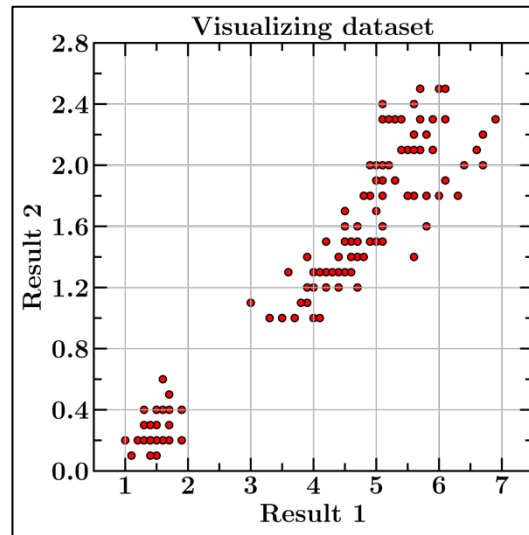1. Plot the data (Data.xlsx) to get an idea of the data distribution. Plot Result 1 on x-axis and Result 2 on y-axis. Report your visual observations.
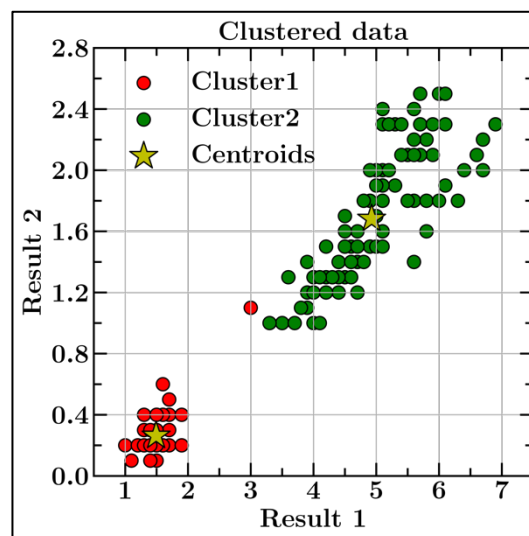
Solution

Observations: The data is distinctly classified into two clusters. One cluster represents vehicles with low emissions and the other cluster represents vehicles with higher emissions of pollutants. K-means algorithm can be used to classify points in different classes.

2. Apply K-Means clustering on the data to find out the two clusters. Make appropriate plots.
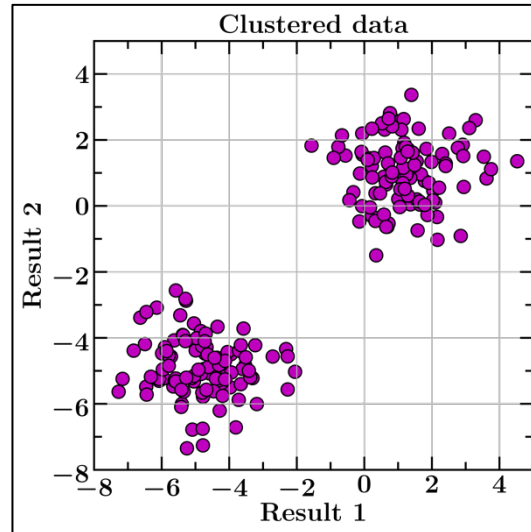
Solution

Centroid of Cluster 1 : [1.492, 0.263]

Centroid of Cluster 2 : [4.925, 1.682]

3. Plot the data (Data_GMM.xlsx) to get an idea of the data distribution. Plot Result 1 on x-axis and Result 2 on y-axis. Report your visual observations.
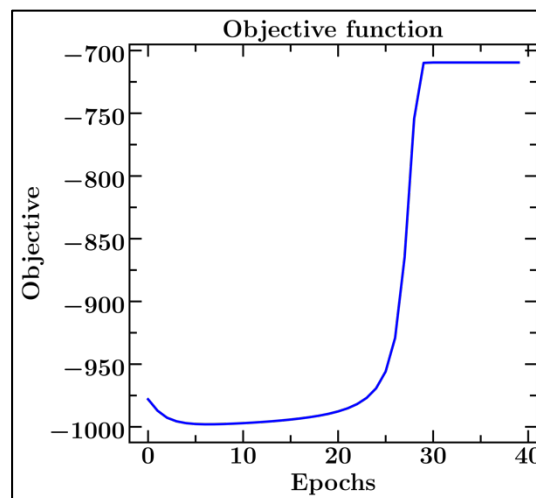
Solution

Observations: The data is distinctly classified into two clusters. K-means and gaussian process seems to be reasonable choices for doing the classification. In case we need to make a classifier, linear classifier or support vector classifier are suitable candidates.

4. Apply Gaussian Mixture Model on the Data_GMM.xlsx to find out the 2 clusters. Make appropriate plots.

Solution
The plot of objective functions vs. epochs shows that the procedure has converged.

The clustering of points along with colourmap in the background of clusters shows the ability of gaussian distributions which have accurately captured the underlying points.
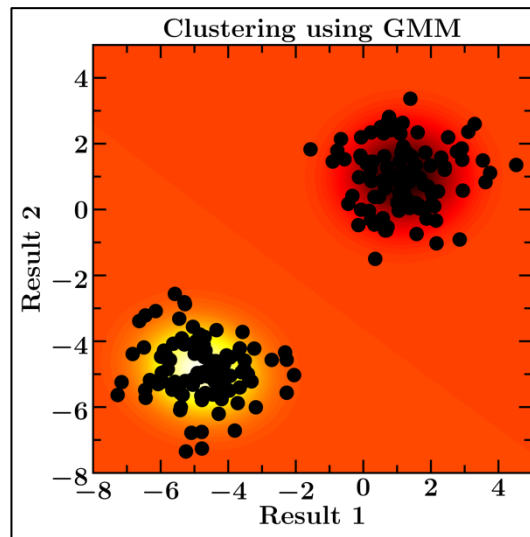
5. Compare the two methods.
   Solution

   Centroid 1 (K-Means) = [-4.818, -4.872]
   Centroid 1 (GMM)   = [-4.818, -4.872]

   Centroid 2 (K-Means) = [1.197, 1.009]
   Centroid 2 (GMM)   = [1.197, 1.009]

   Covariance   = [1.192, -0.105
                            -0.105, 0.800]

   a. The results of both the algorithms are in good agreement with each other.
   b. The K-Means algorithm is faster as compared to GMM.
   c. Gaussian processes are notorious because of increase in complexity of mean and covariance calculation with the increase in number of data samples.