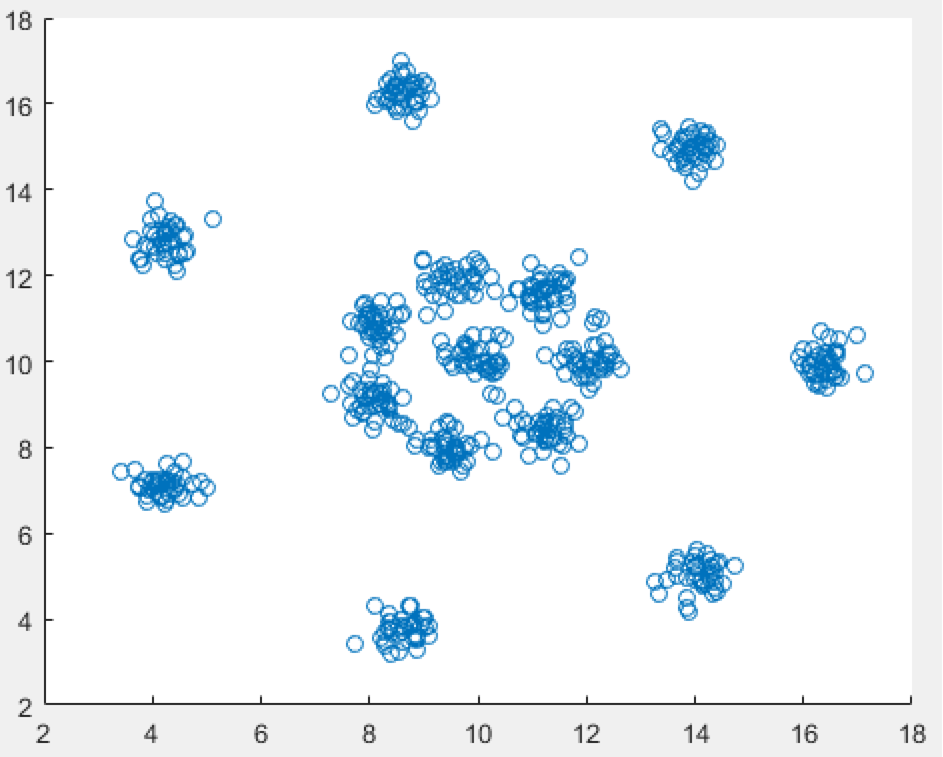
ALI HELMUT
1301154246

(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

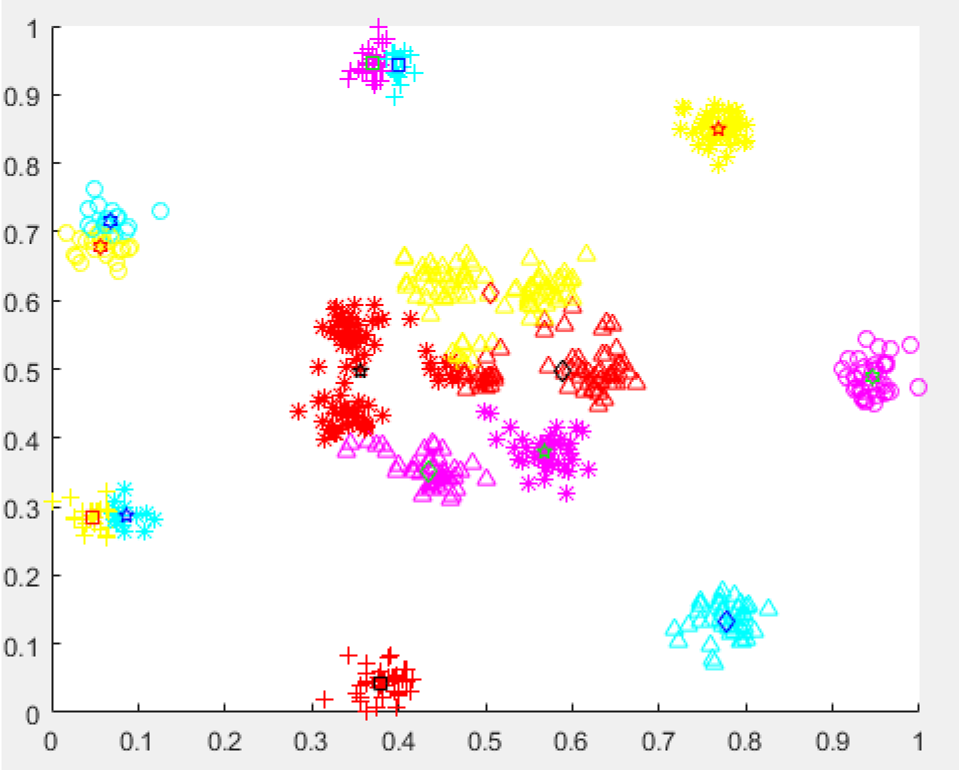
I. FUNCTION THAT TAKES AS INPUTS: THE DATA MATRIX AND INITIAL CENTROIDS, AND AS OUTPUTS: THE fiNAL CENTROIDS AND THE CLUSTER ASSIGNMENTS SPECIFYING WHICH DATA VECTORS ARE ASSIGNED TO WHICH CENTROIDS AFTER CONVERGENCE OF THE ALGORITHM. (USE MATRIX OPERATIONS WHEREVER POSSIBLE, AVOIDING EXPLICIT LOOPS, TO SPEED UP THE ALGORITHM SUffiCIENTLY FOR RUNNING THE ALGORITHM ON THE SELECTED DATA). [10 POINTS]

