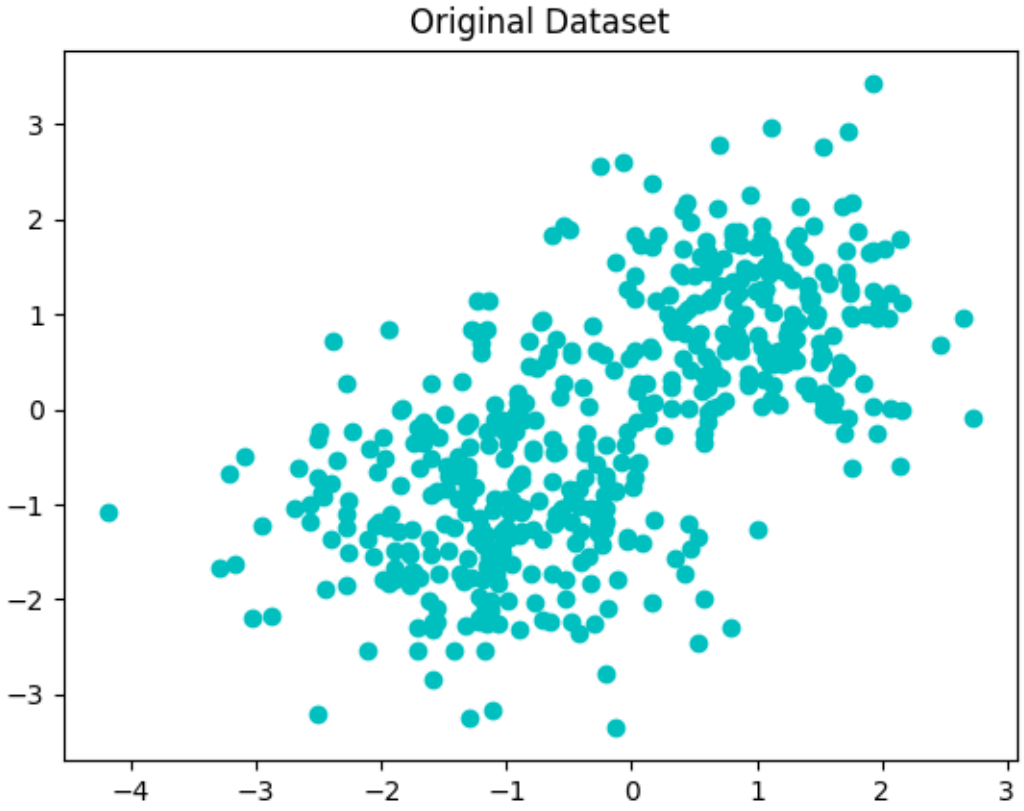
Clustering Project

The aim of this project is to perform clustering techniques on an artificial dataset and then use them on a real problem.

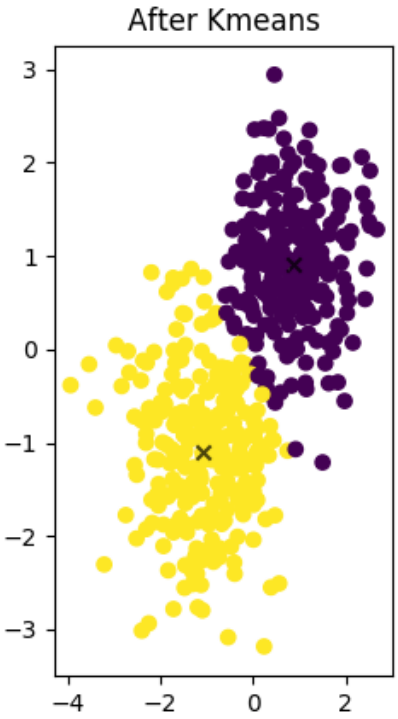
Theoretical part:

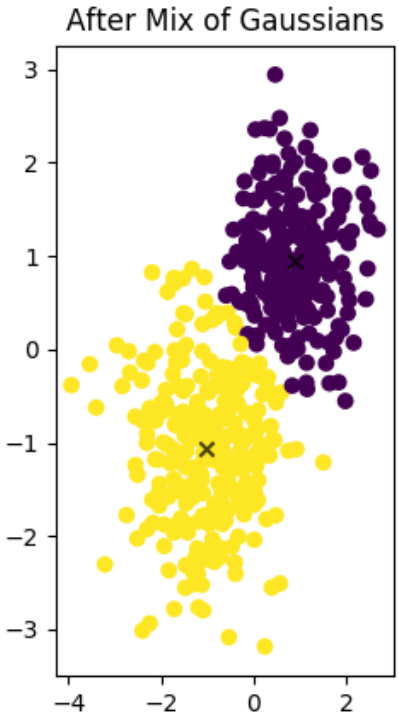
The artificial data was composed as a mix of two random Gaussians, one with mean [-1 -1] and one with mean [1 1]. The covariance matrices were diagonal scaled by 0.5 and 0.75 respectively.



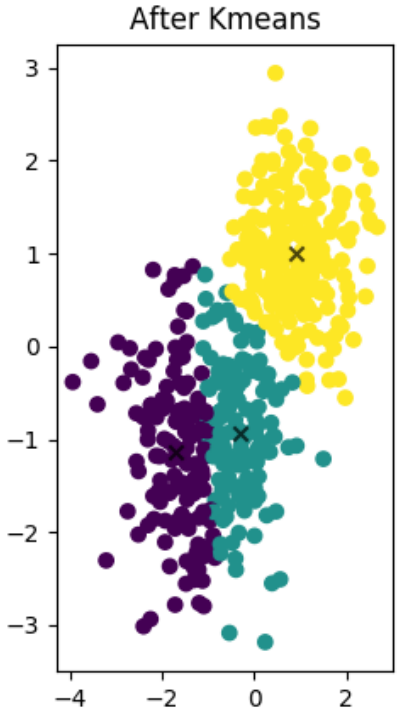
Two algorithms were performed on this data set: K-means and Mixtures of Gaussians. This was done for different initial k clusters, specifically for 2, 3, 4 and 5 clusters. And so we obtain the following picture:

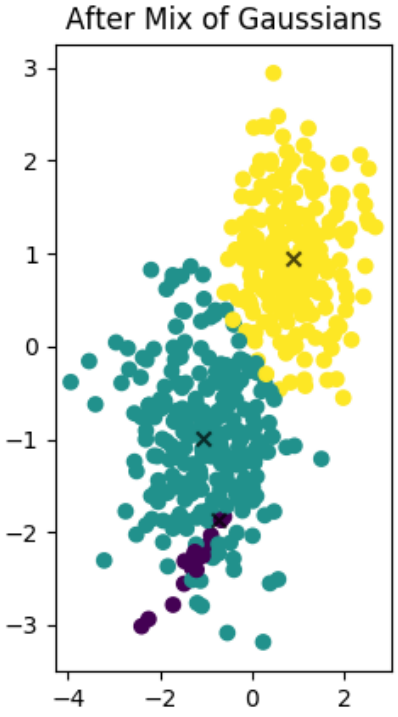
For 2 clusters:



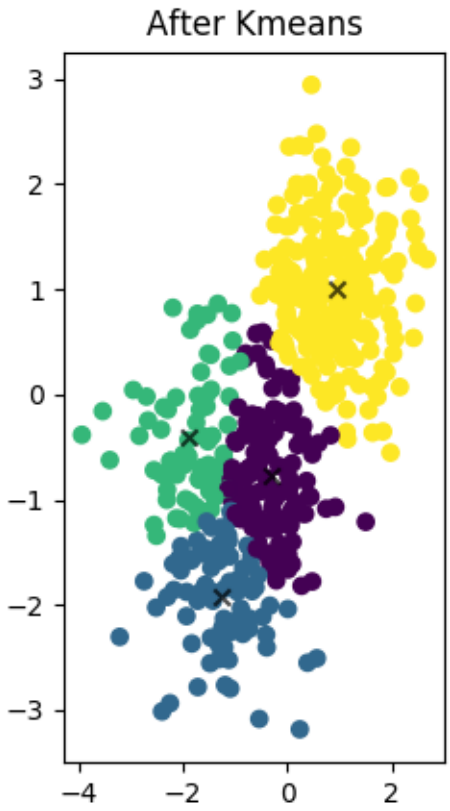


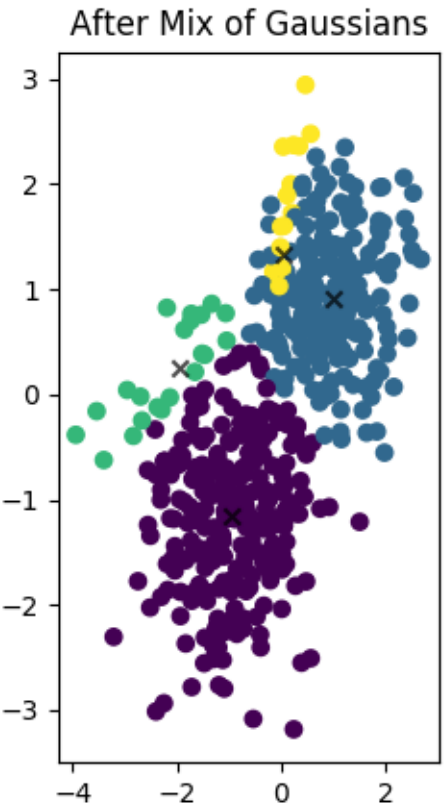
For 3 clusters:



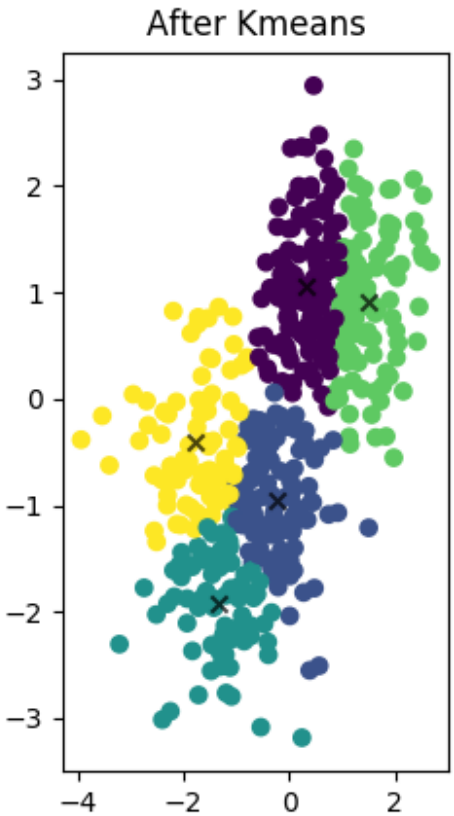


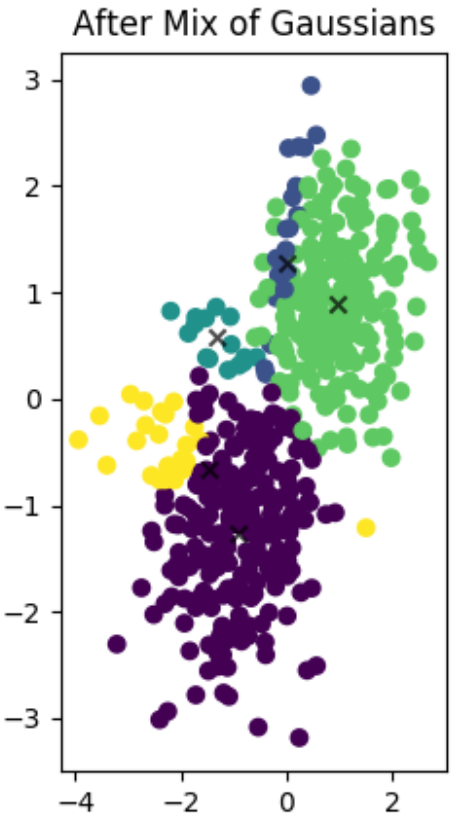
For 4 clusters:





And finally for 5 clusters:



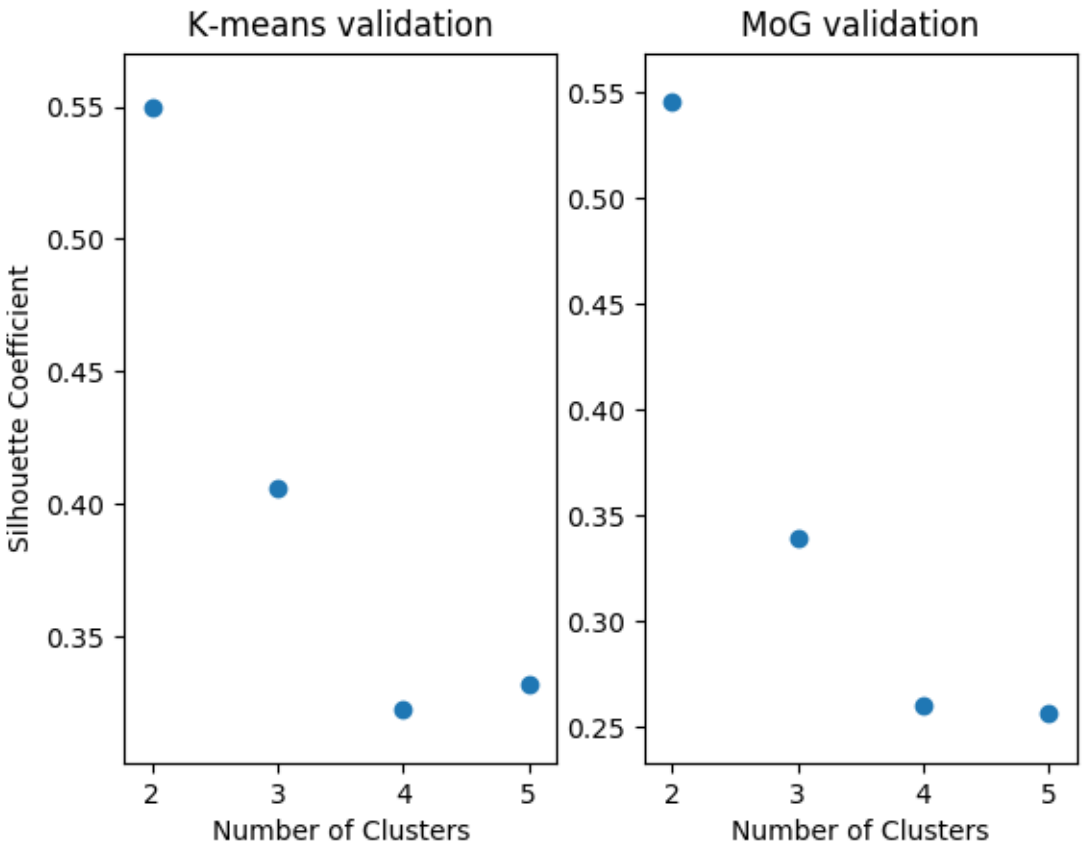


The correct number of clusters should be 2 obviously, because the data is a mix of two distributions. Both algorithms work well under this condition, producing the same results of 2 clusters. Beyond this number though, it becomes apparent that K-means is extremely sensitive to original conditions as it splits each distribution to different clusters of same proportions, whereas Mixtures of Gaussians tries to minimize the range of clusters from 3 and above. For the Mixtures algorithm, it is obvious that the extra clusters are redundant.

To further validate these results, the Silhouette coefficient was used. Which is best described by the following formula:

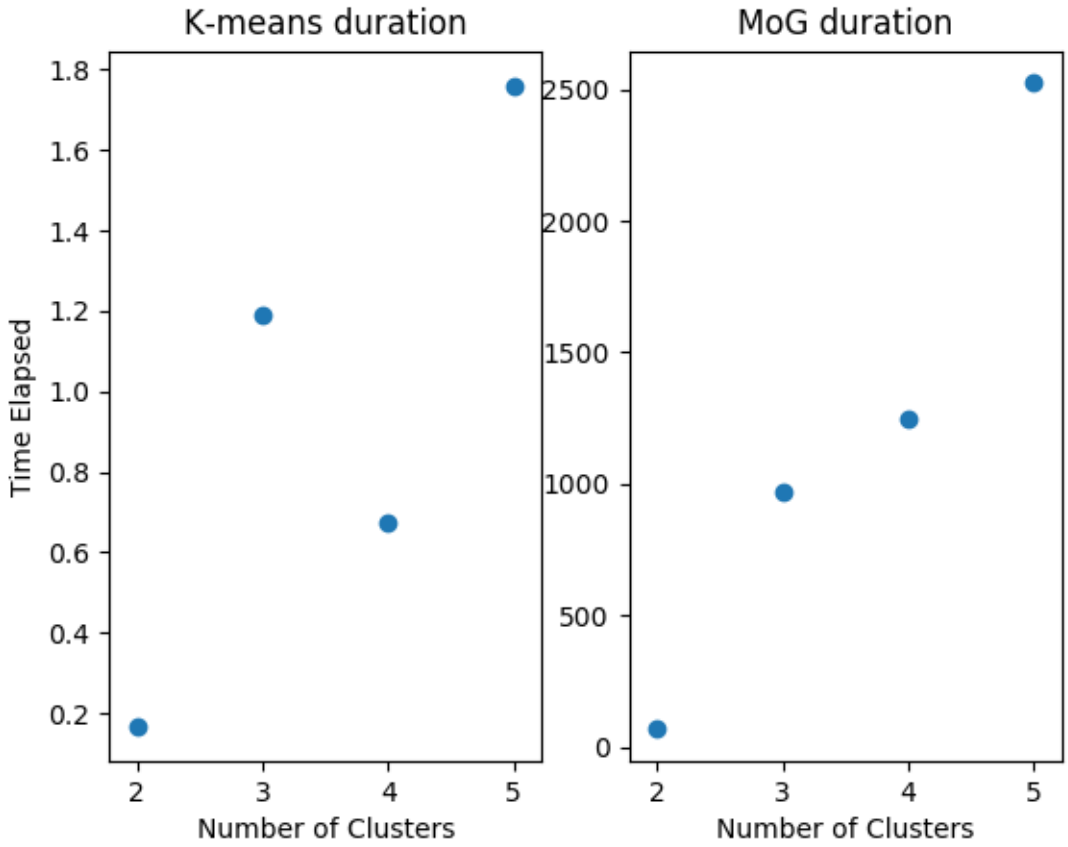
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

For every data point i, a(i) will be the distance of i from all other data points within its cluster, b(i) will be the distance of i from all other data points of the nearest other cluster. The silhouette coefficient s(i) will range from -1 to 1, the higher it is the better the said data point fits the cluster to which it has been assigned. The total coefficient will be the average of all these coefficients.

