Activity 3

Unsupervised-learning

- **Git Repository**

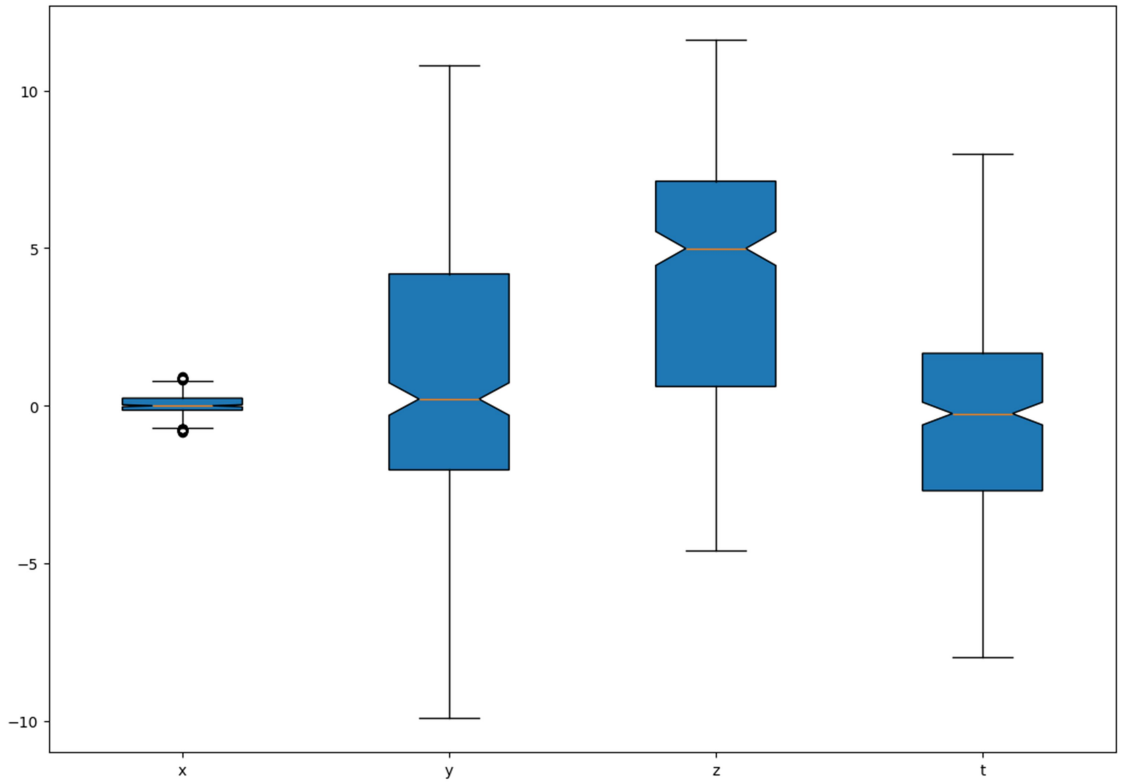
  https://github.com/YoussefEzz/Unsupervised-learning

- **Part 1 : Selecting and analyzing the datasets**

  a) **A3-data.txt analyzed in A3-data.ipynb**

  Features: 4 variables(**x**, **y**, **z**, **t**) of type float, 1 class of type integer(labels 1 : 5)

  Patterns: 360 patterns



  b) **2nd Dataset: Pumpkin_Seeds_Dataset analyzed and preprocessed in pumpkin_seeds.ipynb**

