Nama : Hafidh Fikri Rasyid

NIM : 1301142190

Kelas : IF-38-Gab01

Assignment 4

Machine Learning

Soal :

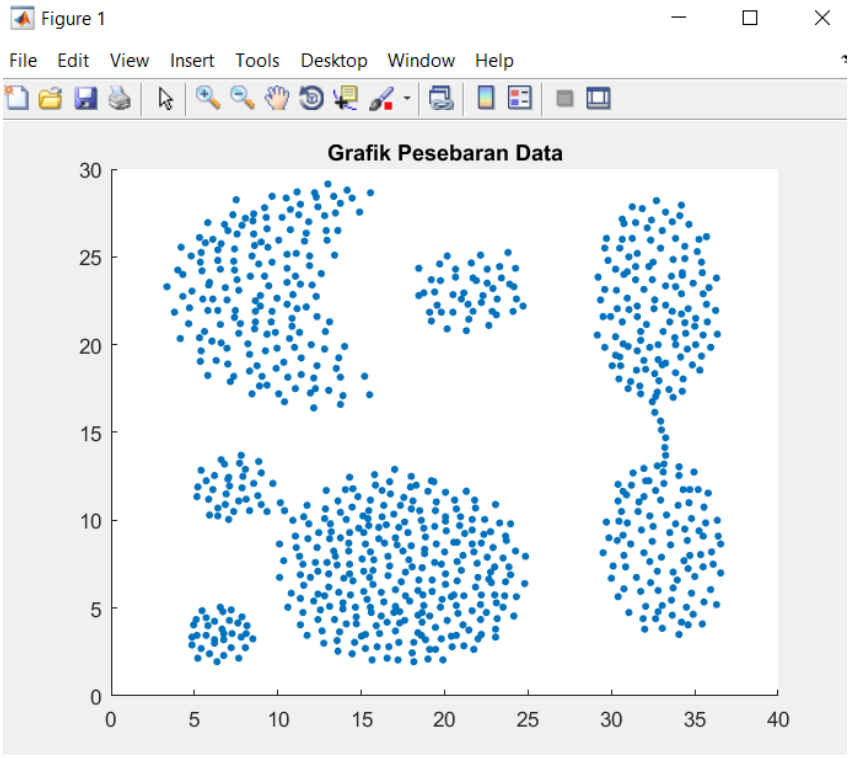
1. **(50 points)** In this exercise, we implement partitional clustering method: K-means algorithm.
   (a) Load the selected data set. Visualize all data points using scatter plot in one color (no need to give different color for each class). Use attribute 1 as x -axis, attribute 2 as y -axis.[**4 points**]
   (b) Apply K-means on the selected data set. Your codes have to clearly contain
      i. Function that takes as inputs: the data matrix and initial centroids, and as outputs: the final centroids and the cluster assignments specifying which data vectors are assigned to which centroids after convergence of the algorithm. (Use matrix operations wherever possible, avoiding explicit loops, to speed up the algorithm sufficiently for running the algorithm on the selected data). [**10 points**]
      ii. Function to calculate the objective function of K-means that is Sum of Squared Errors (SSE). It takes as inputs: the data matrix and final centroids that is resulted from learning. [**6 points**]
   (c) Run your K-means algorithm, using K equals to the number of classes in data set, with the initial centroids taken from randomly selected K data points. After convergence, visualize the centroid of each cluster as well as all data points assigned to that cluster (it should be easily distinguished between the centroids and the data points, also give different colors to different clusters). One may run the algorithm several times in order to obtain the best result (hints: use the SSE as the measure). [**7 points**]
   (d) Load the selected data set. Visualize all data points using scatter plot. Use attribute 1 as x -axis, attribute 2 as y -axis. Use different color and/or different symbol for each class label. The class labels are actually not used in the clustering algorithm, however we use this visualization to get a view/thought of the clustering results. [**4 points**]
   (e) Based on visualization resulted from point 1(c), to what extent do the K clusters correspond to the K different classes? (Hints: Use the visualization from point 1(d) to get a view of clustering results shown by point 1(c).) [**4 points**]
   (f) Re-run K-means but selecting randomly one instance of each class as the initial centroids (so that the initial centroids all represent distinct class). After convergence, visualize the centroid of each cluster as well as all data points assigned to that cluster (it should be easily distinguished between the centroids and the data points, also give different colors to different clusters). One may run the algorithm several times in order to obtain the best result (hints: use the SSE as the measure). [**7 points**]
   (g) Based on visualization resulted from point 1(e), to what extent do the K clusters correspond to the K different classes? (Hints: Use the visualization from point 1(d) to get a view of clustering results shown by point 1(e).) [**4 points**]
   (h) By visually comparing figures created from point 1(c) and 1(e), what do you think of the clustering results? Give explanations. [**4 points**]