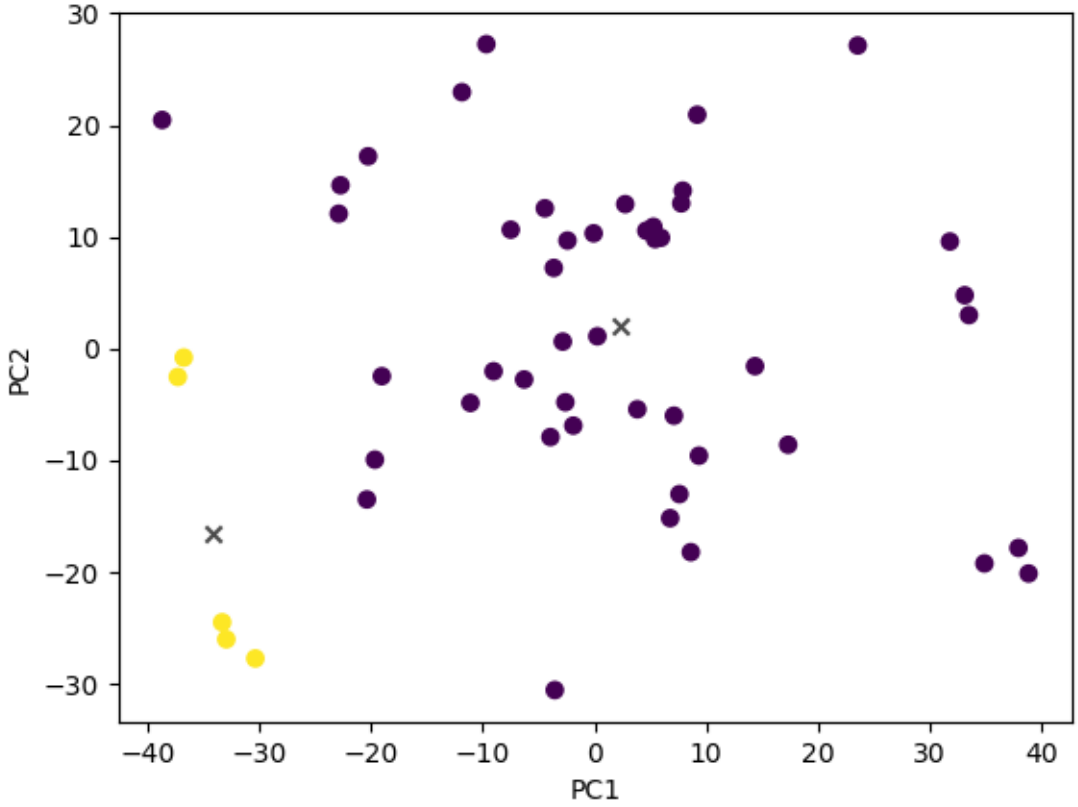


If we weren't aware that the original data was a product of two distributions we could rely on the silhouette coefficient to choose an appropriate number of clusters. As 2 clusters correspond to the highest silhouette coefficient, we would pick that option.

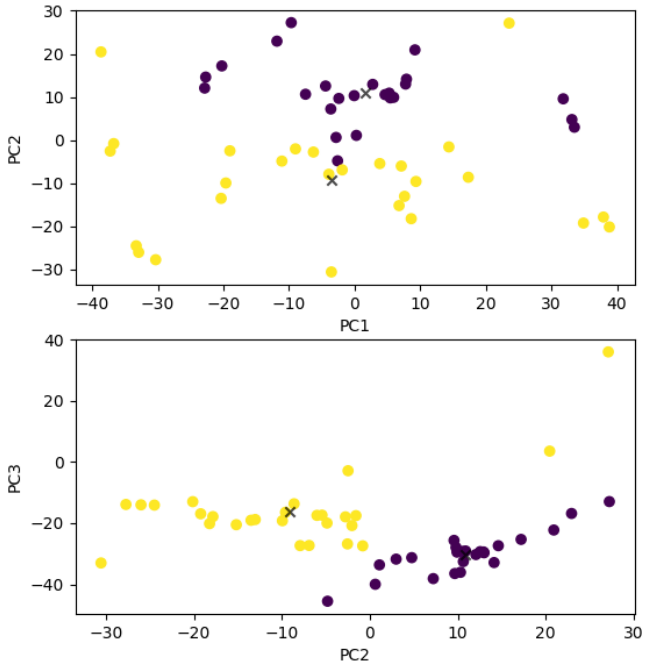
Also a speed test was performed to check how much time elapses for every algorithm. Mixtures of Gaussians takes a lot longer compared to K-means and the time scales greatly with number of clusters.

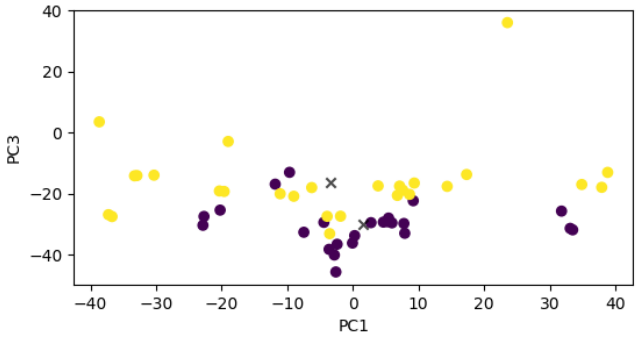


Practical Part:
For the practical part the expression data from the analysis of livers of C57BL/6J mice fed a high fat diet and normal diet was used. Because the data is high dimensional, PCA was performed to reduce it to 2 dimensions and then Mixture of Gaussians was performed. The result was not satisfactory:



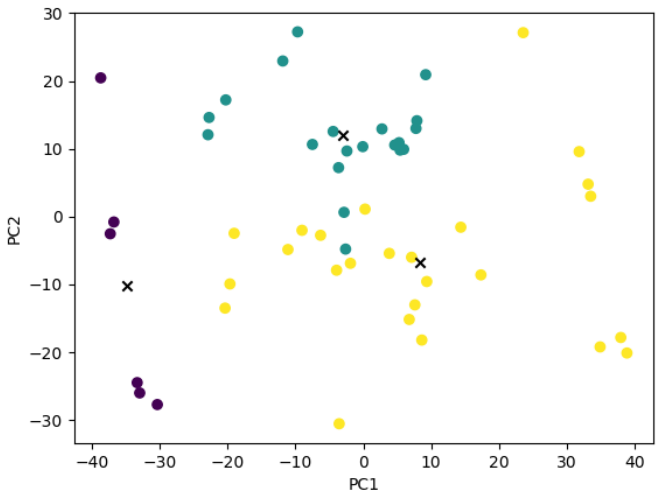
However performing PCA on the data down to 3 dimensions and then performing MoG gives much better results. Below the pairwise plotting of principal components is demonstrated:

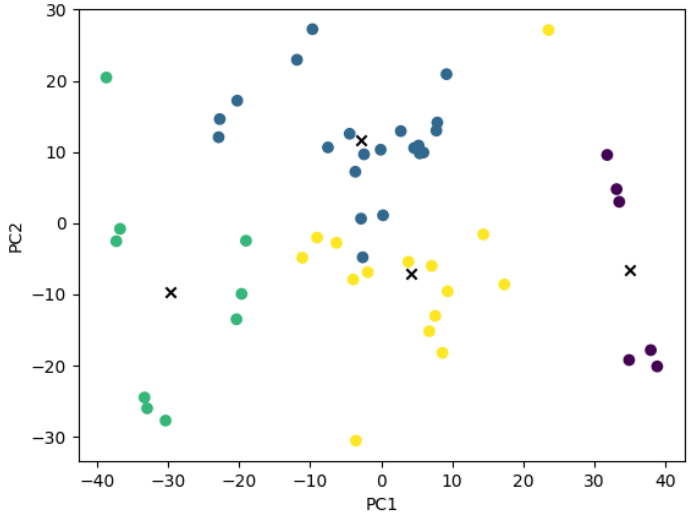




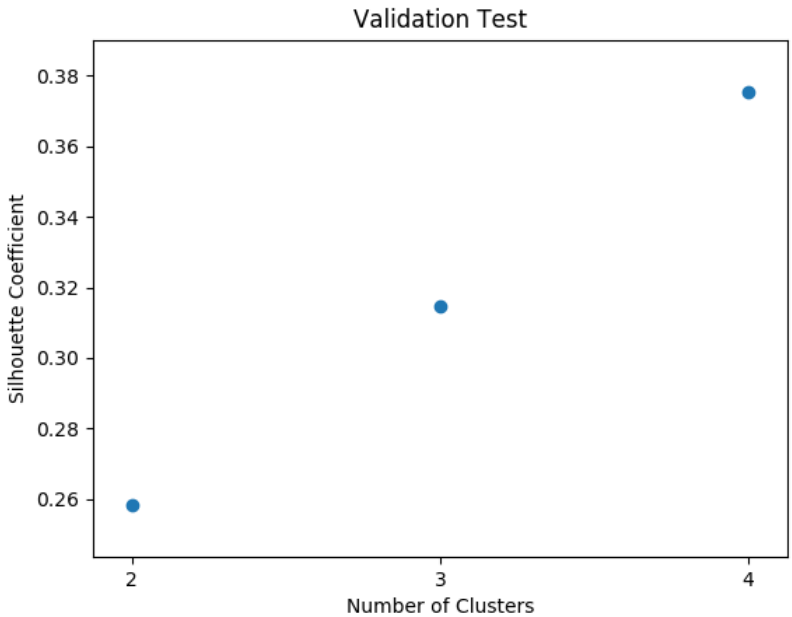
It is clear that the data is distinguishable in 3 dimensions, and MoG can easily cluster it at that space.

