UNIVERSITÀ DI PARMA

Department of Engineering and Architecture
Master of Science in Communication Engineering

Machine Learning

Academic Year 2023-2024

Final Assignment Report

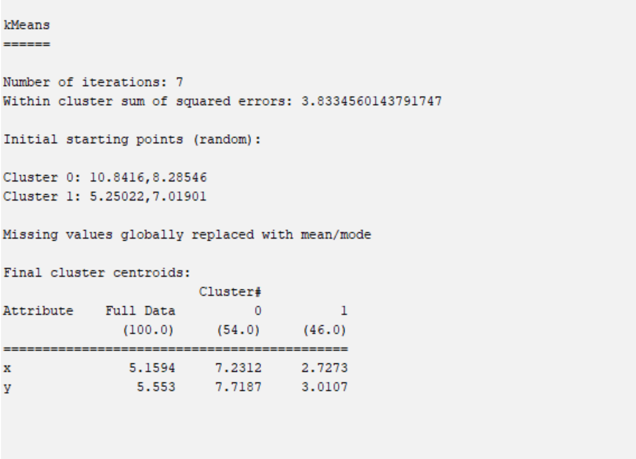
Student: MOHAMMADREZA RAEESOSADATI
Identification number: 351476

I. Introduction

Clustering and classification are fundamental methods in machine learning used to group and differentiate data points. This project aims to identify data points originating from two separate Gaussian distributions, utilizing a variety of clustering and classification techniques. In one part of this project, we apply the K-Means algorithm to our training data. K-Means, an unsupervised learning method, divides data into K clusters based on similarities in features. We will evaluate K-Means' ability to separate the two distributions and validate the clustering results with suitable metrics. Additionally, we will use supervised classification methods to label data points. This process involves training a classification model on a labeled dataset and then predicting labels for a test dataset. To further challenge the algorithms, we will introduce higher variance into the Gaussian distributions. This will help us examine the algorithms' performance and robustness when dealing with data that has increased variability.

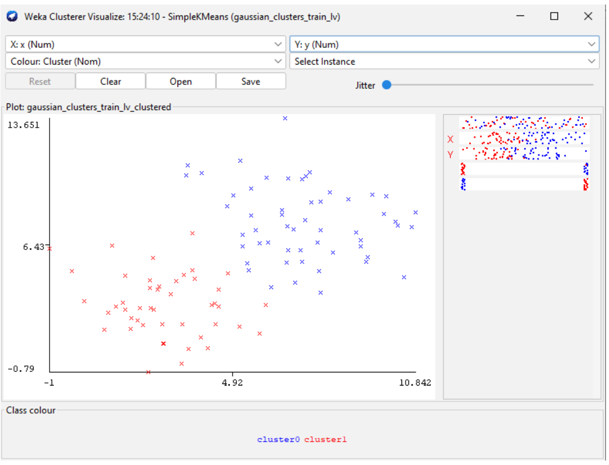
II. Numerical analysis

PART ONE: In the first part of the exercise (0), we applied the K-Means clustering algorithm to a training dataset containing 100 samples from two Gaussian distributions. The Gaussian distributions were centered at (3,3) and (7,7) with equal variances along both axes. Using Weka's SimpleKMeans, we clustered the data into two groups. The result of the clustering is shown in the figure1 and figure2, as it is shown the centroid of this exercise is (7.2,7.7) and (2.7,3) this indicates that the centroids are close to the true centers of the Gaussian distributions. The centroid near (7.2, 7.7) corresponds to the Gaussian distribution centered at (7, 7), and the centroid near (2.7, 3.01) corresponds to the Gaussian distribution centered at (3, 3). This proximity suggests that K-Means effectively identified the underlying structure of the data



```
kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 3.8334560143791747

Initial starting points (random):

Cluster 0: 10.8416,8.28546
Cluster 1: 5.25022,7.01901

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1
                 (100.0)     (54.0)     (46.0)
==============================================
x                 5.1594     7.2312     2.7273
y                  5.553     7.7187     3.0107
```

Figure 1(Centroids)



Weka Clusterer Visualize: 15:24:10 - SimpleKMeans (gaussian_clusters_train_lv)

X: x (Num)

Y: y (Num)