  URL : https://www.kaggle.com/datasets/muratkokludataset/pumpkin-seeds-dataset

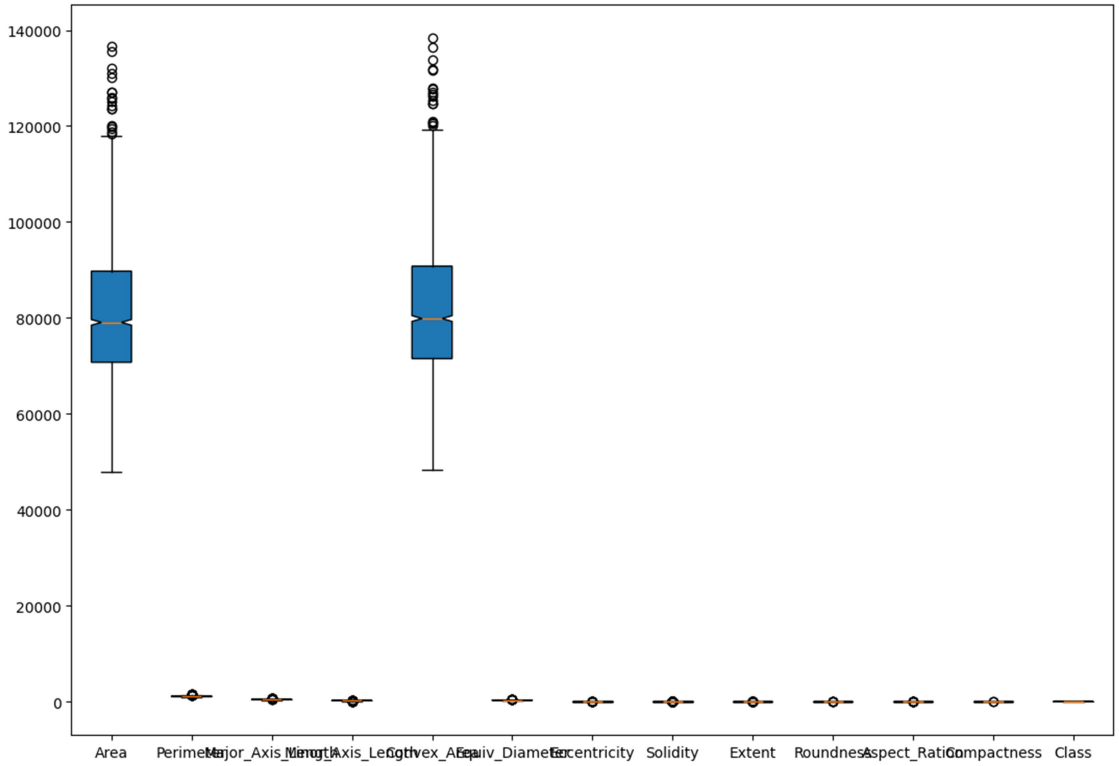
  Features: 12 variables(Area, Perimeter, Major_Axis_Length, Minor_Axis_Length, Convex_Area ,Equiv_Diameter, Eccentricity, Solidity, Extent, Roundness, Aspect_Ration, Compactness) 10 of type float and 2 of type integer, 1 class of type integer(labels 0,1) after preprocessing

