Praktikum 7

Algoritma DBSCAN Dataset Customer Cluster

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, data, dan pre-process

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings as warnings
warnings.filterwarnings("ignore")
import sklearn
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScaler
from scipy.cluster.hierarchy import linkage
import scipy.cluster.hierarchy as sch
from sklearn.cluster import AgglomerativeClustering, KMeans, KMeans
from sklearn.neighbors import NearestNeighbors
from sklearn.cluster import DBSCAN
from sklearn.decomposition import PCA
from sklearn.metrics.pairwise import cosine_similarity

from yellowbrick.cluster import KElbowVisualizer, SilhouetteVisualizer
```

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

Library dimasukan untuk keperluan eksplorasi command atau perintah agar bisa melakukan kerja dengan berbagai macam bentuk dan bisa meluaskan perintah yang kita inginkan. Lalu lakukan import data dan lakukan pre-process.

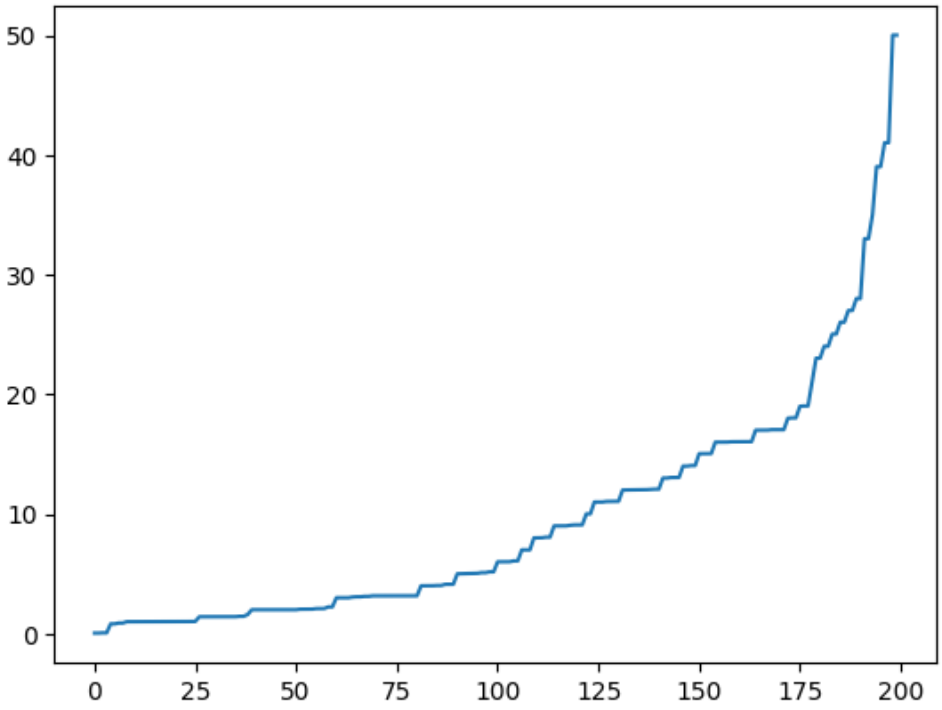
2. Preprocessing Data dengan Nearest Neighbors untuk Analisis DBSCAN

```
[46] nn = NearestNeighbors(n_neighbors=8)
     nn = nn.fit(df)
     distances, indices = nn.kneighbors(df)
```

Kode ini berguna untuk mendapatkan informasi nearest neighbors yang dapat digunakan sebagai langkah awal dalam penerapan algoritma DBSCAN untuk mengidentifikasi cluster berbasis kerapatan dalam data.

3. Grafik Elbow dari Jarak Terdekat Kedua

```
[47] distances = np.sort(distances, axis=0)
     distances = distances[:,1]
     plt.plot(distances)
     plt.show()
```



Kode di atas bertujuan untuk membantu menentukan parameter epsilon dalam algoritma DBSCAN dengan menggunakan plot grafik elbow. Dengan melihat grafik ini, dapat memilih nilai epsilon yang sesuai untuk digunakan dalam algoritma DBSCAN.

4. Penerapan pada DataFrame dan Visualisasi Hasil Clustering

```
[49] import pandas as pd
     from sklearn.cluster import DBSCAN

     # Assuming df is your NumPy array
     # Convert NumPy array to DataFrame
     df = pd.DataFrame(df)

     # Reset the index
     df_reset = df.reset_index(drop=True)

     # Apply DBSCAN
     d_cluster = DBSCAN(eps=1.125, min_samples=5)
     df_reset['cluster'] = d_cluster.fit_predict(df_reset)