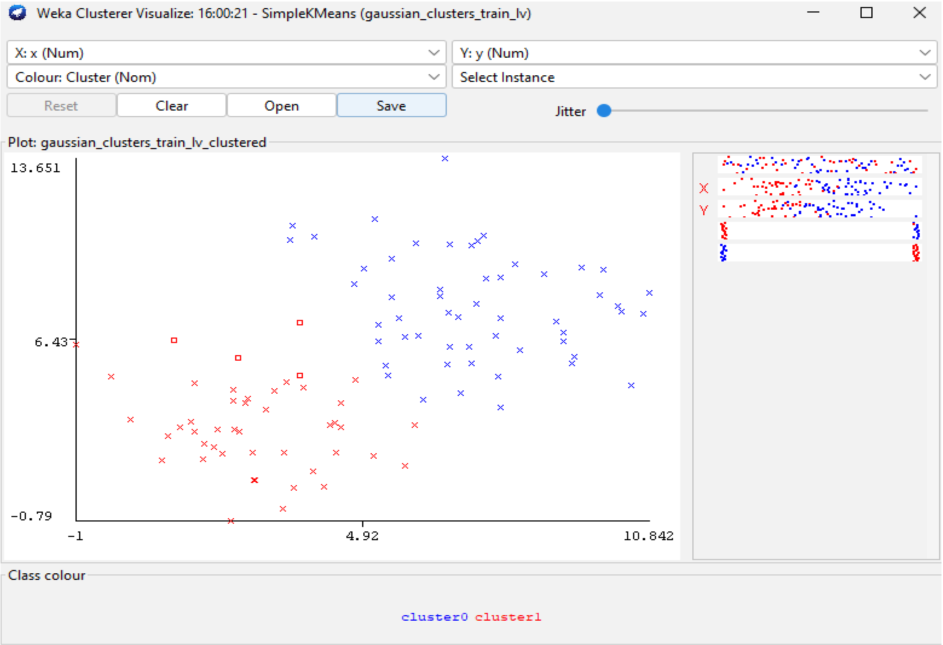
Colour: Cluster (Nom)

Select Instance

Plot: gaussian_clusters_train_lv_clustered

Figure 2(Clustering assignment result)

SECOND PART: In the second part of the exercise (0b) we are trying to implement a supervised classification by the same data from previous exercise, just here we use select "Classes to Clusters evaluation" as "Cluster mode" using "Gaussian" as class attribute so Centroids/Clusters are assigned to the class to which the majority of patterns in the cluster belong, The centroids for the clusters were recalculated, and the results showed that the centroids were at (7.2, 7.7) and (2.7, 3), which are close to the true centres of the Gaussian distributions at (7, 7) and (3, 3), respectively. There were four misclassifications observed in this supervised classification task. Misclassifications occur when data points are assigned to the wrong cluster, meaning they are closer to the centroid of the incorrect Gaussian distribution this is happening often with data points that lie near the boundary between the two Gaussian distributions. These points may not have a clear distinction and can be closer to the centroid of the opposite cluster due to the inherent variability in the data. The equal variances along both axes contribute to overlapping regions where data points from both distributions are intermixed. This overlap can lead to ambiguity in cluster assignment, resulting in misclassifications.



Weka Clusterer Visualize: 16:00:21 - SimpleKMeans (gaussian_clusters_train_lv)

X: x (Num)

