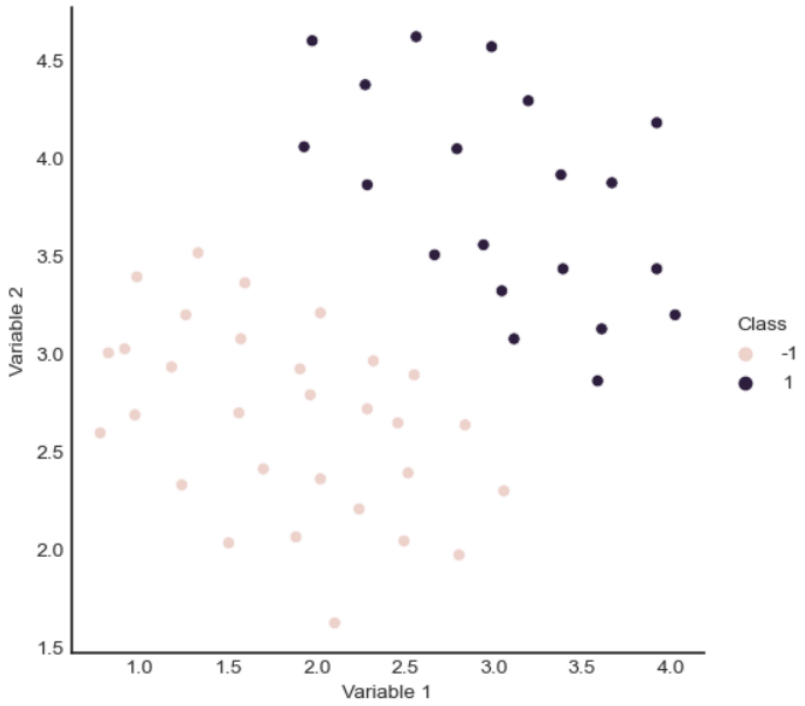
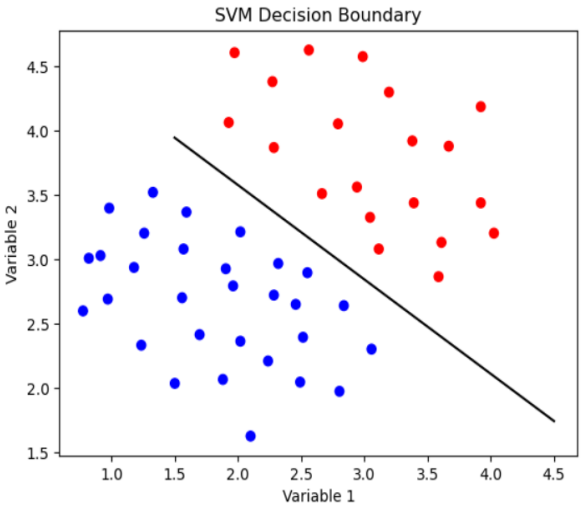
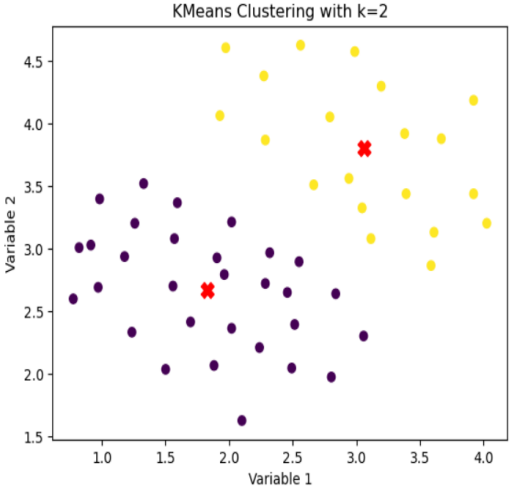
Question SVM

1.Upon visualizing the data in the file "Data 1.xlsx" using Seaborn, it appears that the two sets of features can be separated linearly using Support Vector Machines (SVM). As there is no overlap between the two sets of features, it is not necessary to apply regularization, which means that a value of C=0 can be used.

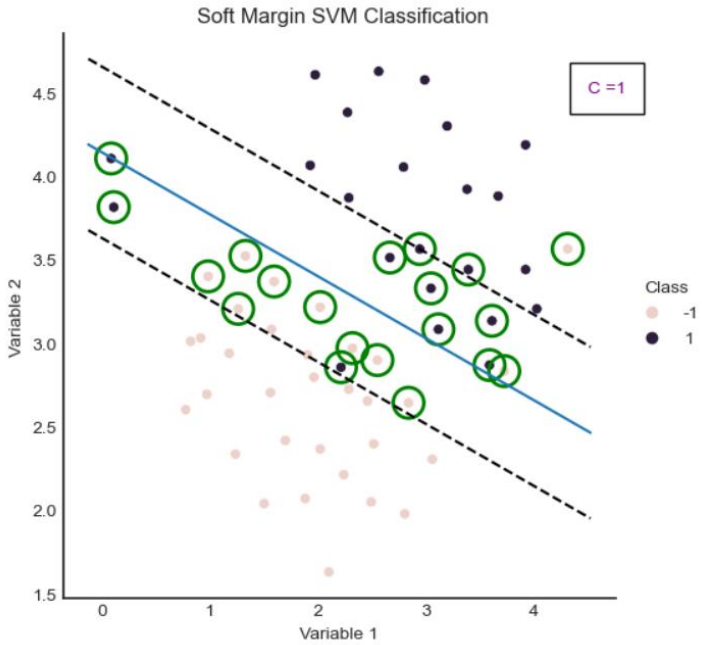


2. The Support Vector Machine (SVM) is trained using the CVXOPT library in Python, which is a quadratic optimizer designed for complex sets. After training and classification using SVM, it is observed that out of 50 sample feature points, only 3 serve as support vectors. Additionally, the decision function parameters can be computed and used to create a visual representation of the SVM. KMeans is faster than SVM.



