Praktikum 5

Clustering Pada Data Customer Menggunakan Algoritma

K-Mean

Disusun Oleh:

Muhammad Risalah Naufal

(21051214008)

S1 Sistem Informasi B

Mata Kuliah: Data Mining

Dosen Pengampu: Wiyli Yustanti, S.Si., M.Kom.

PROGRAM STUDI SISTEM INFORMASI

RUMPUN TEKNIK INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS NEGERI SURABAYA 2023

A. Dataset Customer

(https://www.kaggle.com/datasets/tohuangjia/customer-cluster/data)

Dataset ini merupakan dataset yang diperuntukan untuk clustering. Pada dataset ini dapat dilakukan clustering untuk menentukan segmen pasar yang cocok untuk perusahaan dengan melihat hasil clustering dari data customer perusahaan.

Fitur yang terdapat pada dataset ini adalah:

1. ID

   Diperuntukan untuk mengetahui penomeran dari data yang ada, nantinya akan dihapus atau *drop* dikarenakan tidak memberikan dampak yang signifikan pada clustering data nantinya.

2. Gender

   Jenis kelamin para pembeli pada perusahaan yang telah dilakukan pendataan. nanti akan digunakan untuk dibandingkan dengan fitur lain setelah itu akan menghasilkan clustering yang dapat diambil informasinya untuk menentukan keputusan kedepannya.

3. Age

   Untuk menentukan pelanggan usia berapa saja yang melakukan pembelanjaan pada perusahaan ini dan nanti akan digunakan untuk dibandingkan dengan fitur lain setelah itu akan menghasilkan clustering yang dapat diambil informasinya untuk menentukan keputusan kedepannya.

4. Income

   Untuk menentukan berapa pendapatan pelanggan yang melakukan pembelanjaan pada perusahaan ini dan nanti akan digunakan untuk dibandingkan dengan fitur lain setelah itu akan menghasilkan clustering yang dapat diambil informasinya untuk menentukan keputusan kedepannya.

5. Spending

   Untuk menentukan berapa pengeluaran pelanggan ketika berbelanja disini dan nanti akan digunakan untuk dibandingkan dengan fitur lain setelah itu akan menghasilkan clustering yang dapat diambil informasinya untuk menentukan keputusan kedepannya.

Pada clustering ini nanti diharapkan hasilnya dapat memberikan informasi yang dapat meningkatkan penjualan pada perushaan dan mengetahui segmen apa saja yang perlu ditingkatkan nantinya.

B. Metode Penelitian

1. Masukan library

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from yellowbrick.cluster import KElbowVisualizer, SilhouetteVisualizer
```

Pada code tersebut, Pandas digunakan untuk manipulasi data, NumPy untuk operasi numerik, Matplotlib dan Seaborn untuk visualisasi, serta Yellowbrick untuk visualisasi clustering. Selanjutnya, kita dapat menggunakan KElbowVisualizer dan SilhouetteVisualizer dari Yellowbrick untuk analisis clustering.

2. Import data

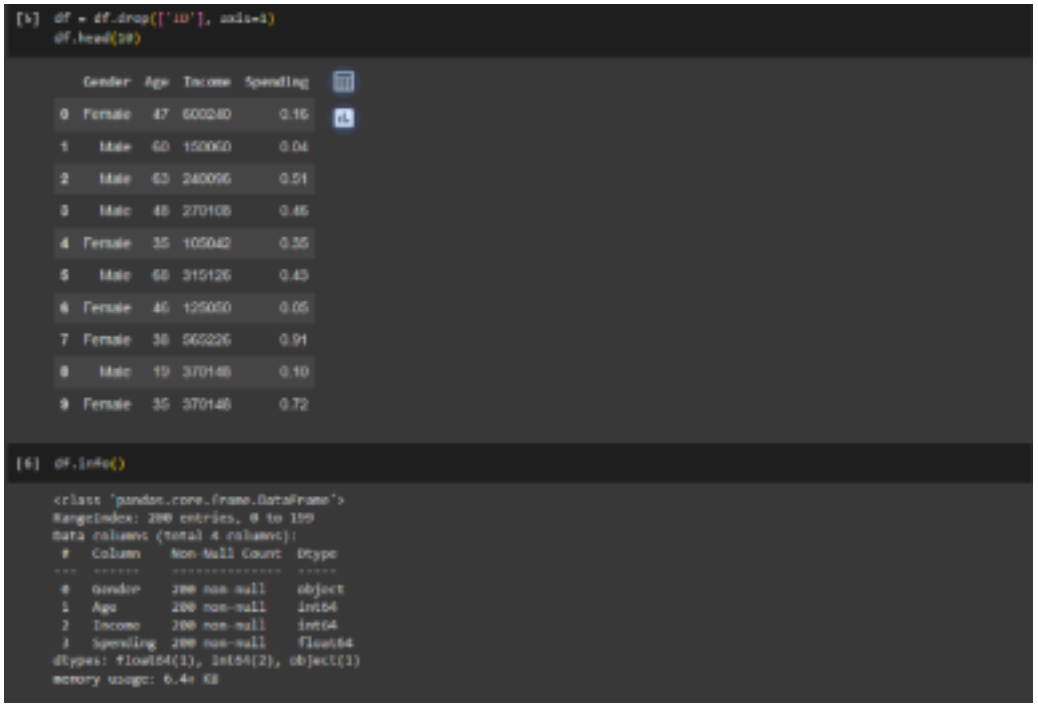
```
[2] nama_kolom = ['ID','Gender','Age','Income','Spending']
```