Keterangan :

1. Untuk menjalankan program, jalankan pada file KMeans.m

Jawaban :

a. Data tersebut di plot sebagai berikut

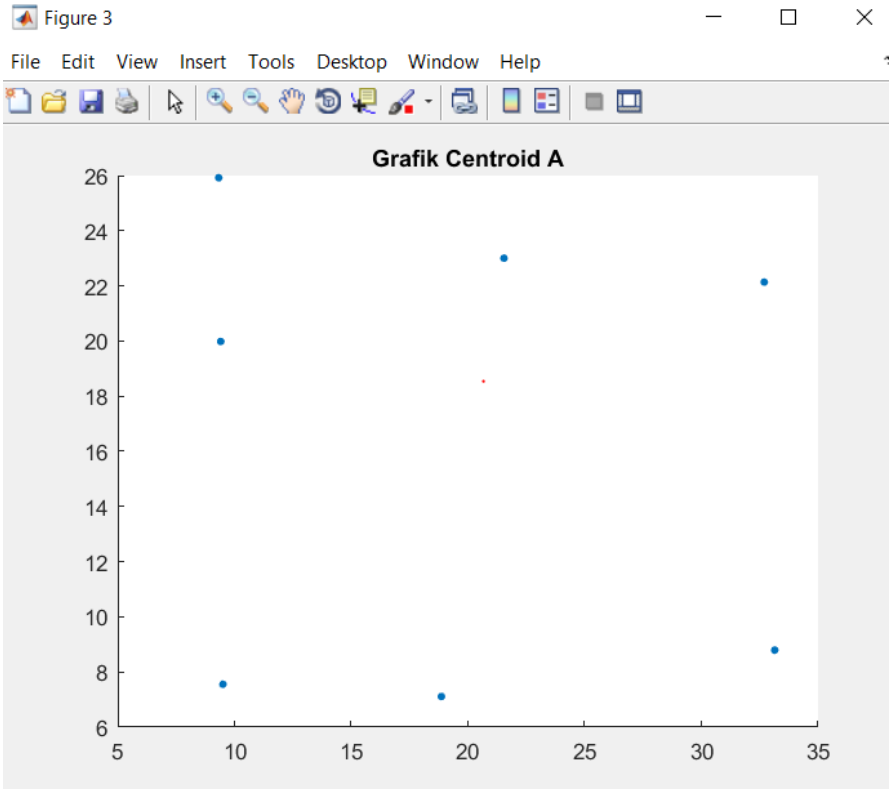



b.