  Patterns: 2500 patterns

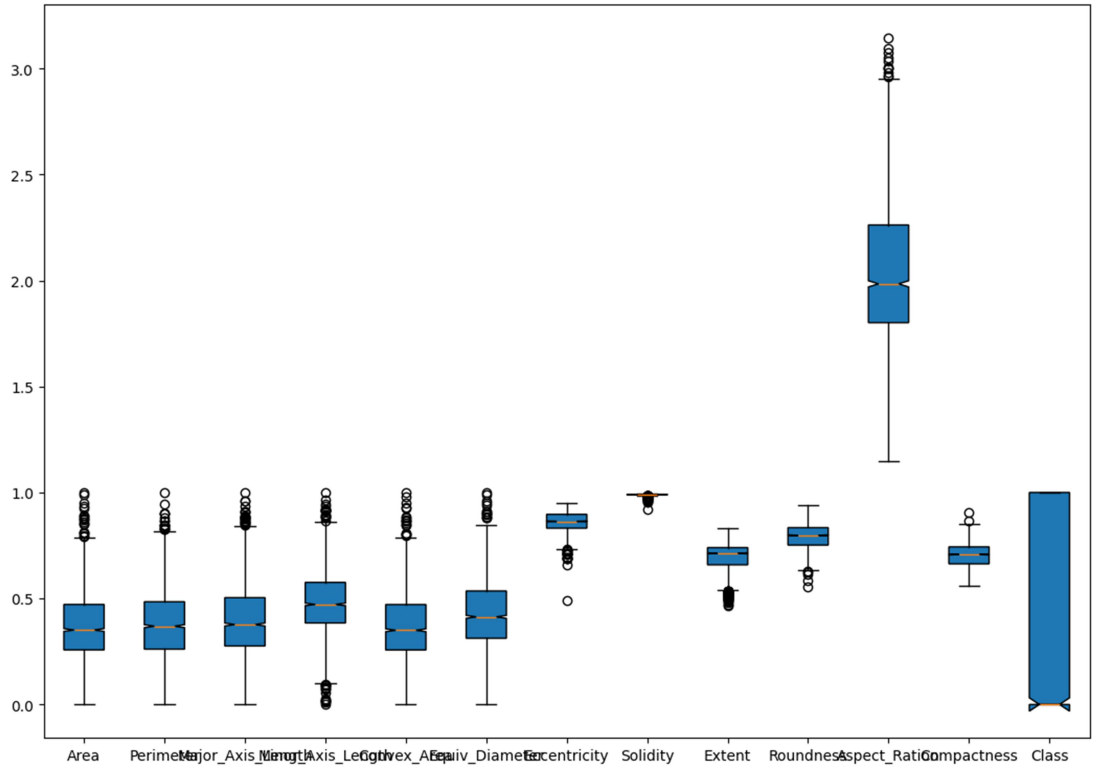
  Preprocessing:

  - Class labels **Çerçevelik** and **Ürgüp Sivrisi** changed using label encoding to 0 and 1 respectively
  - Columns ("Area", "Perimeter", "Major_Axis_Length", "Minor_Axis_Length", "Convex_Area", "Equiv_Diameter") scaled to be from 0 to 1