```
[4] df = pd.read_csv('Customers Cluster.csv')
    df.head()
```

	ID	Gender	Age	Income	Spending
0	1	Female	47	600240	0.16
1	2	Male	60	150060	0.04
2	3	Male	63	240096	0.51
3	4	Male	48	270108	0.46
4	5	Female	35	105042	0.35

Masukan dataset dengan nama “Customer Cluster.csv” untuk menggunakan dataset tersebut pada aplikasi google colab, dan berikan nama masing-masing fiturnya.

3. Pre-Processing



```
[5] df = df.drop(['ID'], axis=1)
    df.head(10)
```

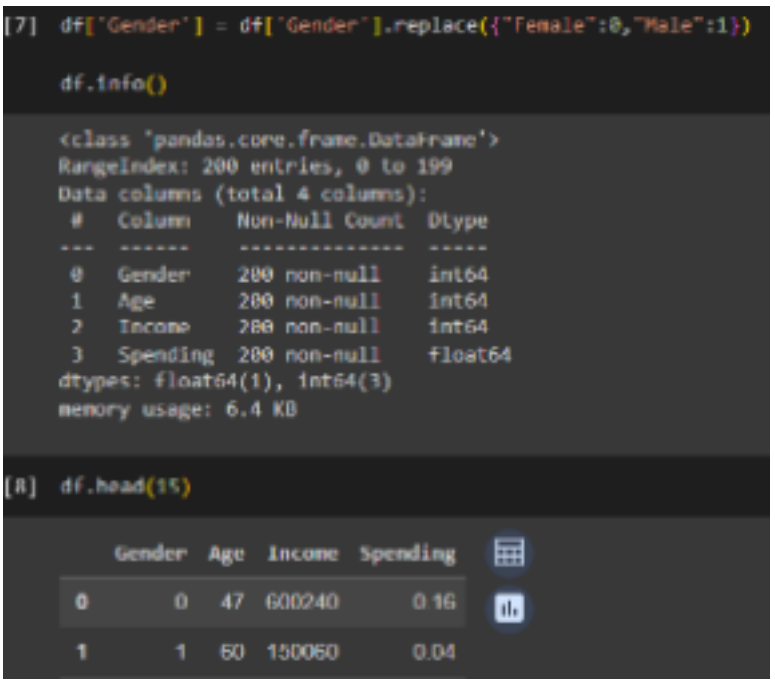
	Gender	Age	Income	Spending
0	Female	47	600240	0.16
1	Male	60	150060	0.04
2	Male	63	240096	0.51
3	Male	48	270108	0.46
4	Female	35	105042	0.35
5	Male	68	315126	0.43
6	Female	46	125000	0.05
7	Female	38	565226	0.91
8	Male	19	370148	0.10
9	Female	35	370148	0.72

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Gender    200 non-null    object
 1   Age       200 non-null    int64
 2   Income    200 non-null    int64
 3   Spending  200 non-null    float64
dtypes: float64(1), int64(2), object(1)
memory usage: 6.4+ KB
```

Untuk membersihkan data agar hasil yang keluar tidak bias. Dilakukan penghapusan fitur serta dilakukan juga pengecekan apakah ada data yang kosong atau tidak.

