Report PA5 Claas Fillies

Task 1: Dimensionality Reduction

Subtask 1: Data Loading and Data Preparation

How many different people are in the data?

34 different people are displayed in the dataset if it is reduced to people with a minimum of 30 pictures per person.

How many images are in the data?

2370 different pictures are in the dataset.

What is the size of the images?

Every image has a shape: (62, 47)

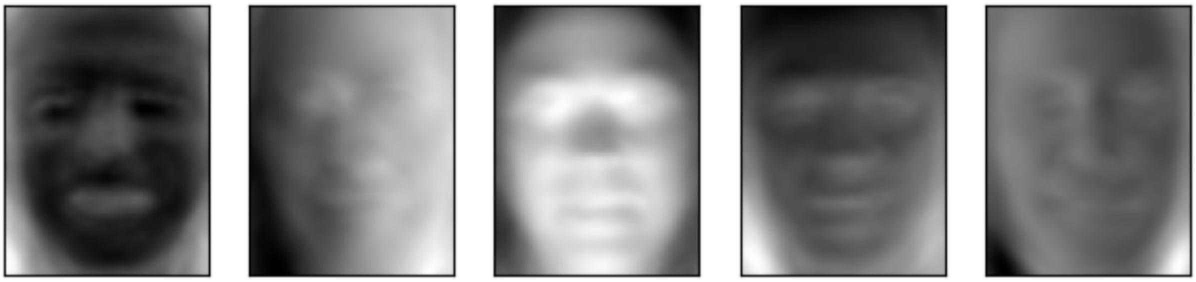
Plot images of ten different people in the data set.

Subtask 2: Dimensionality Reduction Using PCA

Briey describe your implementation in the report.

The PCA class consists of two major functions, the fit function and the reconstruct function. The fit function receives all the pictures in the dataset, substracts the mean of the data, and calculates the covariance matrix as well as the eigenvalues and eigenvectors. The reconstruct function receives an image to reconstruct and a limit of projection eigenvectors. The image gets projected/scored on the sorted amount of eigenvector up to the given limit. From those scores, the projected image can be reconstructed through the formula X^=ZV.T=XVV.T + mean (from lecture). A detailed step-by-step explanation of the implementation is given in the code.

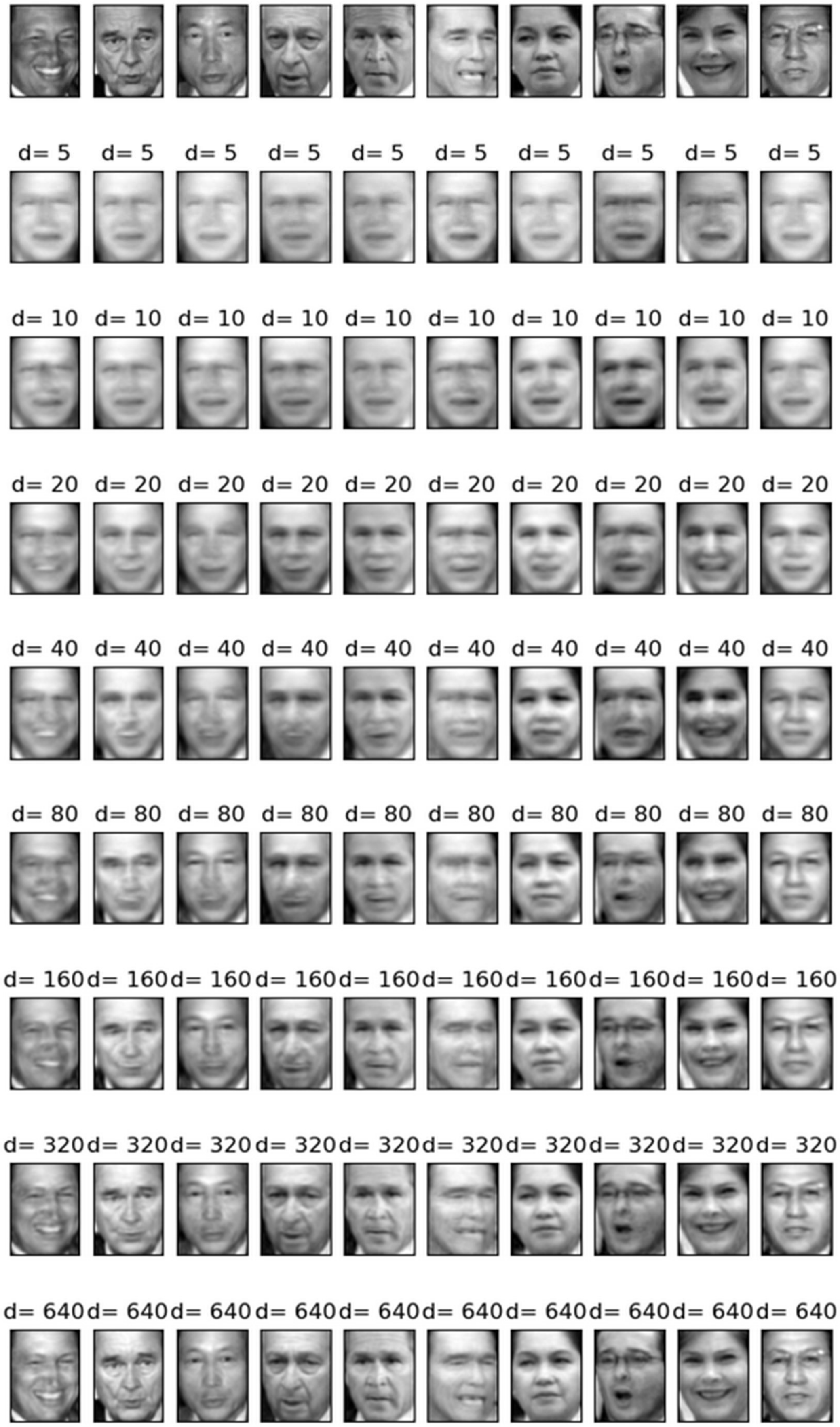
Plot the first 5 principal components as images.