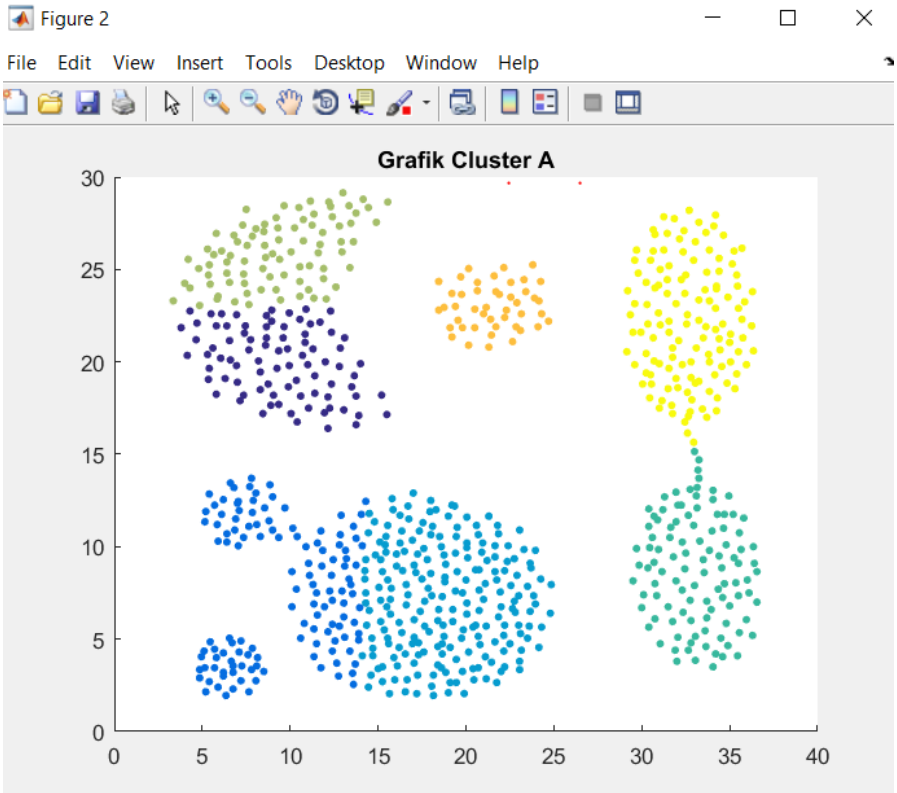
i. fungsi dari perhitungan k means terdapat pada file cariKMeans.m

ii. Fungsi dari perhitungan Sum Square Error terdapat pada file cariSSE.m

c. Aplikasi berhasil dijalankan dengan memilih titik secara random untuk centroid awal. Aplikasi ini berhasil menghasilkan beberapa titik centroid sebagai berikut

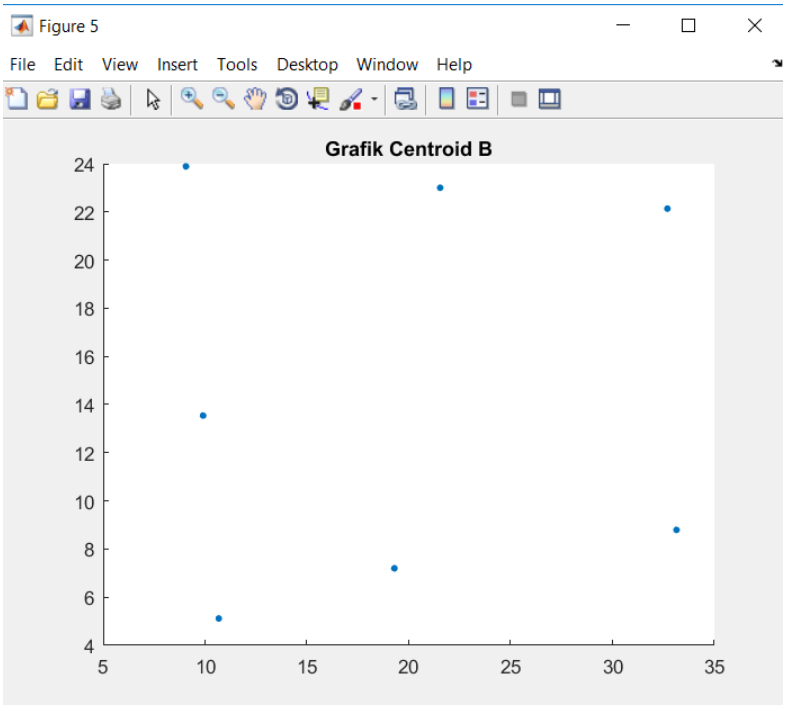



d. Grafik clustering yang dihasilkan dari titik centroid yang ada diatas adalah sebagai berikut :