Before Preprocessing

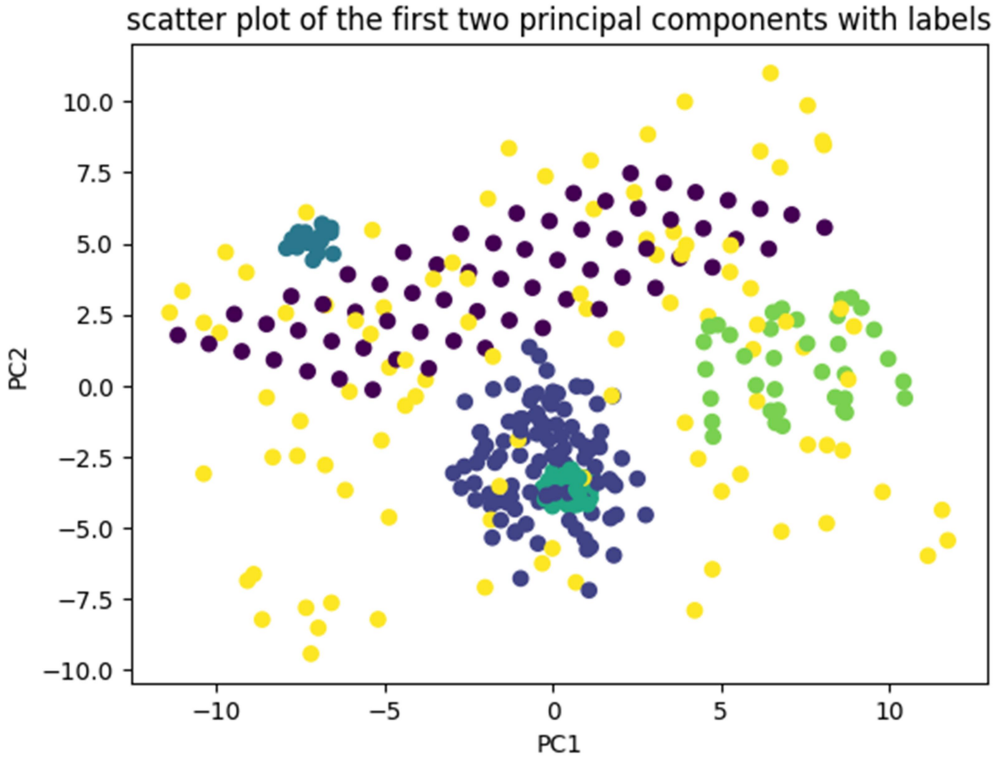


After Preprocessing

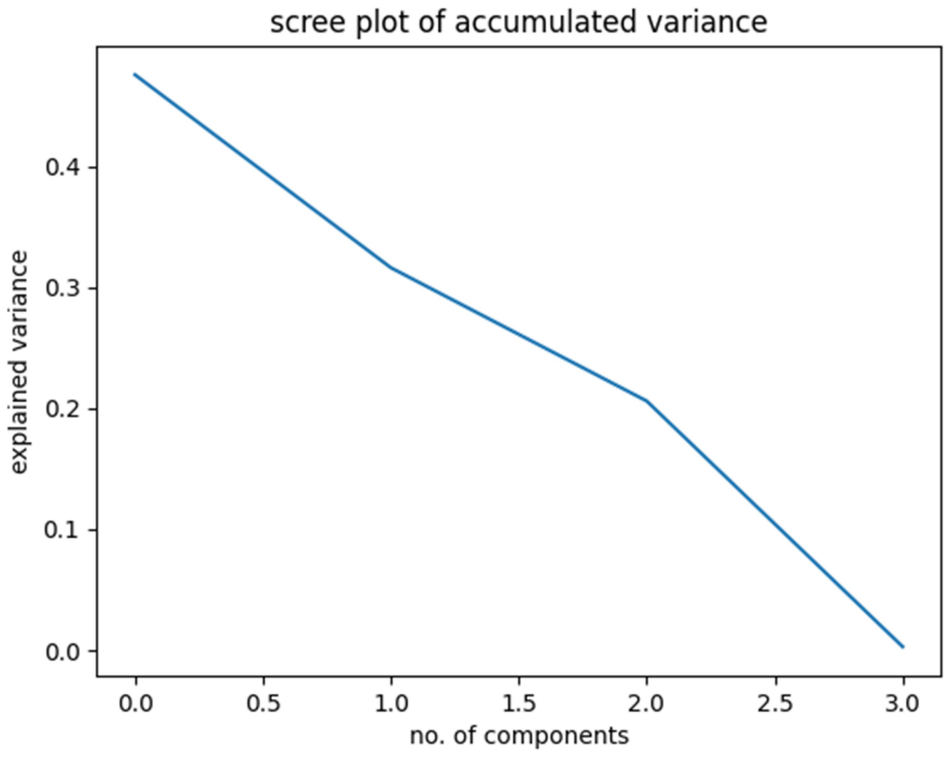


- **Part 2: Comparing unsupervised learning algorithms for A3-data.txt**
  - **a) PCA** analysed in **PCA.ipynb**