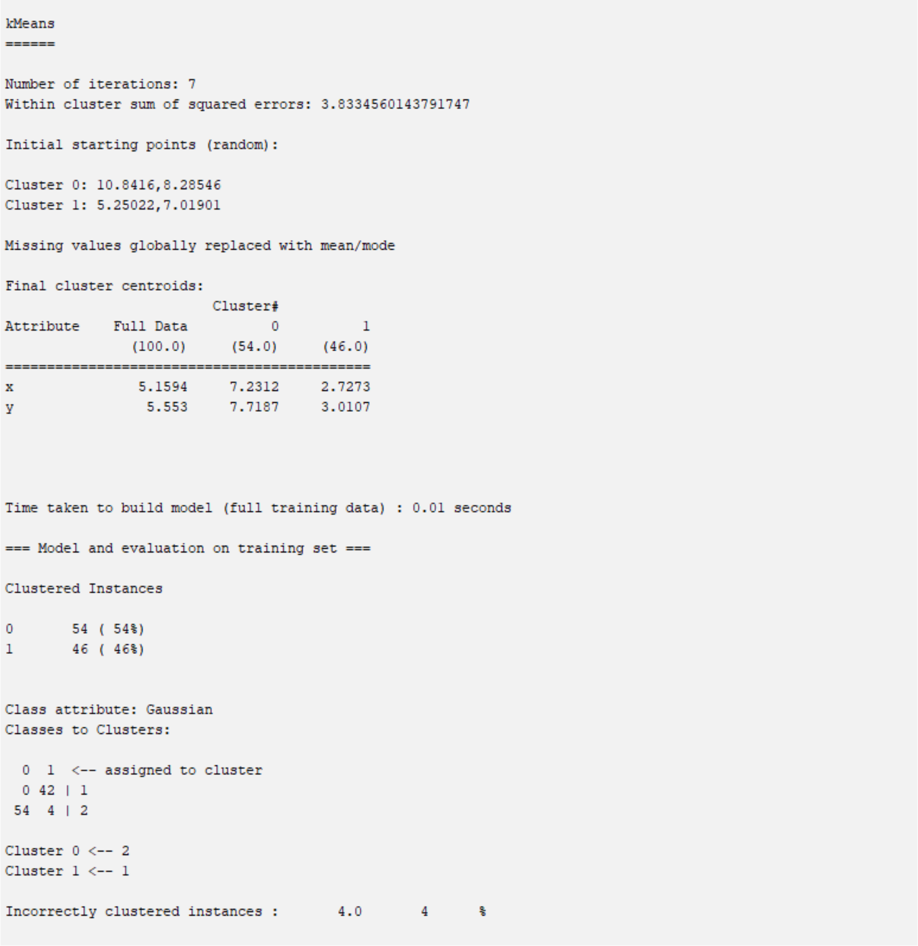
Y: y (Num)

Colour: Cluster (Nom)

Select Instance

Plot: gaussian_clusters_train_lv_clustered

Figure 3(two clusters and misclassification)



```
kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 3.8334560143791747

Initial starting points (random):

Cluster 0: 10.8416,8.28546
Cluster 1: 5.25022,7.01901

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
               (100.0)     (54.0)     (46.0)
============================================
x               5.1594     7.2312     2.7273
y                5.553     7.7187     3.0107




Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       54 ( 54%)
1       46 ( 46%)


Class attribute: Gaussian
Classes to Clusters:

  0  1  <-- assigned to cluster
  0 42 | 1
 54  4 | 2

Cluster 0 <-- 2
Cluster 1 <-- 1

Incorrectly clustered instances :	4.0	 4      %
```

Figure 4(supervised classification)

Figure 3 and figure 4 are showing the result of the clustering with supervised classification and the misclassified location which occurred in this part.

THIRD PART: In this section (0C), we examine the effect of increased variance on the clustering results by using the file gausstrainhv.arff. The Gaussian means remain the same as before, but the variance has increased by 50%. With increased variance, the overlap between the two Gaussian distributions will be larger. Data points near the decision boundary are more likely to be assigned to the "wrong" cluster. Specifically, points that are closer to (5, 5) may frequently be misassigned because they fall within the expanded overlapping region. Is It Really the "Wrong" Cluster?

The concept of a "wrong" cluster becomes more nuanced with increased variance. As the data points spread out, the decision boundary becomes less distinct, and the probability of points being near the boundary increases. Also, the inherent variability in the data due to increased variance means that even though a point may be closer to the centroid of the opposite cluster, it is still a valid representation of the distribution's spread. Therefore, in clustering, particularly K-Means, clusters are formed based on minimizing within-cluster variance. If a data point is closer to a different centroid due to higher variance, it is not necessarily misclassified but rather assigned to the cluster that best represents its location in the feature space.