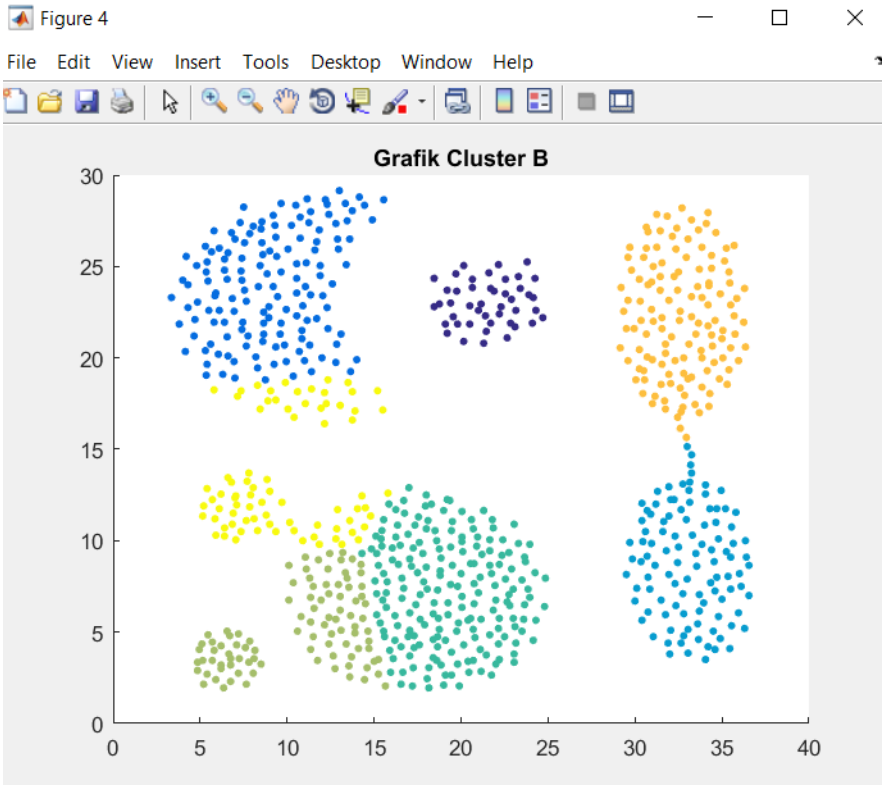



e. Jadi, titik centroid yang ada pada poin c dapat memberikan kelompok/clustering ke kelompok data yang ditampilkan pada poin d. Tetapi, pengelompokan belum terbentuk secara sempurna.

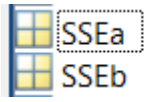
f. Aplikasi berhasil dijalankan dengan memilih titik secara random berdasarkan kelompok data yang ada pada dataset untuk centroid awal. Aplikasi ini berhasil menghasilkan beberapa titik centroid sebagai berikut




g. Grafik clustering yang dihasilkan dari titik centroid yang ada diatas adalah sebagai berikut :




h. Jadi, titik centroid yang ada pada poin f dapat memberikan kelompok/clustering ke kelompok data yang ditampilkan pada poin g. Tetapi, pengelompokan belum terbentuk secara sempurna. Dibandingkan dengan centroid yang digenerate pada soal c, centroid yang digenerate pada soal f lebih baik karena memiliki nilai Sum Square Error yang lebih kecil.



| | |
|---|---|
| SSEa | 2.7546e+03 |
| SSEb | 2.7624e+03 |