

MODELLING - MODUL 8 (TUGAS VERSI A/1)

Nama : Mohammad Delvin Farhan Akbar
Dataset : Mall Customer Segmentation Data (Mall_Customers.csv)
Tools : Python, Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn

Link Notebook : <https://colab.research.google.com/drive/1WslHeoqfQaWPS5OtrbxWL67ZideGRhus-sharing>

Link Dataset : https://www.kaggle.com/datasets/choudhurycustomer-segmentation/tutorial-in-python>Select-Mall_Customers.csv

Dataset yang digunakan berjudul Mall Customers Segmentation Data yang berisi informasi pelanggan dari sebuah pusat perbelanjaan.

Tujuan analisis ini adalah melakukan segmentasi pelanggan menggunakan metode Clustering (K-Means) agar manajemen mall dapat mengidentifikasi kelompok pelanggan dengan karakteristik berbeda.

1. 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
from sklearn.metrics import silhouette_score
```

2. Load Dataset & Data Overview

```
df = pd.read_csv("Mall_Customers.csv")
df.head()
```

Next steps: [Generate code with df](#) [New interactive sheet](#)

```
df.info()
df.describe()
df.isnull().sum()

class pandas.core.frame.DataFrame:
    RangelineIndex: 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
```

3. Data Preparation

```
data = df.drop("CustomerID", axis=1)
data["Gender"] = data["Gender"].map({"Male": 0, "Female": 1})
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data)

scaled_df = pd.DataFrame(scaled_data, columns=data.columns)
scaled_df.head()
```

Next steps: [Generate code with scaled_df](#) [New interactive sheet](#)

1. Dataset tidak memiliki missing value, seluruh 200 baris data lengkap.

2. Kolom CustomerID dihapus karena tidak relevan dalam analisis.

3. Kolom Gender dikonversi ke bentuk numerik.

 • Male → 0

 • Female → 1

4. Semua kolom numerik (Gender, Age, Annual Income, Spending Score) dinormalisasi menggunakan StandardScaler agar tiap fitur memiliki skala yang setara.

4. Exploratory Data Analysis (EDA)

```
sns.countplot(x="Gender", data=df)
plt.title("Distribusi Gender")
fig, axes = plt.subplots(1, 2, figsize=(12, 4))
sns.histplot(df["Age"], bins=20, kde=True, ax=axes[0])
axes[0].set_title("Distribusi Umur")
sns.histplot(df["Annual Income (k$)"], bins=20, kde=True, ax=axes[1])
axes[1].set_title("Distribusi Pendapatan Tahunan")
plt.show()
```

```
plt.figure(figsize=(6, 5))
sns.scatterplot(x="Age", y="Spending Score (1-100)", data=df)
plt.title("Hubungan Umur dan Skor Pengeluaran")
plt.show()
```

Distribusi Gender

Count (0-100)

Male Female

Distribusi Umur

Count (0-25)

Age (20-70)

Distribusi Pendapatan Tahunan

Count (0-25)

Annual Income (k\$) (20-140)

Hubungan Umur dan Skor Pengeluaran

Spending Score (1-100)

Age (20-70)

1. Distribusi Gender:

Jumlah pelanggan perempuan lebih banyak dibanding laki-laki.

2. Distribusi Umur:

Mayoritas pelanggan berusia antara 25-40 tahun.

3. Distribusi Pendapatan:

Pendapatan tahunan pelanggan berkisar antara 15k-140k dollar.

4. Hubungan Umur dan Skor Pengeluaran:

Pelanggan muda cenderung memiliki skor pengeluaran lebih tinggi dibanding pelanggan yang lebih tua.

Visualisasi yang digunakan:

• Histogram untuk umur & pendapatan.

• Scatterplot antara Age dan Spending Score.

5. Clustering dengan K-Means

Algoritma yang digunakan adalah K-Means Clustering, karena cocok untuk data numerik dan mudah diinterpretasikan.

```
wcss = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
    kmeans.fit(scaled_df)
    wcss.append(kmeans.inertia_)

plt.plot(range(1, 11), wcss, marker='o')
plt.title('Elbow Method')
plt.xlabel('Jumlah Cluster (k)')
plt.ylabel('WCSS')
```

Elbow Method

WCSS (200-800)

Jumlah Cluster (k) (2-10)

Dari grafik WCSS, titik siku terletak di k=5.

6. Model K-Means

```
kmeans = KMeans(n_clusters=5, random_state=42, n_init=10)
df['Cluster'] = kmeans.fit_predict(scaled_df)
df.head()
```

CustomerID Gender Age Annual Income (k\$) Spending Score (1-100) Cluster

0 1 Male 19 15 39 3 5

1 2 Male 21 15 81 3 5

2 3 Female 20 16 6 3 5

3 4 Female 23 16 77 3 5

4 5 Female 31 17 40 3 5

Next steps: [Generate code with df](#) [New interactive sheet](#)

7. Evaluasi Model (Silhouette Score)

Evaluasi dilakukan menggunakan Silhouette Score, yang mengukur seberapa baik setiap objek ditempatkan dalam cluster-nya.

```
silhouette_avg = silhouette_score(scaled_df, df['Cluster'])
print(f"Silhouette Score: {silhouette_avg:.3f}")
Silhouette Score: 0.394
```

Nilai ini menunjukkan bahwa hasil clustering cukup baik, meskipun masih terdapat sedikit tumpang tindih antar-cluster.

8. Visualisasi Hasil Clustering

```
plt.figure(figsize=(8, 6))
sns.scatterplot()
```