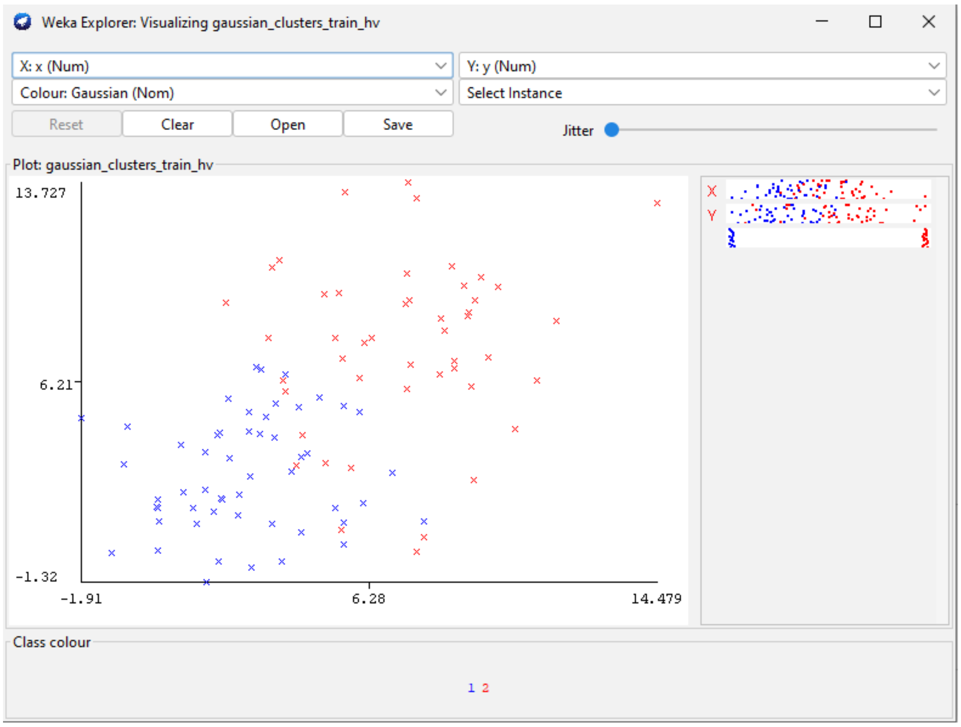
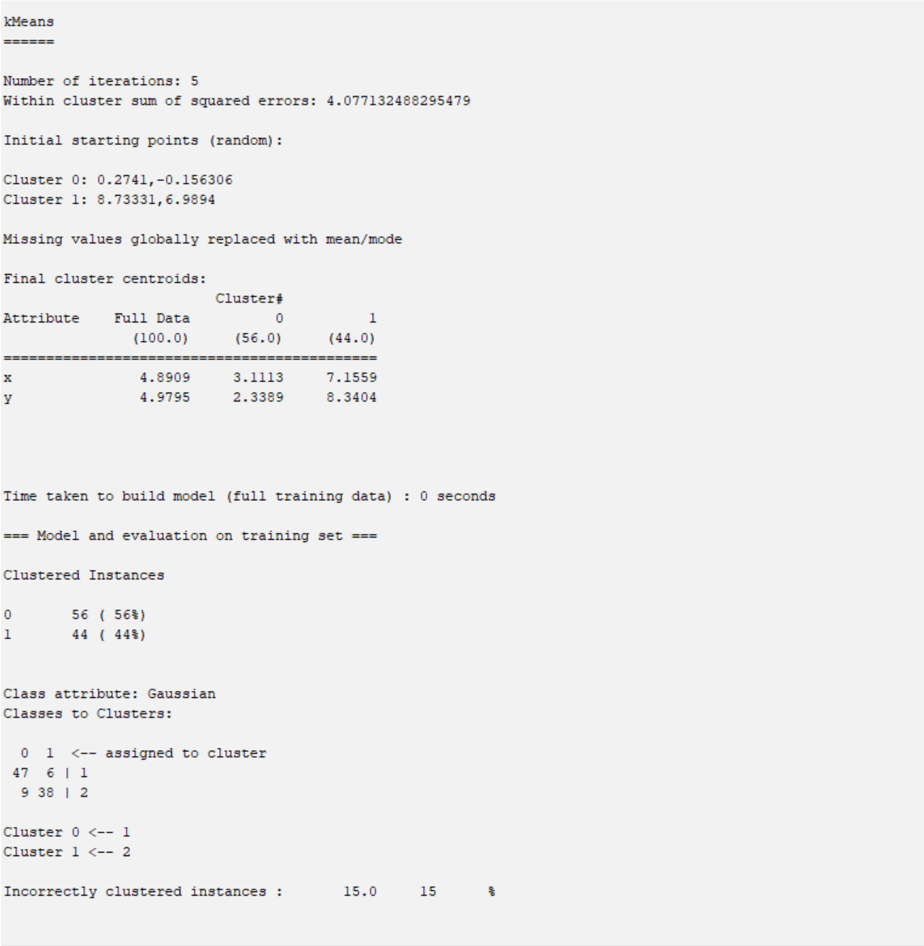


Figure 5(variance increased by 50%)