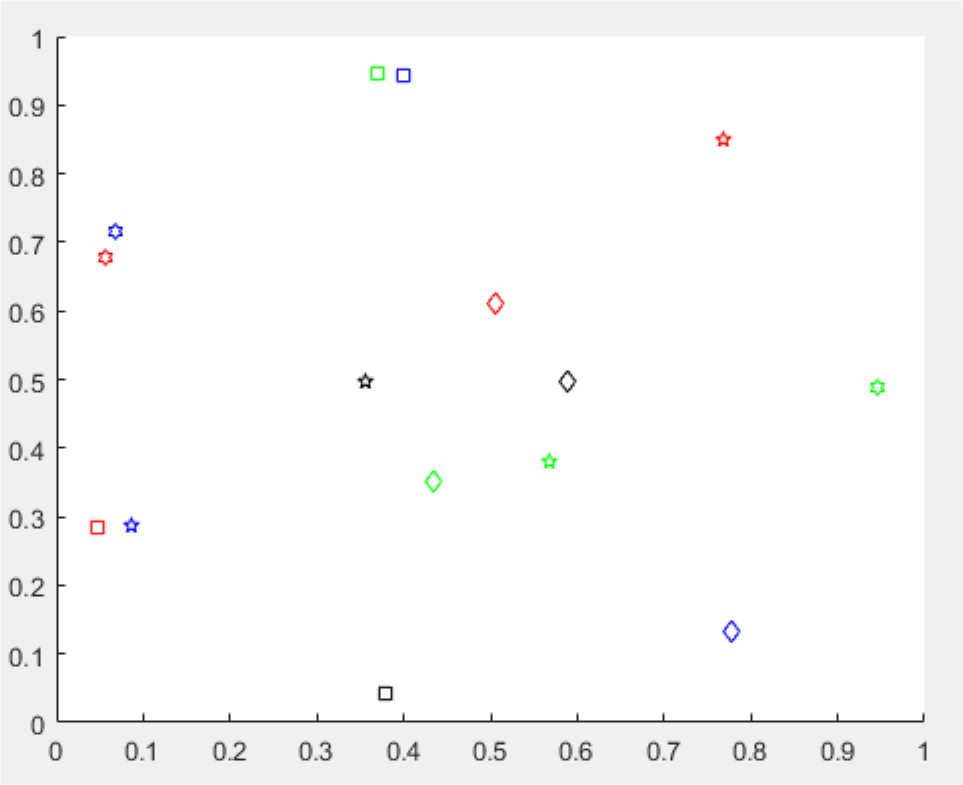
```
function [ x y ] = clustering( dataInput, centroid  )
dataOutput = [];
for i=1 : length(dataInput)
    jarak = [];
    for j=1 : length(centroid)
        jarak = [jarak; norm(dataInput(i,:)-centroid(j,:)) j ];
    end
    jarak = sortrows(jarak,1);
    dataOutput = [ dataOutput; dataInput(i,1) dataInput(i,2) jarak(1,2)];
end

centroidOutput = [];
for i=1 : 15
    data = [];
    a = 0;
    for j=1 : length(dataOutput)
        if(dataOutput(j,3) == i)
            data = [data;  dataOutput(j,1) dataOutput(j,2)];
            a = 1;
        end
    end
    if(a == 1)
        centroidOutput = [centroidOutput; mean(data)];
    else
        centroidOutput = [centroidOutput; rand(1,1) rand(1,1)];
    end
end
x = dataOutput;
y = centroidOutput;
end
```

