
<td>200.000000</td>  
<td>200.000000</td>  
<td>200.000000</td>  
<td>200.000000</td>  
</tr>  
<tr>  
<td>mean</td>  
<td>100.500000</td>  
<td>38.850000</td>  
<td>60.560000</td>  
<td>50.200000</td>  
</tr>  
<tr>  
<td>std</td>  
<td>57.879185</td>  
<td>13.969007</td>  
<td>26.264721</td>  
<td>25.823522</td>  
</tr>  
<tr>  
<td>min</td>  
<td>1.000000</td>  
<td>18.000000</td>  
<td>15.000000</td>  
<td>1.000000</td>  
</tr>  
<tr>  
<td>25%</td>  
<td>50.750000</td>  
<td>28.750000</td>  
<td>41.500000</td>  
<td>34.750000</td>  
</tr>  
<tr>  
<td>50%</td>  
<td>100.500000</td>  
<td>36.000000</td>  
<td>61.500000</td>  
<td>50.000000</td>  
</tr>  
<tr>  
<td>75%</td>  
<td>150.250000</td>  
<td>49.000000</td>  
<td>78.000000</td>  
<td>73.000000</td>  
</tr>  
<tr>  
<td>max</td>  
<td>200.000000</td>  
<td>70.000000</td>  
<td>137.000000</td>  
<td>99.000000</td>  
</tr>  
</tbody>  
</table>
```

4. Membagi dataset yang diperlukan

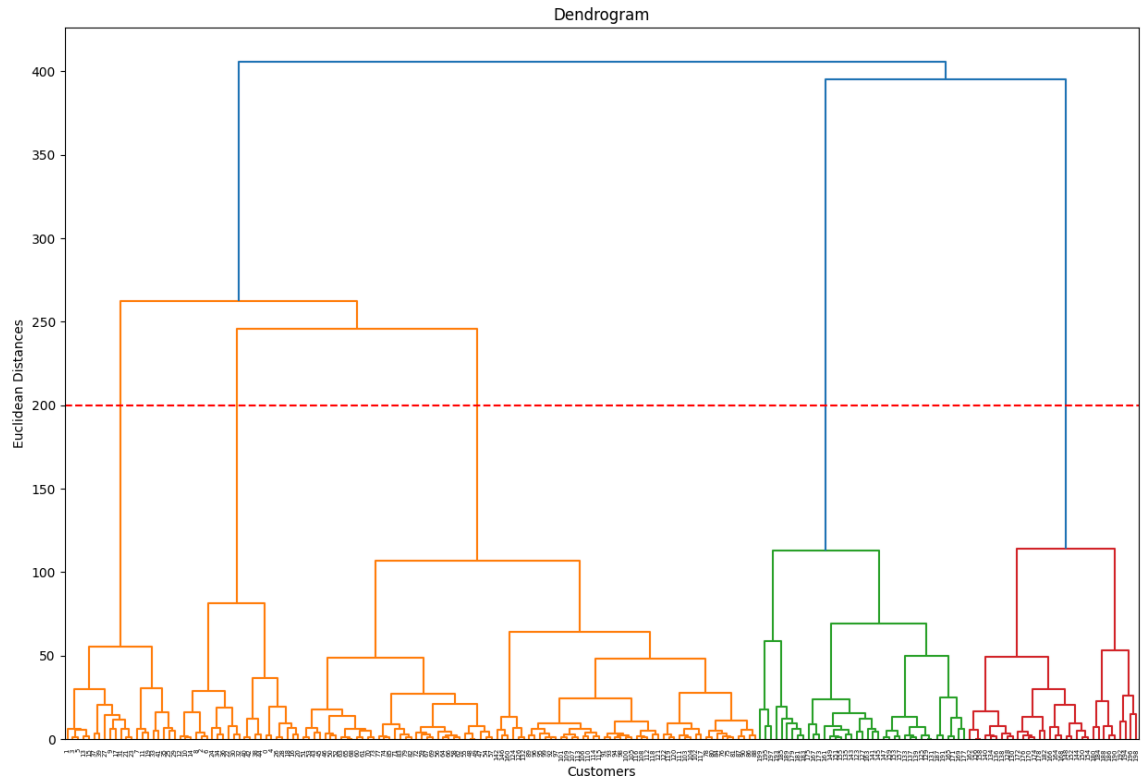
```
X = df.iloc[:, 3:]
X.head()
```

	Annual Income (k\$)	Spending Score (1-100)
0	15	39
1	15	81
2	16	6
3	16	77
4	17	40

5. Membagi dataset yang diperlukan

```
import scipy.cluster.hierarchy as hc
import matplotlib.pyplot as plt
from pylab import rcParams

rcParams['figure.figsize'] = 15, 10
dendrogram = hc.dendrogram(hc.linkage(X, method = 'ward'))
plt.title('Dendrogram')
plt.xlabel('Customers')
plt.ylabel('Euclidean Distances')
plt.axhline(200, c='r', linestyle='--')