In the following we launch the K-means clustering and got the information below:



```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 4.077132488295479

Initial starting points (random):

Cluster 0: 0.2741,-0.156306
Cluster 1: 8.73331,6.9894

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute     Full Data         0          1
                (100.0)    (56.0)     (44.0)
============================================
x                4.8909    3.1113     7.1559
y                4.9795    2.3389     8.3404




Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       56 ( 56%)
1       44 ( 44%)


Class attribute: Gaussian
Classes to Clusters:

  0  1  <-- assigned to cluster
 47  6 | 1
  9 38 | 2

Cluster 0 <-- 1
Cluster 1 <-- 2

Incorrectly clustered instances :       15.0      15      %
```

Figure 6(Cluster centroids by 50% higher variance)

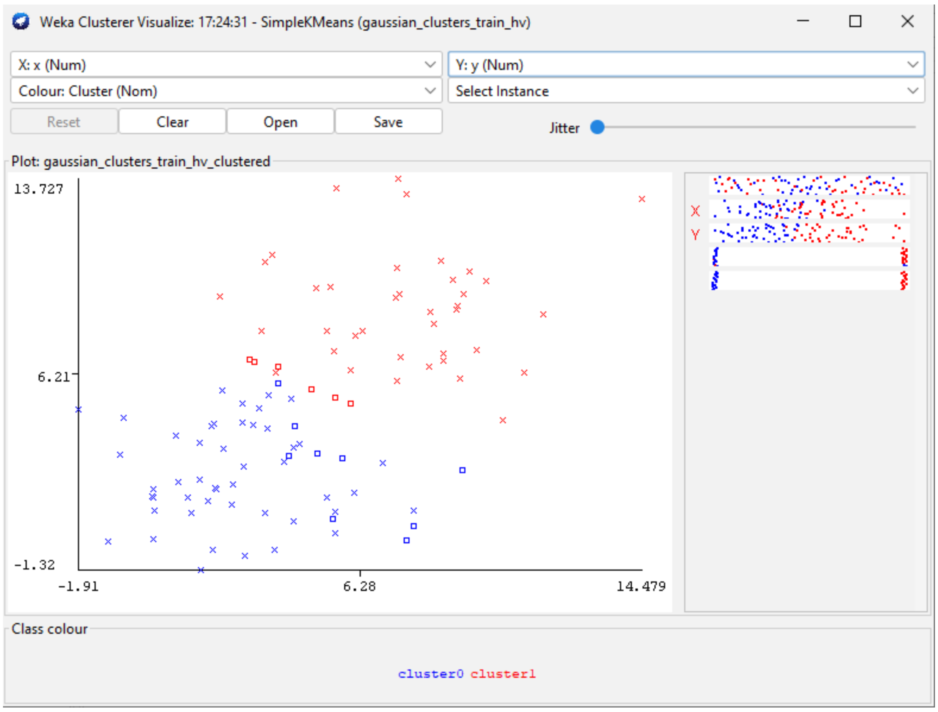
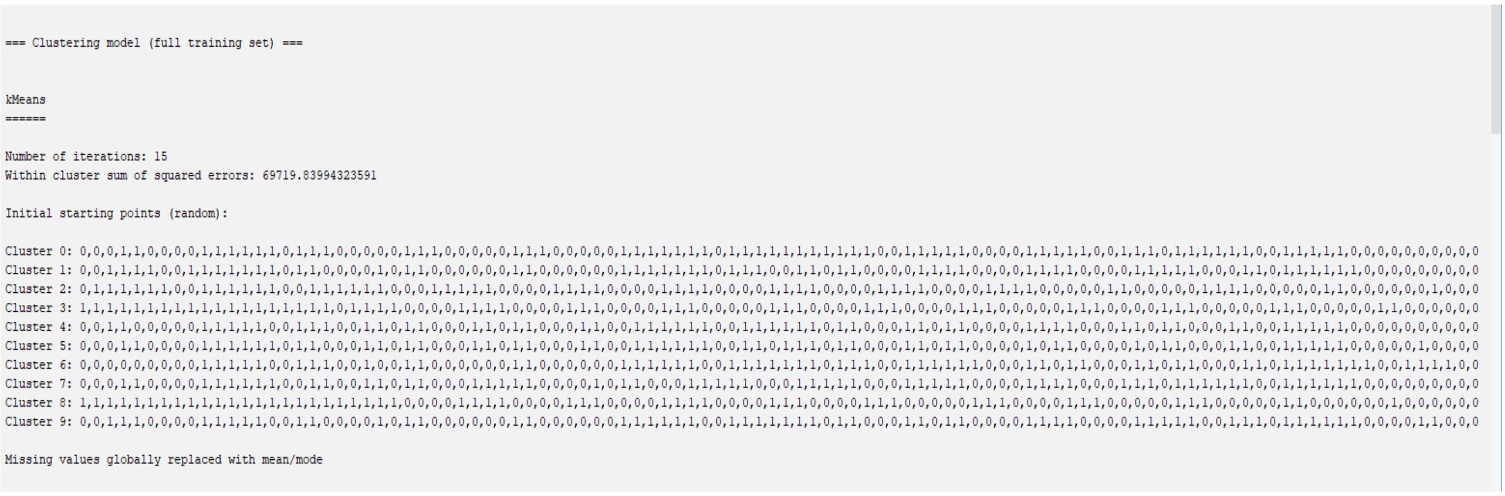


