Lab - 8

CSL2010: Introduction To Machine Learning
AY 2021-22

Task 1

K Means

(i)-downloaded the milk.csv file and read it using pandas

(ii)-imported the Kmeans from sklearn library sklearn.cluster.

Then performed the K Means using n=2,3 and 4.stored the predicted labels(the cluster to which a particular point belongs) in 3 different variables.

(iii)-Printed the cluster_labels for the three cases.

(a)n=2

```
[[62.6625     9.7       22.675      2.3       1.27625  ]
 [85.48823529  4.57058824  4.48823529  4.99411765  0.66882353]]
```

(b)n=3

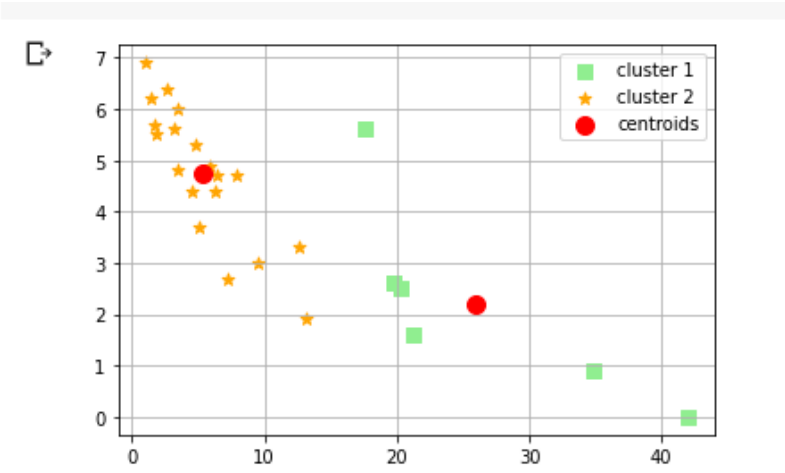
```
[[86.0625     4.275      4.175      5.11875    0.635625 ]
 [69.47142857  9.51428571 16.28571429  2.92857143  1.43285714]
 [45.65       10.15      38.45       0.45       0.69     ]]
```

(c)n=4

```
[[68.33333333  9.55       17.41666667  2.91666667  1.47166667]
 [81.18571429  7.42857143  6.9         4.01428571  0.93142857]
 [88.5         2.57       2.8         5.68        0.485    ]
 [45.65       10.15      38.45       0.45        0.69     ]]
```

These matrices shows the cluster centres obtained for different values of n.

Each contains 5 columns.These five columns show the cluster centre for the five features



Collab link--

https://colab.research.google.com/drive/1BFYUi5AmMBG_jxaw1SVqQpQwGV2qOvrF?usp=sharing

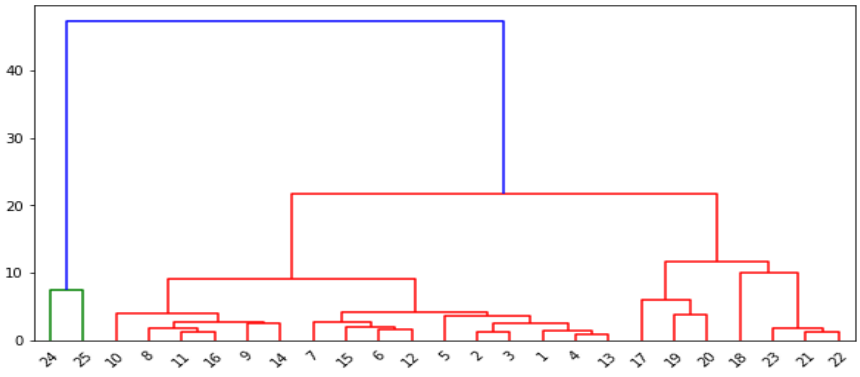
Task 2

Hierarchical Clustering and K Means [w/o inbuilt function]

I.

i.

```
)
plt.show()
```

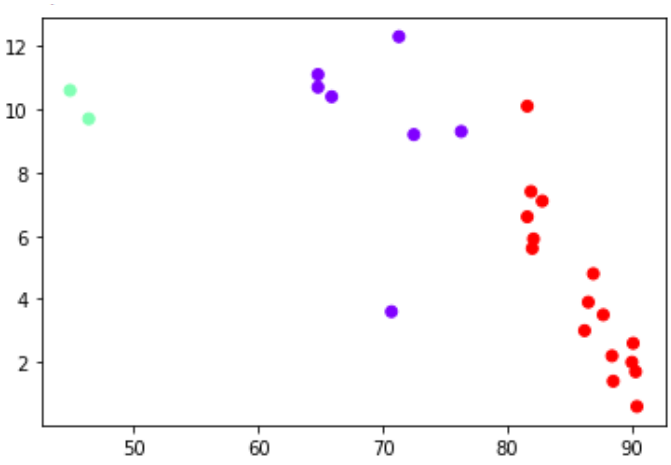



```
241 from sklearn.cluster import AgglomerativeClustering
```