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]

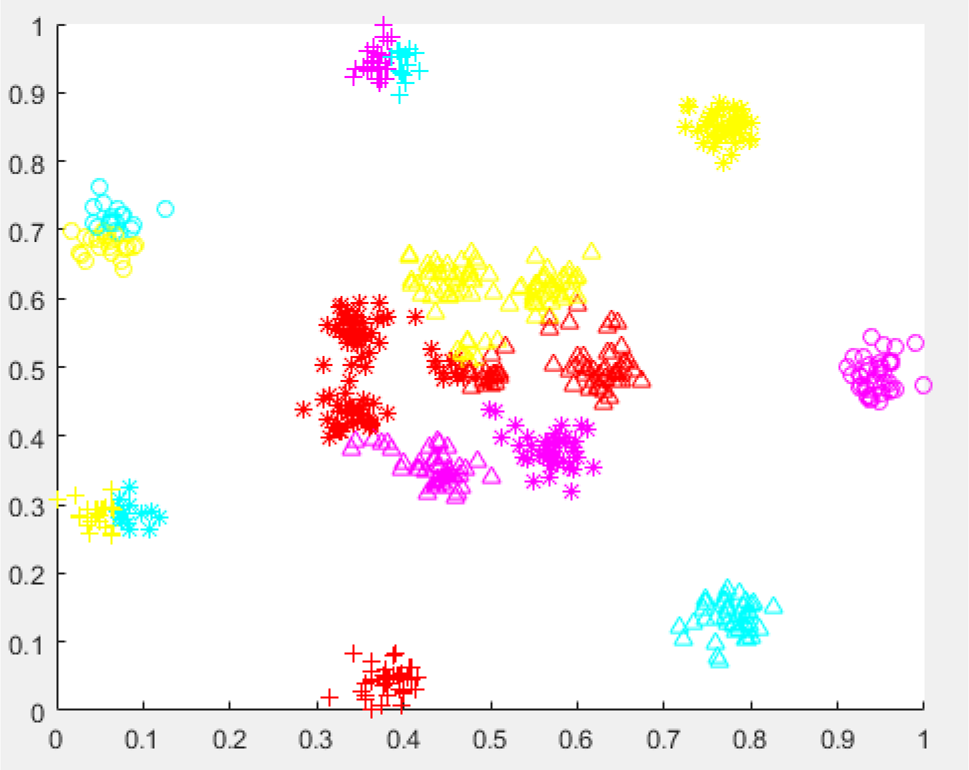
```
function [ output_args ] = getSSE( inp ,centroid )
    a = 0;
    for i=1 : length(inp)
        a = norm((inp(i,1:2))-centroid(inp(i,3)));
    end

output_args = a;
end
```

(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]