Displayed are the first 5 eigenvectors of the given dataset from the left to the right.

Visualize 10 reconstructed images for each d.

reference pictures:



Report the achieved classification accuracy of the classifiers on the training and the validation set for all d.

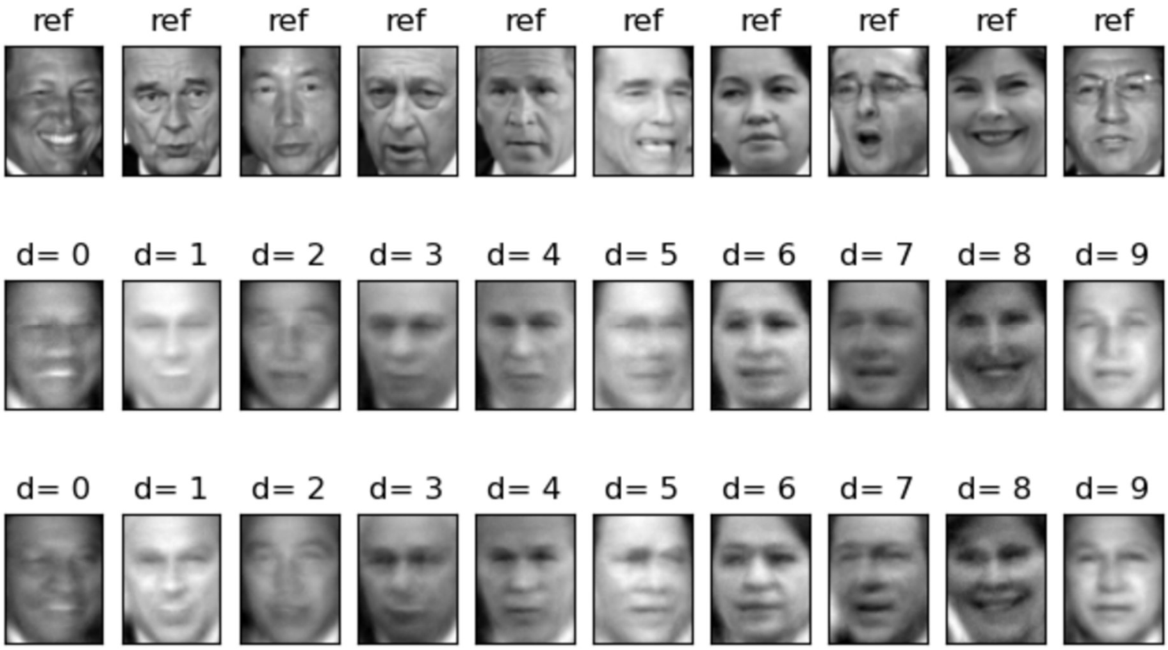
d	Accuracy on training dataset	Accuracy on testing dataset
5	0.033	0.038
10	0.100	0.081
20	0.405	0.349
40	0.678	0.532
80	0.862	0.611
160	0.964	0.660
320	0.989	0.676
640	0.992	0.683
Not projected data as a reference	1	0.708