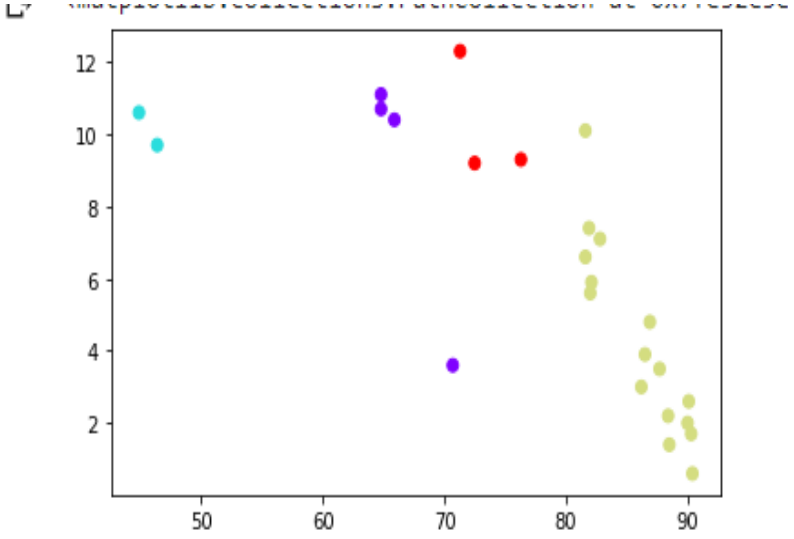
ii. performed agglomerative clustering using n=3,n=4,n=6

iii.

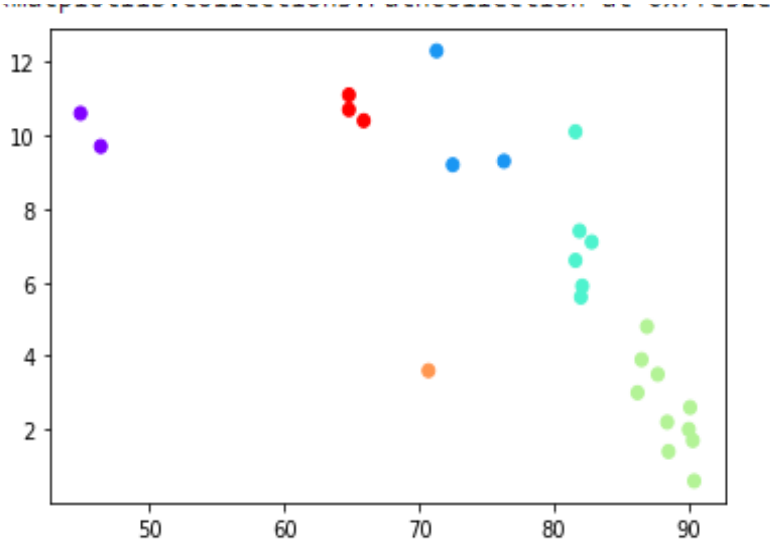
n=3



n=4



n=6



III. Kmeans From Scratch

i. Chosen k=3.

ii. I picked up 3 points randomly from the dataset itself and stored these points as a nested list in a single list ‘centers’

center_1 = [90.1,2.6,1,6.9,0.35]

center_2 = [81.6,10.1,6.3,4.4,0.75]

center_3 = [64.8,11.1,21.2,1.6,1.70]

centers=[center_1,center_2,center_3]

iii.Now , I made another list that contains three lists each for cluster 1,cluster2 and cluster3.For assigning clusters to all the points I took the points one by one and calculated their distances from the centroid points that I initialised in ‘centers’ list. Now for each point there are 3 distances which are stored in another list such that the first value in the list contains distance from the first centroid and so on.
Now, I grouped these points into three clusters in the ‘clusters’ list by taking the minimum distance from the distance list and using its index to assign the cluster to each point.

distance=[](for every point the index value of minimum distance is found and that is used to assign the point to a respective cluster in clusters list.

cluster_1=[]

cluster_2=[]

```
cluster_3=[]
clusters=[cluster_1,cluster_2,cluster_3]
```

iv. The mean of each cluster is calculated and stored in another list.

```
mean=[[],[],[]].
```

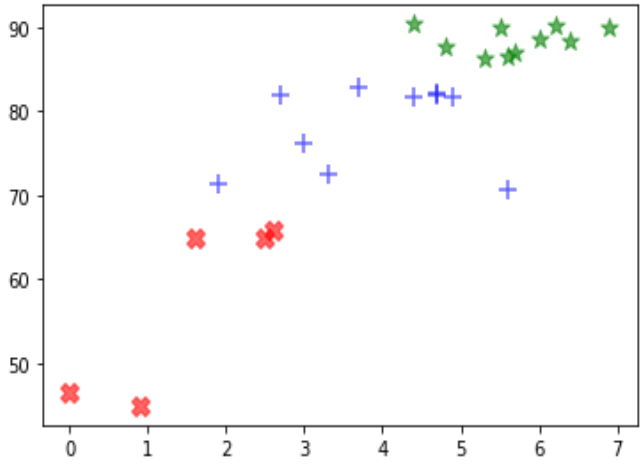
After calculating the mean for each cluster the centres are changed by the new mean values.