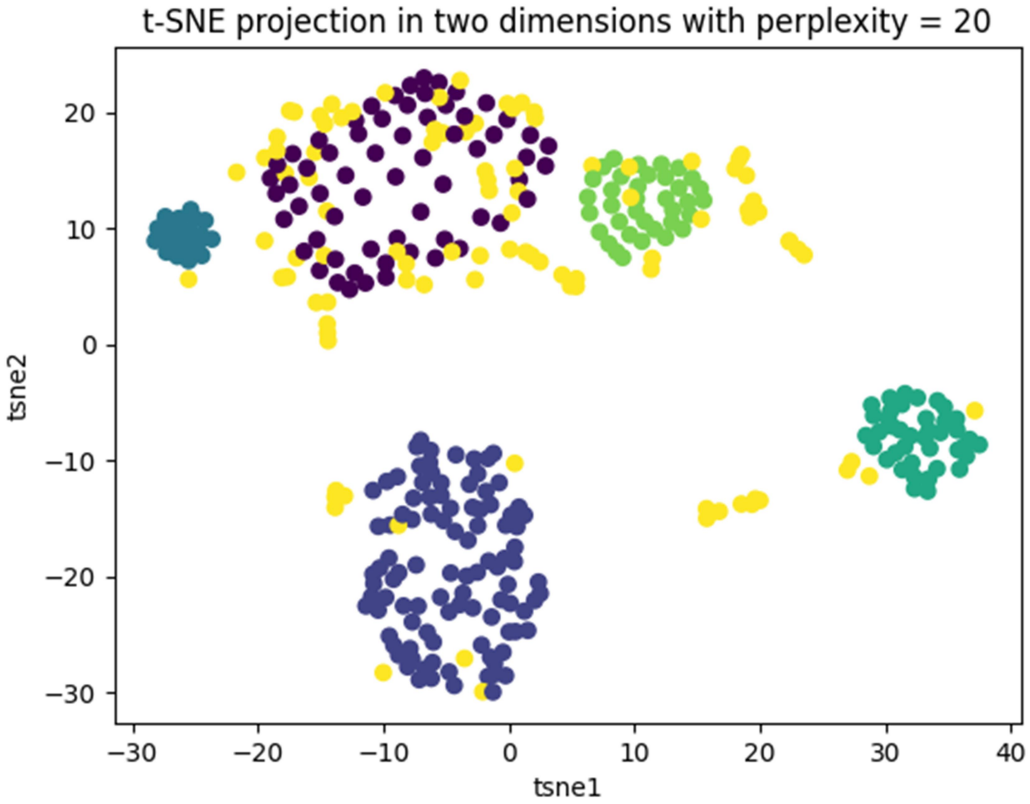
a colored scatter plot of the first two principal components