     # Check the results
     print(df_reset.head(70))
```

```
    Gender  Age  Income  Spending  cluster
0        0   47  600240      0.16       -1
1        1   60  150060      0.04       -1
2        1   63  240096      0.51       -1
3        1   48  270108      0.46       -1
4        0   35  105042      0.35       -1
..     ...  ...     ...       ...      ...
65       1   19  240096      0.59       -1
```

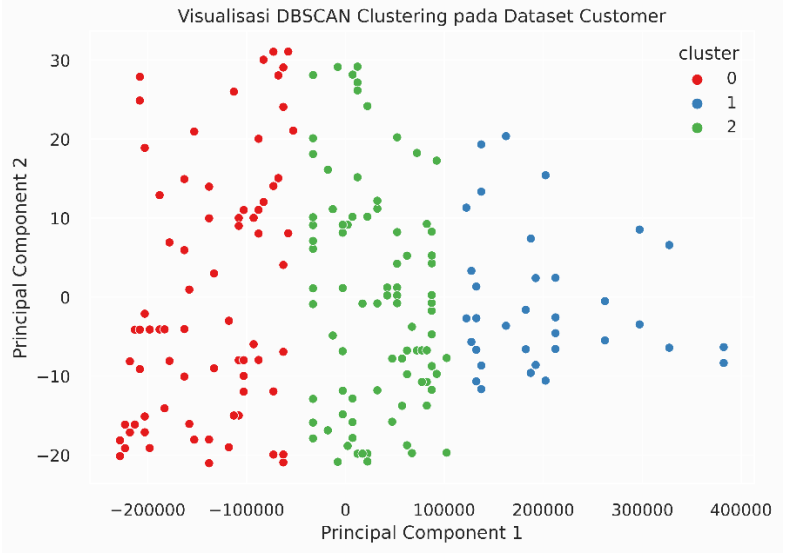
Dengan kode ini dapat melihat hasil dari proses DBSCAN pada data, di mana setiap baris data telah diberi label cluster sesuai dengan pengelompokan yang dilakukan oleh algoritma.

5. Analisis Data dengan DBSCAN setelah Reduksi Dimensi menggunakan PCA

```
[51] pca = PCA(n_components=2)
     data_2d = pca.fit_transform(df)
     df['PCA1'] = data_2d[:, 0]
     df['PCA2'] = data_2d[:, 1]
```

6. Visualisasi Hasil Clustering DBSCAN setelah Reduksi Dimensi PCA pada Dataset Kejahatan

```
[53] sns.scatterplot(x='PCA1', y='PCA2', hue='cluster', data=df, palette='Set1')
     plt.xlabel('Principal Component 1')
     plt.ylabel('Principal Component 2')
     plt.title('Visualisasi DBSCAN Clustering pada Dataset Customer')
     plt.show()
```



Dengan kode ini dapat melihat visualisasi dua dimensi dari hasil clustering yang diterapkan menggunakan algoritma DBSCAN pada data yang sudah direduksi dimensinya dengan PCA.

7. Melihat nilai siluet

```
[85] from sklearn.cluster import DBSCAN
     from sklearn.metrics import silhouette_score
     from sklearn.preprocessing import StandardScaler

     # Assuming scaleddata is your scaled dataset
     d_cluster = DBSCAN(eps=1.125, min_samples=5)
     scaleddata['cluster'] = d_cluster.fit_predict(scaleddata)

     # Check the number of unique labels
     unique_labels = len(set(d_cluster.labels_))

     if unique_labels > 1:
         slht_scr_dbs = silhouette_score(scaleddata, d_cluster.labels_)
         print("Silhouette Score:", slht_scr_dbs)
     else:
         print("Only one cluster is formed. Adjust DBSCAN parameters to form multiple clusters.")

     Silhouette Score: 0.5512063246025819
```

Nilai siluet dari DBSCAN adalah 0.5512063246025819

8. Komparasi nilai siluet dari K-Mean, Kluster Hirarki, dan DBSCAN

```
[86] print("The Silhouette score of KMeans Clustering:",slht_scr_km)
     print("The Silhouette score of Hierarchical Clustering:",slht_scr_hrc)
     print("The Silhouette score of DB Scan Clustering:",slht_scr_dbs)

     The Silhouette score of KMeans Clustering: 0.5630723128969357
     The Silhouette score of Hierarchical Clustering: 0.5025440766591031
     The Silhouette score of DB Scan Clustering: 0.5512063246025819
```

Terlihat bahwa nilai siluet paling tinggi adalah menggunakan K-Mean Clustering.

C. Kesimpulan

- Nilai siluet tertinggi DBSCAN ada pada eps=1.125, min_samples=5, dengan skor 0.5512063246025819
- Pada DBSCAN terdapat 3 cluster
- Data yang digunakan adalah data yang bersih dan sudah layak untuk dipakai
- Dari nilai siluet antar algoritma yang telah dijalankan, algoritma K-Means adalah algoritma dengan nilai siluet paling baik yaitu 0.5630723128969357

D. Daftar Pustaka

Colab:
https://colab.research.google.com/drive/1GxIdSv4YQl9rEvaPp2I5Vvy7IBQNY0zW?usp=sharing

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