1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

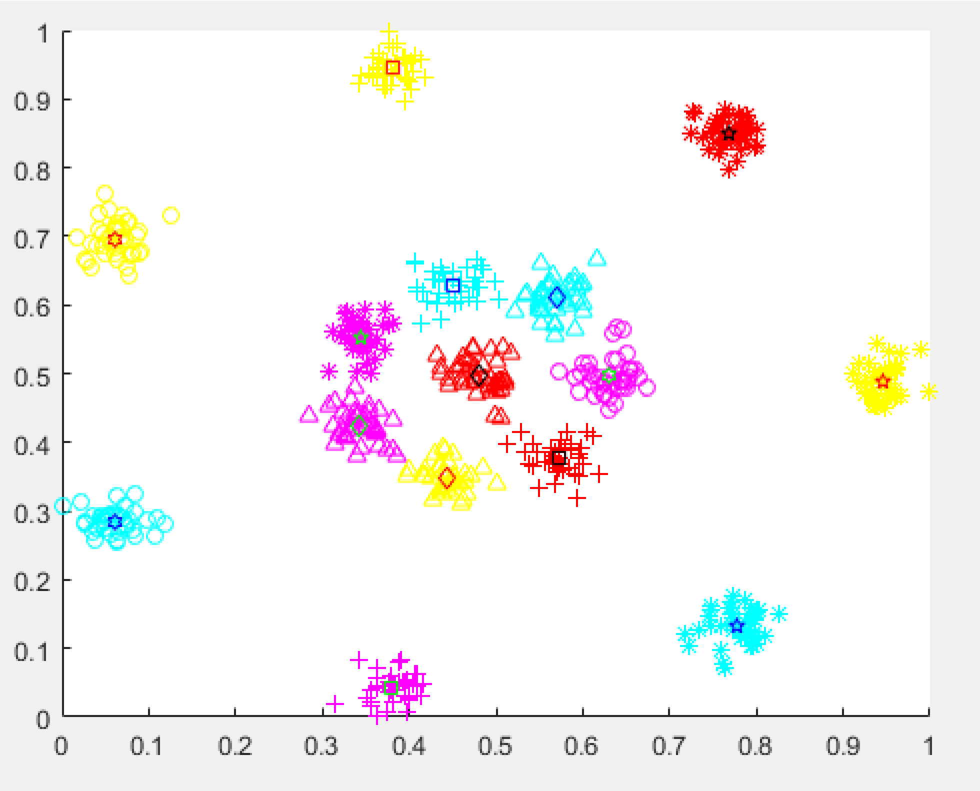
(D) LOAD THE SELECTED DATA SET. VISUALIZE ALL DATA POINTS USING SCATTER PLOT. USE ATTRIBUTE 1 AS X -AXIS, ATTRIBUTE 2 AS Y -AXIS. USE DIFFERENT COLOR AND/OR DIFFERENT SYMBOL FOR EACH CLASS LABEL. THE CLASS LABELS ARE ACTUALLY NOT USED IN THE CLUSTERING ALGORITHM, HOWEVER WE USE THIS VISUALIZATION TO 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]

OBJEK SELALU DIKLASIFIKASI TERHADAP CENTROID TERDEKAT, NAMUN DIKARENAKAN CENTROID DIAMBIL SECARA ACAK, MAKA TERDAPAT CLUSTER YANG BERDEKATAN, DAN TERDAPAT OBJEK YANG BERJAUHAN TAPI DALAM CLUSTER YANG SAMA.

(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]

OBJEK SELALU DIKLASIFIKASI TERHADAP CENTROID TERDEKAT, DAN CENTROID TEPAT PADA PUSAT PERKUMPULAN OBJEK YANG SANGAT BERDEKATAN. INI DIKARENAKAN CENTORID DIAMBIL SECARA RANDOM TIAP TIAP CLASS, SEHINGGA CLUSTER AWAL BERADA DEKAT DENGAN PUSAT MASA OBJEK YANG SANGAT BERDEKATAN

(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]

HASIL CLUSTERING PERCOBAA F LEBIH BAIK DIBANDING HASIL CLUSTERING PERCOBAAN C, DIKARENAKAN CENTROID PADA F MENGGUNAKAN DATA TIAP TIAP KELAS SEBAGAI CENDROID AWAL, SEDANGKAN PADA PERCOBAAN C CENTROID MENGAMBIL DATA SECARA BENAR BENAR ACAK SEBAGAI CENTROID AWAL, SEHINGGA TIDAK MENUTUP KEMUNGKINAN TERDAPAT LEBIH DARI DUA CENTROID AWAL DARI KELAS YANG SAMA SEHINGGA SANGAT MEMUNGKINKAN KESALAHAN CLUSTERING/ CLUSTERING YANG TIDAK OPTIMAL.