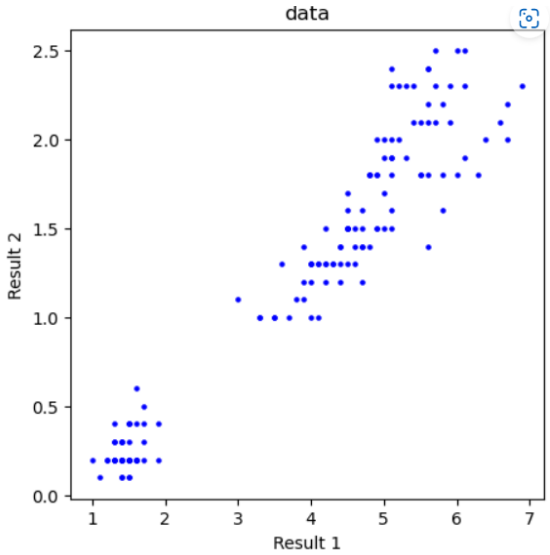


3. To address the issue of feature overlapping in the samples points of Data2.xslsx, it is recommended to train a SVM with a soft margin and set the regularization parameter C accordingly. A high value of C will lead to smaller margins and fewer support vectors, while a low value of C will result in larger margins and a higher number of support vectors. Upon analyzing the data set, it can be observed that setting C=0.1 yields 21 support vectors, whereas C=1 results in 14 support vectors.



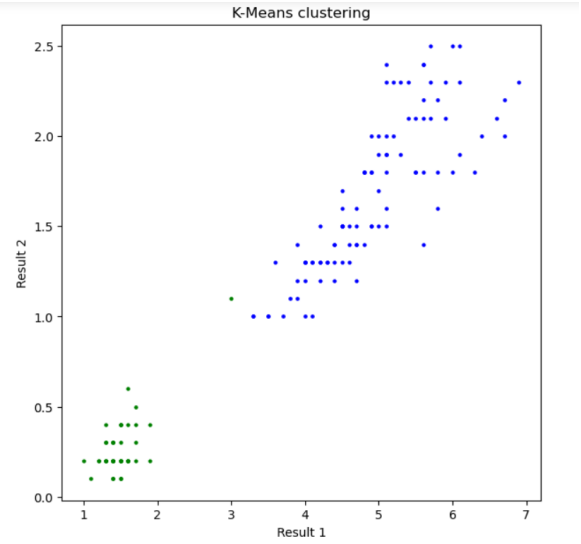
Question: GMM

1. Plotted the Data.xlsx data to get an idea of the data distribution. Roughly one can say it is two cluster.

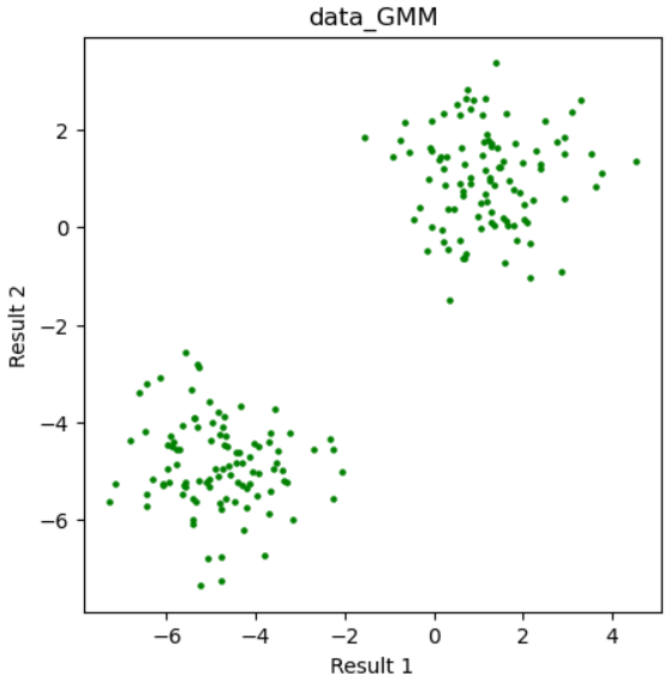


2. Applied K-Means clustering on the data.xlsx to find out the 2 clusters. Basic algo of this-

Select the number of clusters (K) you want to form. Randomly initialize K points as cluster centers. Assign each data point to the nearest cluster center. This forms K clusters. Calculate the new cluster centers by taking the average of all the data points assigned to each cluster. Repeat steps 3 and 4 until convergence is achieved (i.e. the cluster centers no longer move significantly).



3. Plotted the data (Data_GMM.xlsx) to get an idea of the data distribution. Clearly one can observe it show two clusters.



4. Applied Gaussian Mixture Model on the Data_GMM.xlsx to find out the 2 clusters.

Rough algo of this method-

Initialize the means, covariances, and mixing coefficients for each cluster.Repeat until convergence:

a. Evaluate the responsibilities of each data point for each cluster.

b. Update the means, covariances, and mixing coefficients for each cluster.

Output the final means, covariances, and mixing coefficients for each cluster.The initialization of the parameters affect the performance and stability of the algorithm, and multiple runs with different initializations needed to obtain a good solution. Convergence is determined by monitoring the change in log-likelihood of the data under the current parameter estimates.