Comment on your observations.

The results show that it is possible to project the images on their eigenvector reconstruct them and train LRC on them. The data which was reconstructed from 160 eigenvectors could be classified with an accuracy comparable to an LRC which was trained on the original Data. For that reason, it is proved, that most of the information needed to classify the images is contained in a subspace of the first 160 eigenvalues. The accuracy of the LRC increases as soon as more Eigenvalues are considered than different people are contained in the dataset.

Subtask 3: Dimensionality Reduction Using Autoencoders

Visualize 10 reconstructed images for each d.



Report the achieved classiffcation accuracy of the classiffers on the training and validate set for all d.

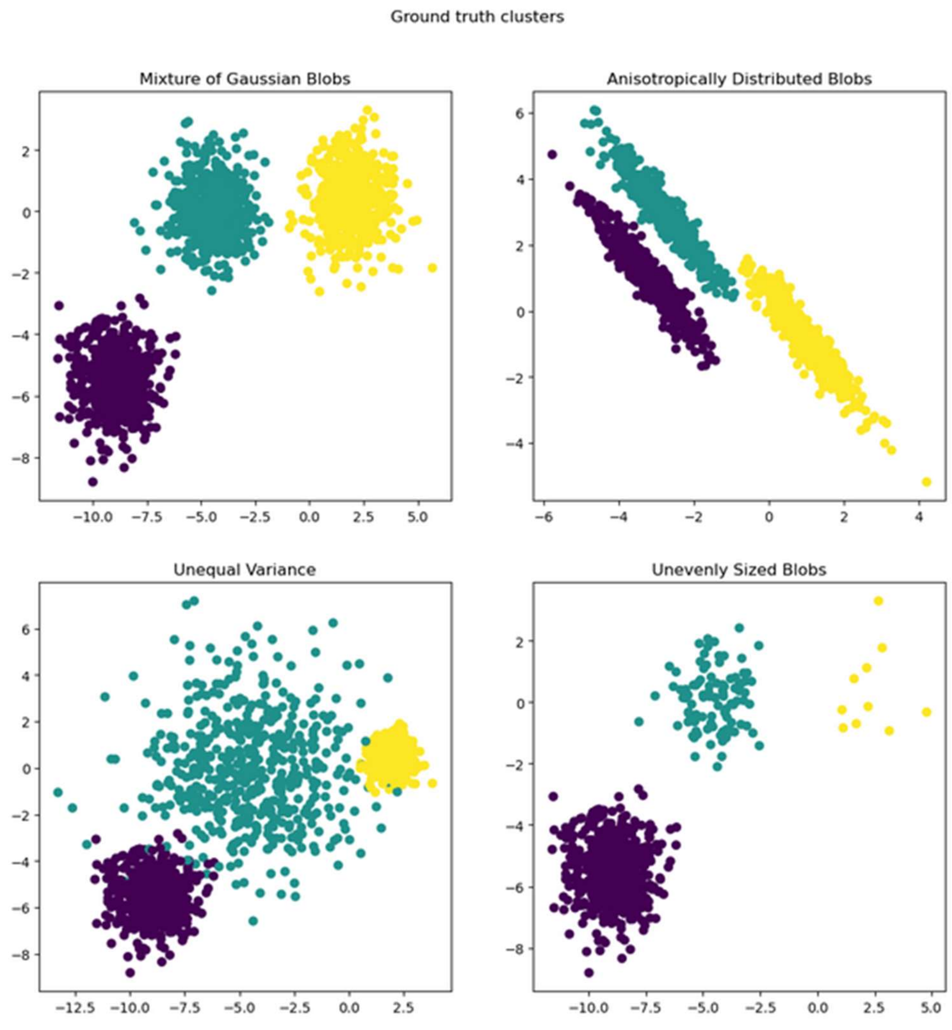
d	Accuracy on training dataset	Accuracy on testing dataset
40	0.296	0.287
80	0.317	0.312

Comment on your observations and compare your results to those of the previous task.

