MODUL 7
CLUSTERING

Clustering adalah seperangkat teknik yang digunakan untuk mempartisi data ke dalam kelompok, atau cluster. Ini sering digunakan sebagai teknik analisis data untuk menemukan pola menarik dalam data, seperti kelompok pelanggan berdasarkan perilaku mereka ataupun melakukan analisis sentiment pada suatu hal. Cluster didefinisikan sebagai kelompok objek data yang lebih mirip dengan objek lain di cluster mereka daripada objek data di cluster lain.
K-Means adalah salah satu metode clustering. Dengan python digunakan library scikit-learn yang diimplementasikan di sklearn.cluster.KMeans.

Penggunaan library scikit-learn untuk clustering

1	`# import tools`
	`import numpy as np`
	`import pandas as pd`
	`import matplotlib.pyplot as plt`
	`import seaborn as sns`
	`from sklearn.cluster import KMeans`

2	`# import data`
	`df = pd.read_csv(r'Mall_Customers.csv')`
	`df.head()`

3	`# amati` `df.shape`
4	`df.describe()`
5	`# cek null data` `df.isnull().sum()`
6	`# tingkatkan visualisasi data` `plt.style.use('fivethirtyeight')`

7	`# amati masing-masing fitur`
	`plt.figure(1 , figsize = (15 , 6))`
	`n = 0`
	`for x in ['Age' , 'Annual Income (k$)' , 'Spending Score (1-100)']:`
	`n += 1`
	`plt.subplot(1 , 3 , n)`
	`plt.subplots_adjust(hspace =0.5 , wspace = 0.5)`
	`sns.distplot(df[x] , bins = 20)`
	`plt.title('Distplot of {}'.format(x))`
	`plt.show()`

8	

	# Ploting untuk mencari relasi antara Age , Annual Income and Spending Score
	plt.figure(1 , figsize = (15 , 7))
	n = 0
	for x in ['Age' , 'Annual Income (k$)' , 'Spending Score (1-100)']:
	for y in ['Age' , 'Annual Income (k$)' , 'Spending Score (1-100)']:
	n += 1
	plt.subplot(3 , 3 , n)
	plt.subplots_adjust(hspace = 0.5 , wspace = 0.5)
	sns.regplot(x = x , y = y , data = df)
	plt.ylabel(y.split()[0]+' '+y.split()[1] if len(y.split()) > 1 else y)
	plt.show()

9	plt.show()

10	# plot Age dan Annual Income
	plt.figure(1 , figsize = (15 , 6))
	for gender in ['Male' , 'Female']:
	plt.scatter(x = 'Annual Income (k$)',y = 'Spending Score (1-100)' ,
	data = df[df['Gender'] == gender] ,s = 200 , alpha = 0.5 ,
	label = gender)
	plt.xlabel('Annual Income (k$)'), plt.ylabel('Spending Score (1-100)')
	plt.title('Annual Income vs Spending Score')
	plt.legend()
	plt.show()

11	# rancang K-Means untuk spending score vs annual income
	# Kmeans, menentukan jumlah kluster dengan elbow
	X1 = df[['Annual Income (k$)' , 'Spending Score (1-100)']].iloc[: ,
	:].values
	inertia = []
	for n in range(1 , 11):
	algorithm = (KMeans(n_clusters = n ,init='k-means++', n_init = 10
	,max_iter=300,
	random_state= 111))
	algorithm.fit(X1)
	inertia.append(algorithm.inertia_)

12	# plot elbow
	plt.figure(1 , figsize = (15 ,6))
	plt.plot(np.arange(1 , 11) , inertia , 'o')
	plt.plot(np.arange(1 , 11) , inertia , '-' , alpha = 0.5)
	plt.xlabel('Number of Clusters') , plt.ylabel('Inertia')
	plt.show()

13	# bangun K-Means
	algorithm = (KMeans(n_clusters = 5 ,init='k-means++', n_init = 10
	,max_iter=300,
	tol=0.0001, random_state= 111 , algorithm='elkan'))
	algorithm.fit(X1)
	labels2 = algorithm.labels_
	centroids2 = algorithm.cluster_centers_

14	# siapkan data untuk plot dan imshow
	labels2 = algorithm.labels

	centroids2 = algorithm.cluster_centers_ step = 0.02 x_min, x_max = X1[:, 0].min() - 1, X1[:, 0].max() + 1 y_min, y_max = X1[:, 1].min() - 1, X1[:, 1].max() + 1 xx, yy = np.meshgrid(np.arange(x_min, x_max, step), np.arange(y_min, y_max, step)) Z1 = algorithm.predict(np.c_[xx.ravel(), yy.ravel()]) # array diratakan 1D
15	plt.figure(1 , figsize = (15 , 7)) plt.clf() Z1 = Z1.reshape(xx.shape) plt.imshow(Z1 , interpolation='nearest', extent=(xx.min(), xx.max(), yy.min(), yy.max()), cmap = plt.cm.Pastel2, aspect = 'auto', origin='lower') plt.scatter(x = 'Annual Income (k$)' ,y = 'Spending Score (1-100)' , data = df , c = labels2 , s = 200) plt.scatter(x = centroids2[: , 0] , y = centroids2[: , 1] , s = 300 , c = 'red' , alpha = 0.5) plt.ylabel('Spending Score (1-100)') , plt.xlabel('Annual Income (k$)') plt.show()

16	# coba prediksi
	data = [[15, 39],[15, 20], [20, 80]]
	print(data)
	print(algorithm.predict(data))

Latihan

1. Cobalah semua source code diatas
2. Lihat hasil outputnya dan berikan penjelasan
3. Buatlah laporan yang berisi penjelasan dari source code tersebut, kumpulkan di Hebat dengan nama PrakClustering_NIM, maksimal pukul 11.00 (27-04-2023)

Tugas

1. Kerjakan tugas secara kelompok
2. Buatlah program clustering dengan data yang lain (data silahkan dicari sendiri)
3. Buatlah laporan dari program yang dibuat
4. File yang dikumpulkan data asli, program phyton, dan laporan
5. Penamaan Tugas TugasClustering_Kelompok XXX.Zip
6. Dikumpulkan paling lambat Senin / 1 Mei 2023 Pukul 23.59 Wib