We know that our data is made mainly of 2 groups, the high fat diet and normal (well there is baseline but it's kinda redundant due to its small size, might as well be clustered along with the normal diet). However let's try to set the initial conditions to more clusters:



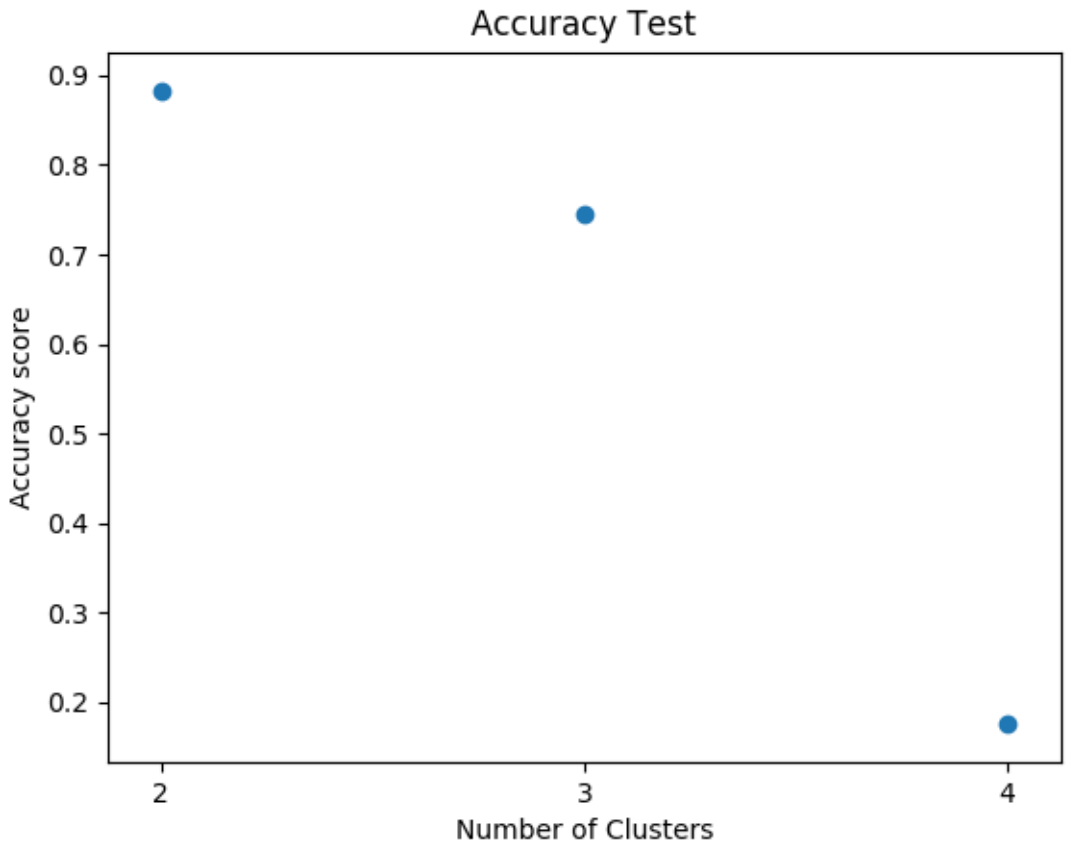


And check for the silhouette coefficient



Even though 2 clusters should be the optimal, we get higher silhouette coefficients for more clusters. This is probably an indication that something is wrong with the dataset.

Lastly since the true labels are known (3 baseline, 24 normal diet, 24 high fat diet) an accuracy test is possible. For this purpose the built in sklearn function “accuracy_score” was used, the results are presented below:



As expected the most accurate one is 2 clusters with almost 90% accuracy, but 3 clusters are also pretty good with around 75% accuracy so maybe the baseline can be quite distinguished. Of course 4 clusters causes an accuracy disaster because there is not even a 4th label.