4. Mengubah string ke number



```
[7] df['Gender'] = df['Gender'].replace({"Female":0,"Male":1})

    df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Gender    200 non-null    int64
 1   Age       200 non-null    int64
 2   Income    200 non-null    int64
 3   Spending  200 non-null    float64
dtypes: float64(1), int64(3)
memory usage: 6.4 KB
```

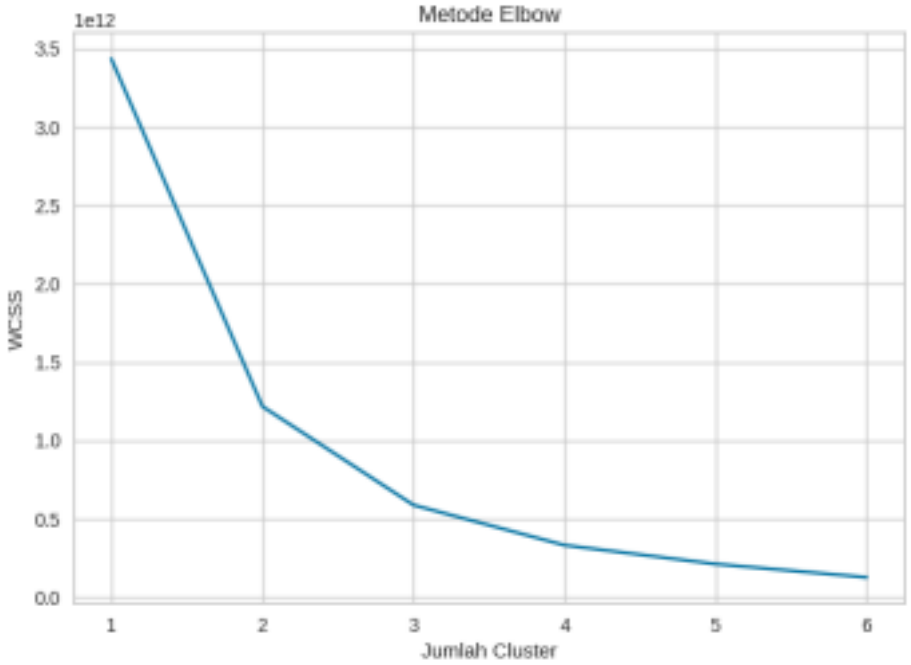
```
[8] df.head(15)
```

	Gender	Age	Income	Spending
0	0	47	600240	0.16
1	1	60	150060	0.04

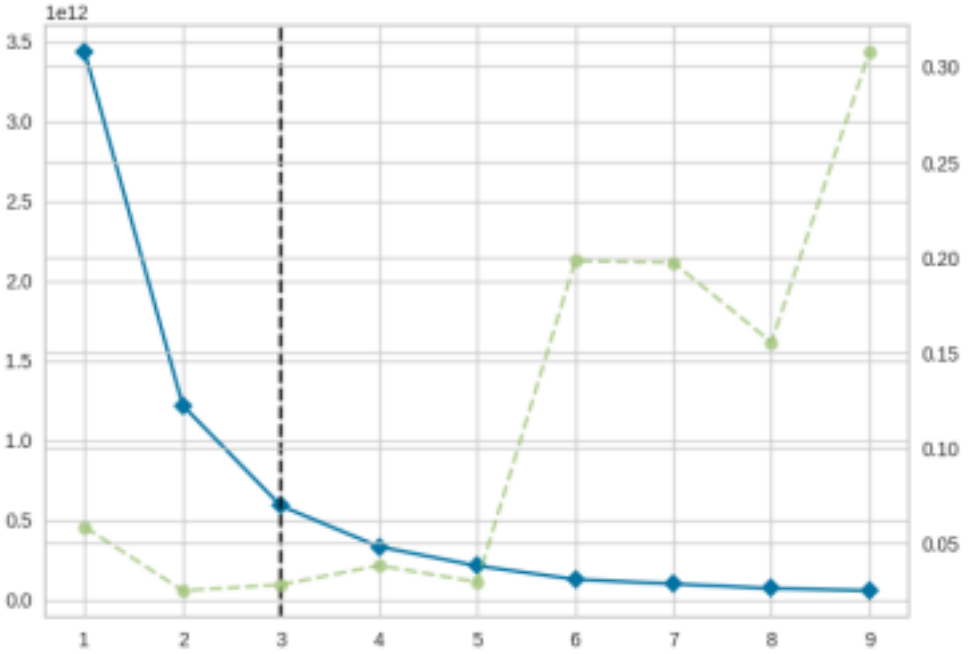
Dilakukan untuk mempermudah clustering dengan data berupa angka.