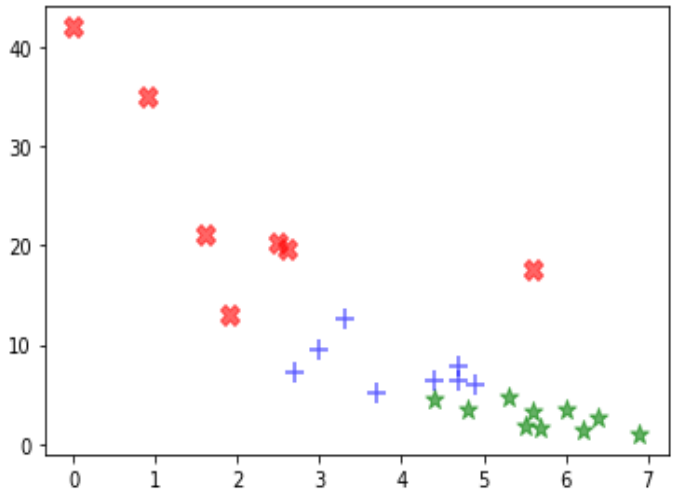
v.Finally for observing the convergence a list stores difference between the previous and new centres(the centers list and mean list).When it becomes 0 means Kmeans has converged.In this case it converged in 4 iterations.

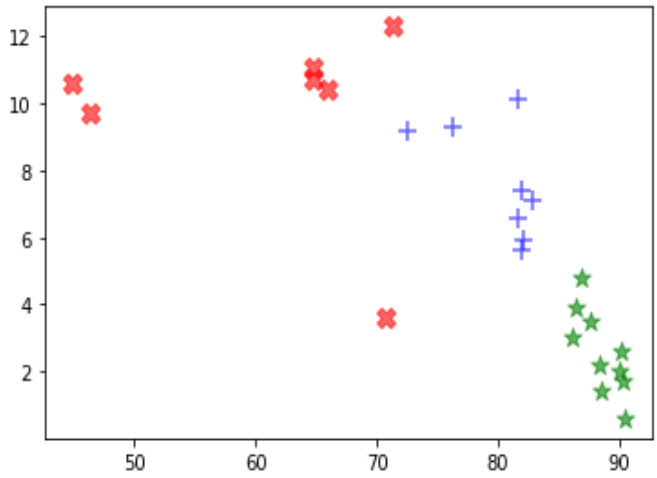
Final clusters-

```
[[array([90.1 ,  2.6 ,  1.  ,  6.9 ,  0.35]),
  array([88.5 ,  1.4 ,  3.5 ,  6.  ,  0.24]),
  array([88.4 ,  2.2 ,  2.7 ,  6.4 ,  0.18]),
  array([90.3,  1.7,  1.4,  6.2,  0.4]),
  array([90.4,  0.6,  4.5,  4.4,  0.1]),
  array([87.7 ,  3.5 ,  3.4 ,  4.8 ,  0.71]),
  array([86.9,  4.8,  1.7,  5.7,  0.9]),
  array([86.5,  3.9,  3.2,  5.6,  0.8]),
  array([90.  ,  2.  ,  1.8 ,  5.5 ,  0.47]),
  array([86.2,  3. ,  4.8,  5.3,  0.7])],
 [array([82.1 ,  5.9 ,  7.9 ,  4.7 ,  0.78]),
  array([81.9 ,  7.4 ,  7.2 ,  2.7 ,  0.85]),
  array([81.6 , 10.1 ,  6.3 ,  4.4 ,  0.75]),
  array([81.6 ,  6.6 ,  5.9 ,  4.9 ,  0.93]),
  array([82.8,  7.1,  5.1,  3.7,  1.1]),
  array([82.  ,  5.6 ,  6.4 ,  4.7 ,  0.91]),
  array([76.3,  9.3,  9.5,  3. ,  1.2]),
  array([72.5,  9.2, 12.6,  3.3,  1.4])],
 [array([70.7 ,  3.6 , 17.6 ,  5.6 ,  0.63]),
  array([71.3, 12.3, 13.1,  1.9,  2.3]),
  array([65.9, 10.4, 19.7,  2.6,  1.4]),
  array([64.8, 10.7, 20.3,  2.5,  1.4]),
  array([64.8, 11.1, 21.2,  1.6,  1.7]),
  array([46.4 ,  9.7 , 42.  ,  0.  ,  0.85]),
  array([44.9 , 10.6 , 34.9 ,  0.9 ,  0.53])]]
```

vi. Scatter graph taking different features


90
80
70
60
50
0
1
2
3
4
5
6
7


40
30
20
10
0
0
1
2
3
4
5
6
7



Collab link--
https://colab.research.google.com/drive/1qUAnD6_hzH-cKEYp1BqS04cKlSBnT9w8?usp=sharing