The reconstructed images from the smallest hidden layer size of 40 and 80 were not able to match the accuracy of the data reconstructed from a similar amount of eigenvectors. For that reason, it is advisable to rather use an eigenvalue decomposition to classify these images.

Task 2: Clustering

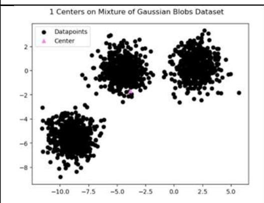
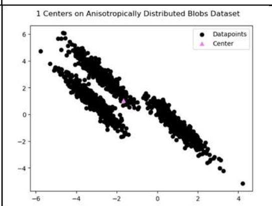
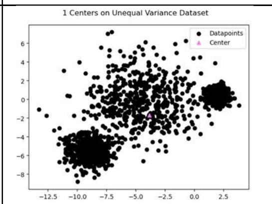
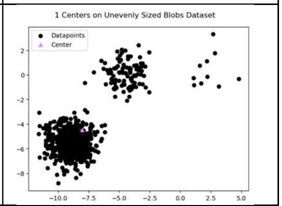
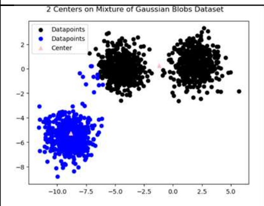
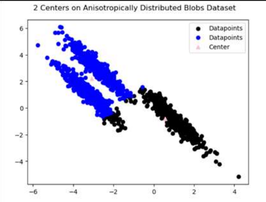
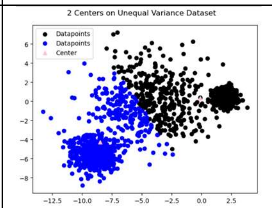
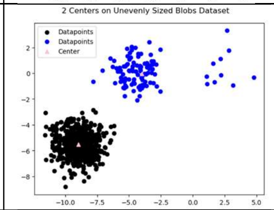
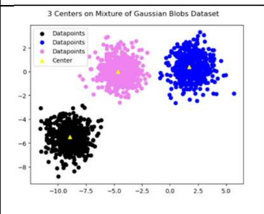
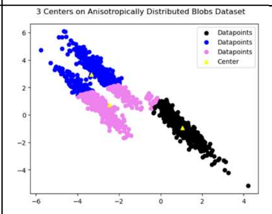
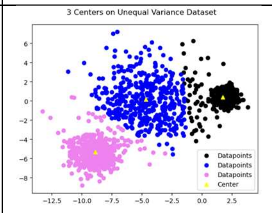
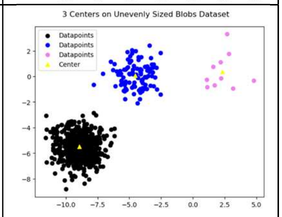
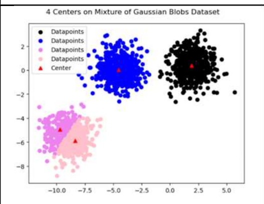
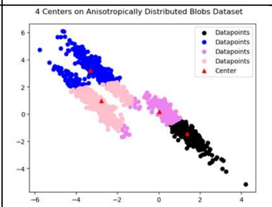
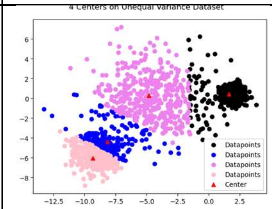
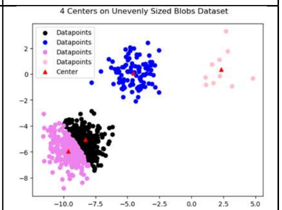
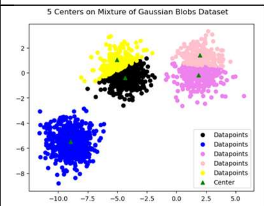
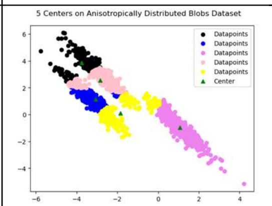
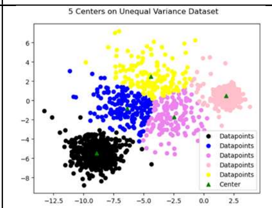
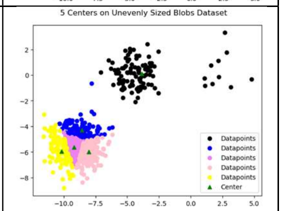
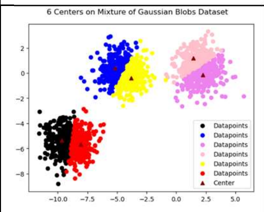
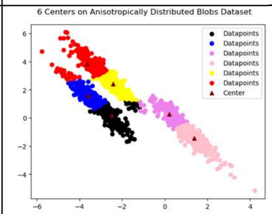
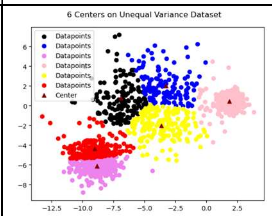
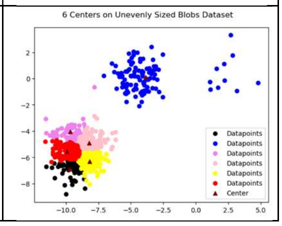
Subtask 1: Generate data