5. Metode Elbow

```
[11] from sklearn.cluster import KMeans
     wcss = []
     for i in range(1, 7):
         kmeans = KMeans(n_clusters=i, init='k-means++', n_init=10, random_state=42)
         kmeans.fit(df)
         wcss.append(kmeans.inertia_)
     plt.plot(range(1, 7), wcss)
     plt.title('Metode Elbow')
     plt.xlabel('Jumlah Cluster')
     plt.ylabel('WCSS')
     plt.show()
```



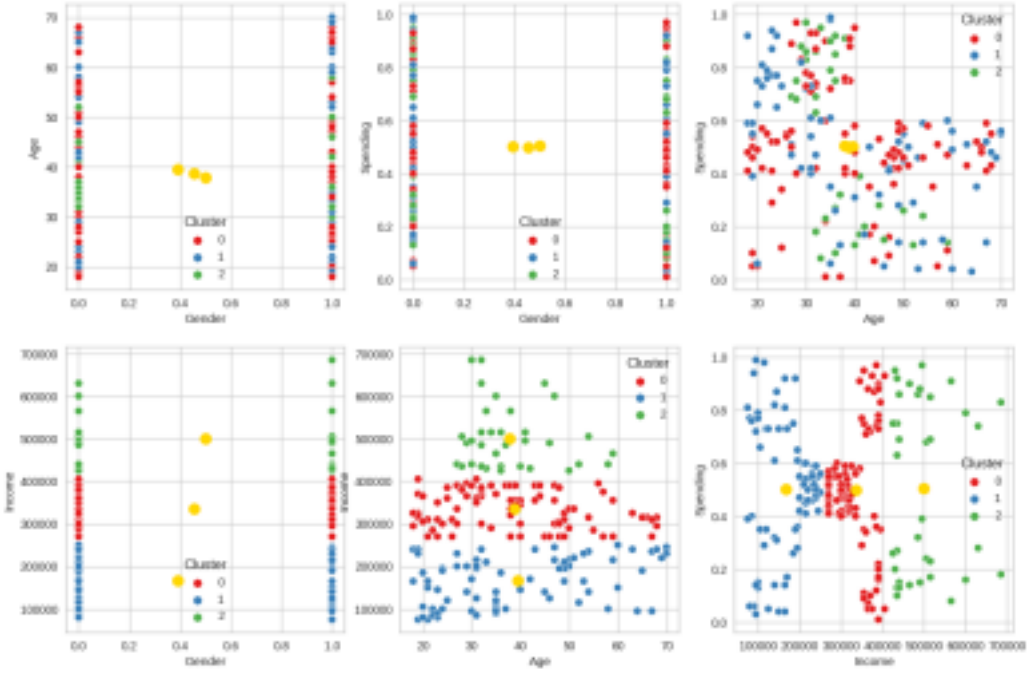

```
[12] kmeans = KMeans(n_clusters=4, random_state=0)
     vis_elbow = KElbowVisualizer(kmeans, k = (1,10))
     vis_elbow.fit(df)
     vis_elbow.poof
```



Metode "elbow" digunakan dalam algoritma k-means untuk membantu menentukan jumlah optimal dari cluster (kelompok) yang harus dibentuk dari data. Tujuan dari metode ini adalah menemukan "siku" (elbow) pada grafik jumlah cluster versus nilai inersia (inertia).