plt.show()
```

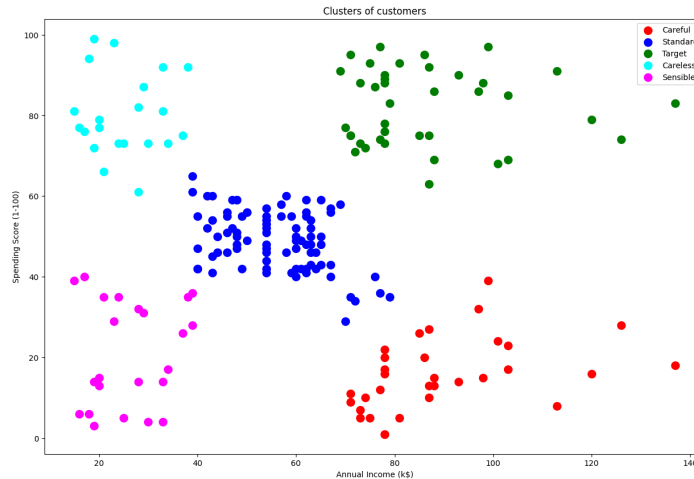


6. Menggunakan *Agglomerative hierarchical clustering Approach*

```
from sklearn.cluster import AgglomerativeClustering
hc_Agg = AgglomerativeClustering(n_clusters = 5, linkage = 'ward')
y_hc = hc_Agg.fit_predict(X)
```

7. Visualisasi

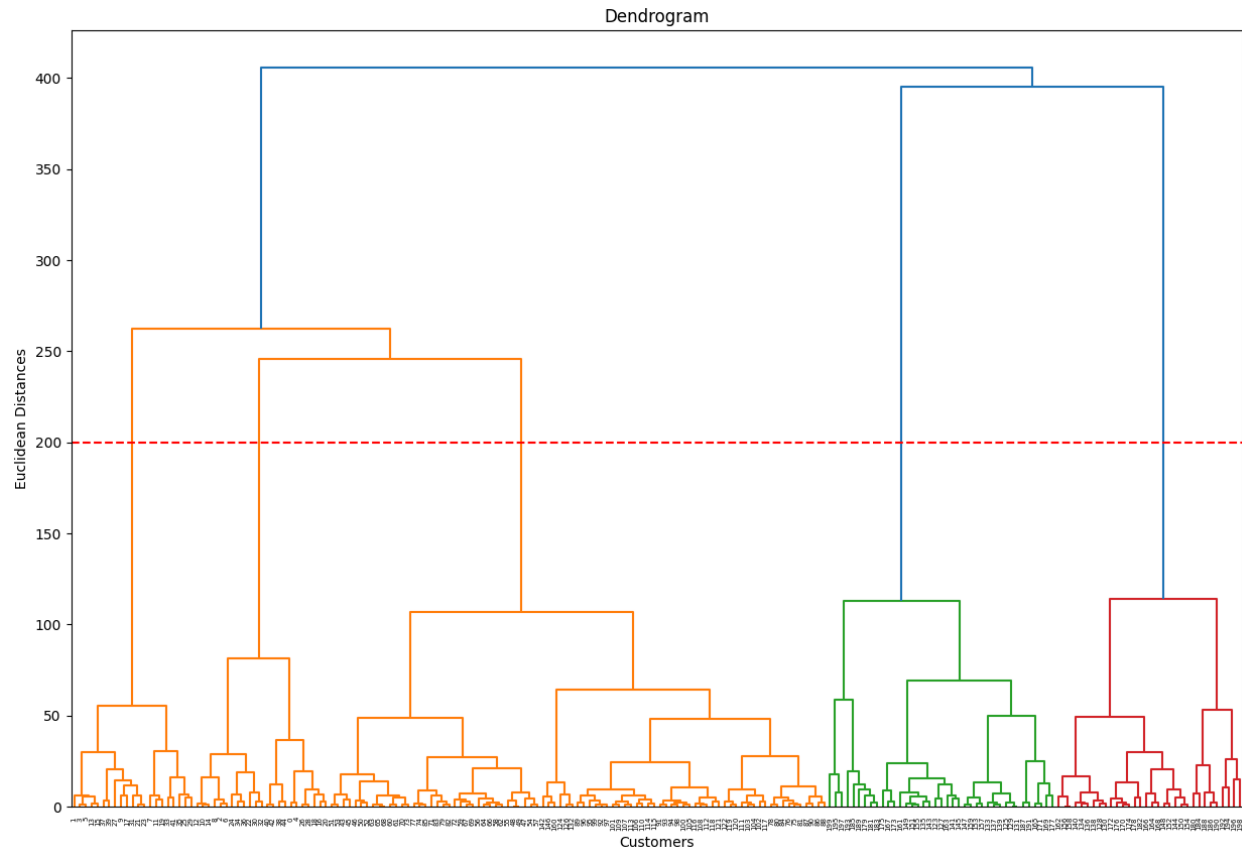
```
# Visualizing the clusters
plt.scatter(X.iloc[y_hc == 0, 0], X.iloc[y_hc == 0, 1], s = 100, c = 'red', label = 'Careful')
plt.scatter(X.iloc[y_hc == 1, 0], X.iloc[y_hc == 1, 1], s = 100, c = 'blue', label = 'Standard')
plt.scatter(X.iloc[y_hc == 2, 0], X.iloc[y_hc == 2, 1], s = 100, c = 'green', label = 'Target')
plt.scatter(X.iloc[y_hc == 3, 0], X.iloc[y_hc == 3, 1], s = 100, c = 'cyan', label = 'Careless')
plt.scatter(X.iloc[y_hc == 4, 0], X.iloc[y_hc == 4, 1], s = 100, c = 'magenta', label = 'Sensible')
plt.title('Clusters of customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()
```



D. ANALISIS DENDOGRAM

Dendrogram adalah representasi visual berbentuk diagram pohon yang digunakan dalam *hierarchical clustering* untuk menunjukkan hubungan hirarkis antara objek-objek data. Setiap cabang (atau garis) dalam dendrogram menghubungkan dua kelompok (atau data) berdasarkan tingkat kesamaan atau jarak antar mereka. Kegunaan utama dendrogram adalah untuk mencari cara terbaik untuk mengalokasikan objek ke dalam kluster. Berikut fungsi dan cara membaca dendrogram:

- Akar, bagian bawah dendrogram adalah data individual.