Figure 7(Classification and missclassification by 50% higher variance)

The centroids were found at (3.1, 2.3) and (7.1, 8.3), with the second centroid being further from its true Gaussian center compared to the first. The 15 misclassifications highlight the challenge posed by the increased overlap between distributions. These points tend to lie in regions of higher ambiguity, where the increased spread in the data results in more overlap and less distinct cluster boundaries. This analysis underscores the importance of considering data variability when interpreting clustering results.

FOURTH PART: In this part we are working with a dataset "**Bigtest1_104.arff**" consisting of 10 different patterns representing the digits 0 through 9. Each pattern is an image of size 13x8 pixels, flattened into a 104-dimensional vector. The goal is to perform clustering using K-Means and then visualize the centroids as grayscale images to interpret the clustering results. So in this case each centroid is represented as a 13x8 matrix (104-dimensional vector) where each element (pixel) corresponds to a grayscale value proportional to the probability of the pixel being "ON" (i.e., 1) within the cluster. Each centroid is the average of the patterns in the cluster, which means that the centroid represents the most common features of the digits in that cluster. If the clustering algorithm worked correctly, each cluster should contain patterns of the same digit, leading to a centroid that looks like that digit. The grayscale level of each pixel indicates the probability that a pixel is "ON" across the patterns in that cluster. For example, in a cluster representing the digit "0," the pixels forming the circular shape of "0" will be closer to "ON" (white) in the centroid image, while other pixels will be closer to "OFF" (black) and then by using MATLAB code we should visualize the centroids as grayscale images. When observing these centroids, we should see that they closely resemble the digits they represent. Figure8 shows the clusters values in the WEKA and figure9 represent the sample result of the centroids as grayscale images. Generally, the centroids resemble the digits they represent because K-Means clustering groups similar patterns together. Therefore, the centroid, which is the average pattern of a cluster, should look like a blurry or averaged version of the digits within that cluster.



```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 15
Within cluster sum of squared errors: 69719.83994323591

Initial starting points (random):

Cluster 0: 0,0,0,1,1,0,0,0,0,1,1,1,1,1,1,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,0,0,1,1,1,1,1,0,0,0,0,1,1,1,1,1,0,0,1,1,1,0,1,1,1,1,1,1,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0
Cluster 1: 0,0,1,1,1,1,0,0,1,1,1,1,1,1,1,0,1,1,0,0,0,0,1,0,1,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,1,1,1,1,1,1,0,1,1,1,0,0,1,1,0,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,1,0,0,0,1,1,0,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0
Cluster 2: 0,1,1,1,1,1,1,0,0,1,1,1,1,1,1,0,0,1,1,1,1,1,1,0,0,0,1,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,0,1,1,0,0,0,0,0,1,1,1,1,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0
Cluster 3: 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,1,1,1,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,0,1,1,0,0,0,0,0,0
Cluster 4: 0,0,1,1,0,0,0,0,0,1,1,1,1,1,0,0,1,1,1,0,0,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,1,1,0,0,1,1,1,1,1,1,0,0,1,1,1,1,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,0,1,1,1,1,0,0,0,1,1,0,1,1,0,0,0,1,1,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0
Cluster 5: 0,0,0,1,1,0,0,0,0,1,1,1,1,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,1,1,0,0,1,1,1,1,1,1,0,0,1,1,0,1,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,0,1,0,1,1,0,0,0,0,1,0,1,1,0,0,0,1,1,0,0,1,1,1,1,1,0,0,0,0,0,1,0,0,0,0
Cluster 6: 0,0,0,0,0,0,0,0,0,1,1,1,1,1,0,0,1,1,1,0,0,1,0,0,1,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,0,1,1,1,0,0,1,1,1,1,1,1,0,0,0,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,1,1,0,1,1,1,1,1,1,1,0,0,1,1,1,1,0,0
Cluster 7: 0,0,0,1,1,0,0,0,0,1,1,1,1,1,1,0,0,1,1,0,0,1,1,0,1,1,0,0,0,1,1,1,1,1,0,0,0,0,1,0,1,1,0,0,0,1,1,1,1,1,0,0,0,1,1,1,1,1,0,0,0,1,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,1,1,1,0,1,1,1,1,1,1,0,0,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0
Cluster 8: 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,0,0
Cluster 9: 0,0,1,1,1,0,0,0,0,1,1,1,1,1,0,0,1,1,0,0,0,0,1,0,1,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,1,0,0,1,1,1,0,1,1,1,1,1,1,0,0,0,0,1,1,0,0,0

Missing values globally replaced with mean/mode
```