Subtask 2: Lloyd's algorithm

Use your algorithm to perform clustering for k = (1; 2; 3; 4; 5; 6) cluster centers. Report plots of your clustering results. What can you observe regarding the clustering results?

Random distribution of centers

Number (n) of centers	Gaussian Blobs	Anisotropic Blobs	Unequal Variance	Unevenly Sized Blobs
1				
2				
3				
4				
5				
6				

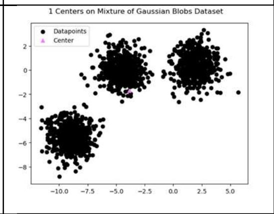
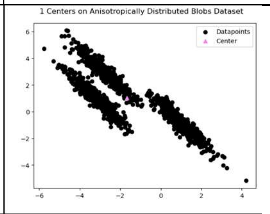
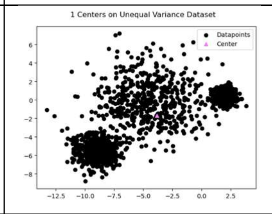
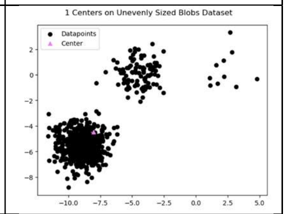
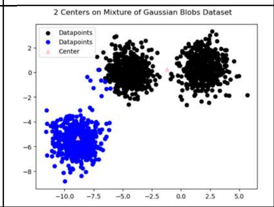
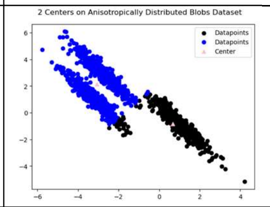
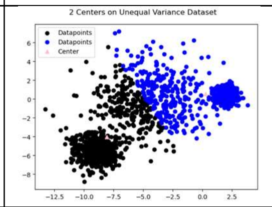
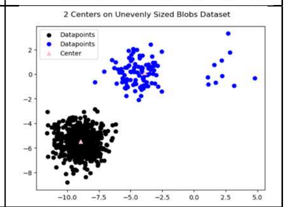
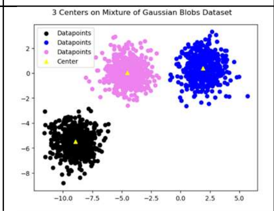
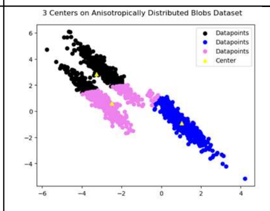
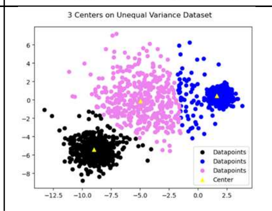
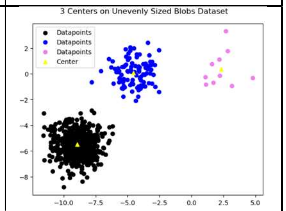
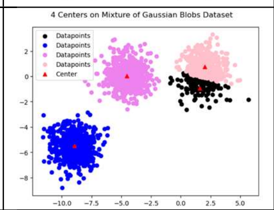
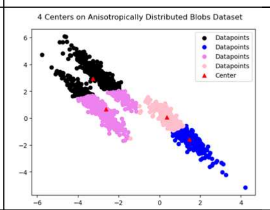
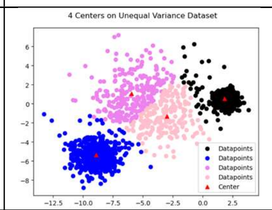
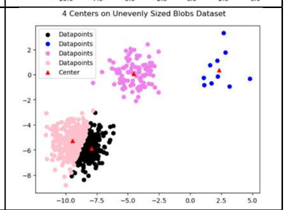
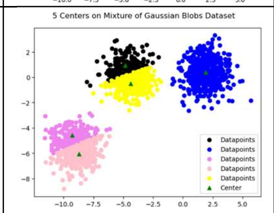
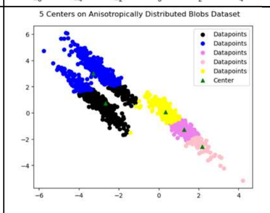
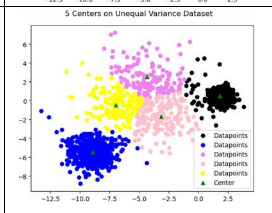
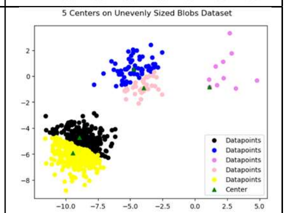
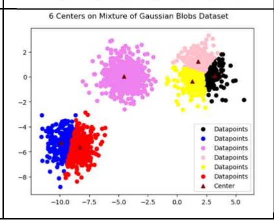
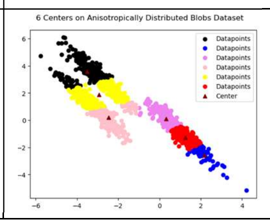
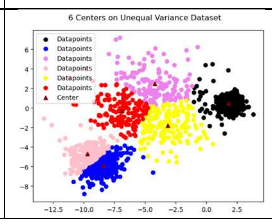