- Cabang, ketika dua data atau dua cluster digabungkan, terbentuk cabang.
- Ketinggian sambungan (*linkage height*), menunjukkan jarak atau dissimilarity antar kelompok yang digabungkan. Semakin tinggi titik penggabungan, semakin tidak mirip dua cluster yang disatukan.
- Dengan memotong dendrogram pada ketinggian tertentu, bisa menentukan jumlah cluster yang dihasilkan.



Pada visualisasi yang ditampilkan, terlihat dengan jelas pembagian pelanggan menjadi tiga cluster utama yang dibedakan dengan warna oranye, hijau, dan merah (walaupun garis threshold memotong 5 cluster sebagai optimal cluster). Ketiga cluster ini menunjukkan pengelompokan pelanggan berdasarkan tingkat kesamaan karakteristik mereka. Pada bagian bawah dendrogram, terlihat akar-akar yang merepresentasikan data individual pelanggan, sementara cabang-cabang yang menghubungkan mereka menunjukkan penggabungan berdasarkan kesamaan. Garis putus-putus merah horizontal yang memotong dendrogram pada ketinggian sekitar nilai 200 pada skala Euclidean Distance berfungsi sebagai threshold yang menentukan jumlah cluster yang terbentuk. Ketinggian sambungan (linkage height) yang mencapai nilai 400 pada beberapa titik menandakan tingginya perbedaan antar kelompok-kelompok tersebut. Cluster oranye di sisi kiri terlihat memiliki populasi pelanggan terbesar dengan beberapa sub-cluster, sedangkan cluster hijau dan merah di sisi kanan memiliki struktur yang lebih sederhana. Pola pengelompokan ini memberikan wawasan berharga untuk segmentasi pelanggan yang dapat digunakan dalam pengembangan strategi bisnis yang lebih tepat sasaran dan personalisasi layanan.

REFERENSI

- DQLab. 2024. Model *Machine Learning Hierarchical Clustering*. Dari <https://dqlab.id/model-machine-learning-hierarchical-clustering>. Diakses pada 2 Mei 2025.
- GeeksforGeeks. (2023). *Hierarchical clustering in data mining*. Dari <https://www.geeksforgeeks.org/hierarchical-clustering-in-data-mining/>. Diakses pada 2 Mei 2025.
- GeeksforGeeks. (2025). *ML | Types of linkages in clustering*. Dari <https://www.geeksforgeeks.org/ml-types-of-linkages-in-clustering/>. Diakses pada 2 Mei 2025.
- Inayatus, S., & Nabiilah, A. F. (2021). *Introduction to Hierarchical Clustering*. Dari <https://algotech.netlify.app/blog/introduction-to-hierarchical-clustering/>. Diakses pada 2 Mei 2025.
- Irwansyah, Edy. 2017. *Clustering*. Dari <https://socs.binus.ac.id/2017/03/09/clustering/>. Diakses pada 2 Mei 2025.
- Karabiber, Fatih. 2024. *Hierarchical Clustering*. Dari <https://www.learn datasci.com/glossary/hierarchical-clustering>. Diakses pada 2 Mei 2025.
- Keita, Zoumana. 2023. *An Introduction to Hierarchical Clustering in Python*. Dari <https://www.datacamp.com/tutorial/introduction-hierarchical-clustering-python>. Diakses pada 2 Mei 2025.
- Noble, Joshua. 2024. *What is hierarchical clustering?* Dari <https://www.ibm.com/think/topics/hierarchical-clustering>. Diakses pada 2 Mei 2025.
- Supianto, Afif. 2014. *Pengenalan Pola Hierarchical Clustering*. Dari <http://afif.lecture.ub.ac.id/files/2014/05/Slide-12-Klasterisasi-Hierarchical-Clustering.pdf>. Diakses pada 2 Mei 2025.