Figure 8(Clusters values)

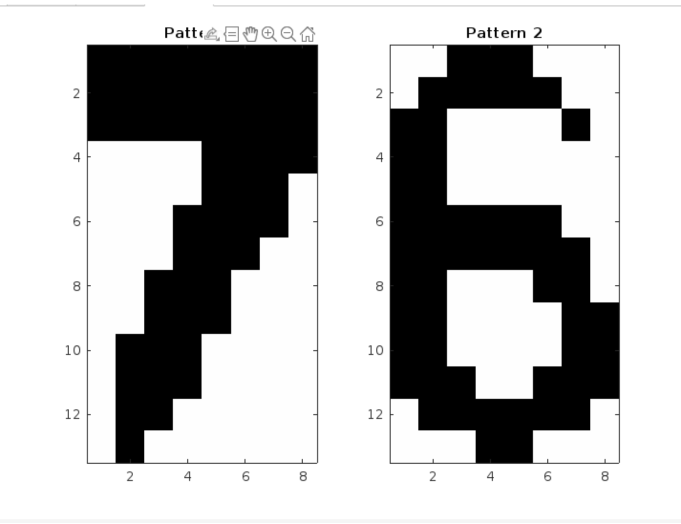


Figure 9(sample result of the centroids)

Fifth Part: Basically, **X-Means** is an extension of K-Means that automatically determines the number of clusters. It is based on a model selection criterion to choose the optimal number of clusters. So, in this case we can automatically determines the optimal number of clusters.
Here in our exercise XMeans is expected to find an optimal number of clusters between 10 and 40. If it correctly identifies clusters representing different digits, it indicates that XMeans has effectively captured the structure of the data. The accuracy of the clustering can be evaluated by comparing the clusters to the actual digit classes. XMeans should ideally produce clusters that closely align with the actual classes, leading to a high accuracy rate. If XMeans finds more or fewer clusters than expected, it indicates that the inherent structure of the data might be more complex or simpler than the predefined range.

III. Conclusions

This project delved into the use of clustering and classification techniques on datasets generated from Gaussian distributions and digit patterns. The findings highlight both the capabilities and the challenges associated with these methods in different scenarios.

In the initial phase, the K-Means algorithm proved effective in identifying clusters corresponding to the two Gaussian distributions with low variance, as the centroids were found to be close to the true centers. However, the supervised classification task exposed the issue of misclassification, especially near the decision boundaries where the distributions overlap. When the variance was increased, this problem became more pronounced, leading to more ambiguous cluster boundaries and a rise in misclassifications. This points to the critical need to account for data variability when interpreting clustering outcomes.

In the analysis of digit patterns, K-Means was able to cluster similar patterns effectively, with the resulting centroids closely resembling the digits when displayed as grayscale images. This suggests that K-Means can capture the key features of each digit cluster, although the centroids tend to represent an averaged version of the digits due to the nature of the algorithm.

The application of X-Means introduced a dynamic clustering approach, showing promise in determining the optimal number of clusters without prior assumptions. The ability of X-Means to adjust the number of clusters based on the data's structure indicates its potential usefulness for more complex datasets.

In summary, this project demonstrates that while K-Means and its extensions like X-Means are valuable tools for clustering and classification, their effectiveness can be significantly influenced by factors such as data variance and the underlying structure of the data. A deep understanding of these factors is essential for the successful application of these algorithms to real-world datasets.