6. Clustering

```
[13] X_numerics = df[['Gender','Age','Income','Spending']]

[14] KM_5_clusters = KMeans(n_clusters=3, init='k-means++', n_init=10).fit(X_numerics) # initialize and fit K-Means model

     KMS_clustered = X_numerics.copy()
     KMS_clustered.loc[:,'Cluster'] = KM_5_clusters.labels_ # append labels to points

[15] fig1, axes = plt.subplots(2, 3, figsize=(15, 10))

     sns.scatterplot(x='Gender', y='Age', data=KMS_clustered,
                     hue='Cluster', palette='Set1', ax=axes[0, 0], legend='full')

     sns.scatterplot(x='Gender', y='Income', data=KMS_clustered,
                     hue='Cluster', palette='Set1', ax=axes[1, 0], legend='full')

     sns.scatterplot(x='Gender', y='Spending', data=KMS_clustered,
                     hue='Cluster', palette='Set1', ax=axes[0, 1], legend='full')

     sns.scatterplot(x='Age', y='Income', data=KMS_clustered,
                     hue='Cluster', palette='Set1', ax=axes[1, 1], legend='full')

     sns.scatterplot(x='Age', y='Spending', data=KMS_clustered,
                     hue='Cluster', palette='Set1', ax=axes[0, 2], legend='full')

     sns.scatterplot(x='Income', y='Spending', data=KMS_clustered,
                     hue='Cluster', palette='Set1', ax=axes[1, 2], legend='full')


     axes[0, 0].scatter(KM_5_clusters.cluster_centers_[:,0],KM_5_clusters.cluster_centers_[:,1], marker='o', s=100, c="gold")
     axes[1, 0].scatter(KM_5_clusters.cluster_centers_[:,0],KM_5_clusters.cluster_centers_[:,2], marker='o', s=100, c="gold")
     axes[0, 1].scatter(KM_5_clusters.cluster_centers_[:,0],KM_5_clusters.cluster_centers_[:,3], marker='o', s=100, c="gold")
     axes[1, 1].scatter(KM_5_clusters.cluster_centers_[:,1],KM_5_clusters.cluster_centers_[:,2], marker='o', s=100, c="gold")
     axes[0, 2].scatter(KM_5_clusters.cluster_centers_[:,1],KM_5_clusters.cluster_centers_[:,3], marker='o', s=100, c="gold")
     axes[1, 2].scatter(KM_5_clusters.cluster_centers_[:,2],KM_5_clusters.cluster_centers_[:,3], marker='o', s=100, c="gold")
     plt.show()
```



Setelah mendapatkan hasil dari clustering maka kita dapat mengambil informasi yang ada dan diolah untuk menentukan tujuan perusahaan selanjutnya harus melakukan apa agar perusahaan mendapatkan keuntungan.

C. Kesimpulan

- Dataset Customer Clustering dapat dilakukan clustering dengan algoritma K-Mean dan menghasilkan perbandingan clustering antar fitur.
- Hasil dari diagram elbow menunjukan bahwa pembagian menjadi 3 cluster adalah pembagian yang paling optimal daripada pembagian lainnya.
- Pada Clustering dimana membandingkan fitur "Gender" dengan fitur yang lainnya hanya memunculkan dua garis lurus yang dimana 0.0 adalah data pembeli wanita dan 1.0 adalah pembeli pria.
- Dari cluster perbandingan fitur "Gender" dan "Age" menunjukan bahwa rata-rata pembeli berusia 40 tahun dan cenderung berkelamin wanita.
- Dari cluster perbandingan fitur "Gender" dan "Spending" menunjukan bahwa rata-rata pembelian sebesar 0,5 Juta dan cenderung berkelamin wanita.
- Dari cluster perbandingan fitur "Age" dan "Spending" menunjukan bahwa rata-rata pembelian sebesar 0,5 Juta dan pembeli rata rata berusia 40 tahun.
- Dari cluster perbandingan fitur "Gender" dan "Income" menunjukan bahwa pengklasteran gaji pembeli beragam, mulai dari 2 juta, 3,5 juta, dan 5 juta. Untuk gaji atau pendapatan 2-3,5 juta cenderung dimiliki oleh pembeli wanita.
- Dari cluster perbandingan fitur "Age" dan "Income" menunjukan bahwa pengklasteran gaji pembeli yang rata-rata berusia 40 tahun adalah beragam, mulai dari 2 juta, 3,5 juta, dan 5 juta.
- Dari cluster perbandingan fitur "Income" dan "Spending" menunjukan bahwa pengklasteran gaji pembeli adalah 2 juta, 3,5 juta, dan 5 juta. Dimana ketiga cluster tersebut memiliki nilai pembelian pada perusahaan dengan rata rata 0,5 juta
- Dari hasil clustering ini dapat dipastikan bahwa perusahaan memiliki pengunjung yang rata-rata berusia 40 tahun dengan pendapatan yang beragam serta pemilihan atau pembelian barang oleh customer yang memiliki rata rata 0,5 juta disaat pembelian. Dengan data tersebut, perusahaan dapat menentukan apa yang harus ditingkatkan dan apa yang harus diganti guna memuaskan pelanggan pada perusahaan tersebut.

D. Daftar Pustaka

Colab:
https://colab.research.google.com/drive/1Sq5ll_2lPey_VzjZnrEs9XJ5oAgxvsV_#scrollTo=F Yqg8ZXN_WQd
Dataset:
https://www.kaggle.com/datasets/tohuangjia/customer-cluster/data