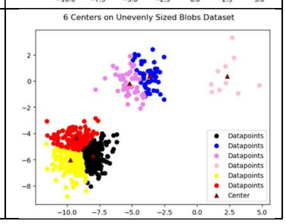
From the results of the clustering, it can be observed, that for n = 3, the Gaussian Blobs and the Unevenly Sized Blobs Datasets can be correctly classified. The Unequal Variance Dataset is to majority correctly classified. However, the grounded truth shows overlapping clusters which makes a perfect classification impossible for the Lloyds algorithm. The Anisotropic Blobs Dataset can not be correctly classified with any number of centers. Even for n = 6, some of the clusters convey points from two ground truth clusters. As a consequence, the Dataset could be kernelized and then clustered to archive a higher accuracy

Report plots showing the k-means objective for the four datasets and k 2= {1; 2; 3; 4; 5; 6}

Not done.

Subtask 3: K-means++ initialization

Report the same plots as in the previous task. Are there any differences?

Numbe r (n) of centers	Gaussian Blobs	Anisotropic Blobs	Unequal Variance	Unevenly Sized Blobs
1				
2				
3				
4				
5				
6				

Up to n = 3, there are no differences in the centering results. However, the needed steps to reach the results are fewer because the inial centers are already farther spread out. For n > 3 the likelihood of split grounded center increases because the initial centers are spread farther out.