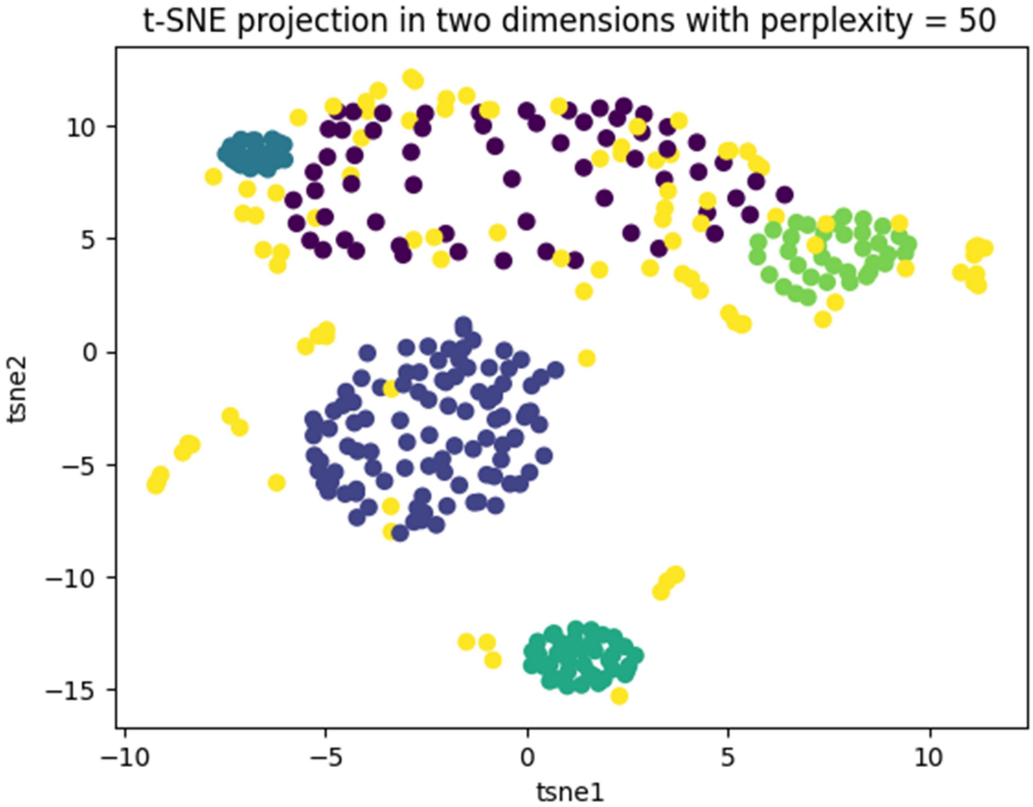


Scree plot of explained variance and no. of principal components

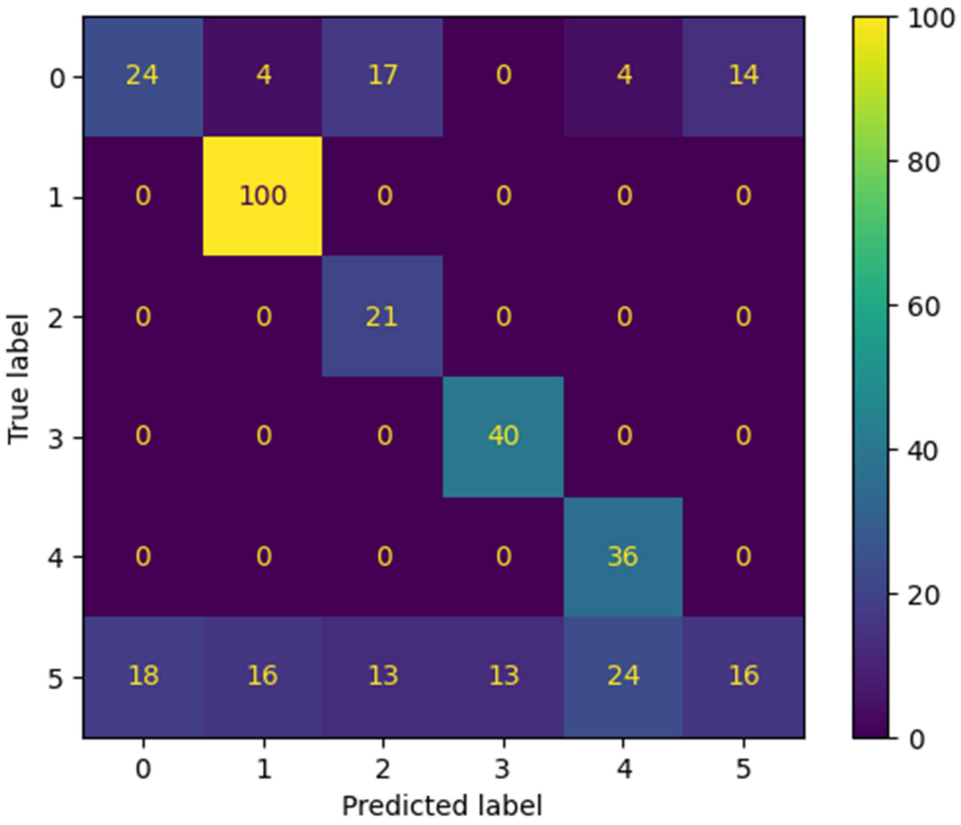


b) **t-SNE** analysed in **t-SNE.ipynb**

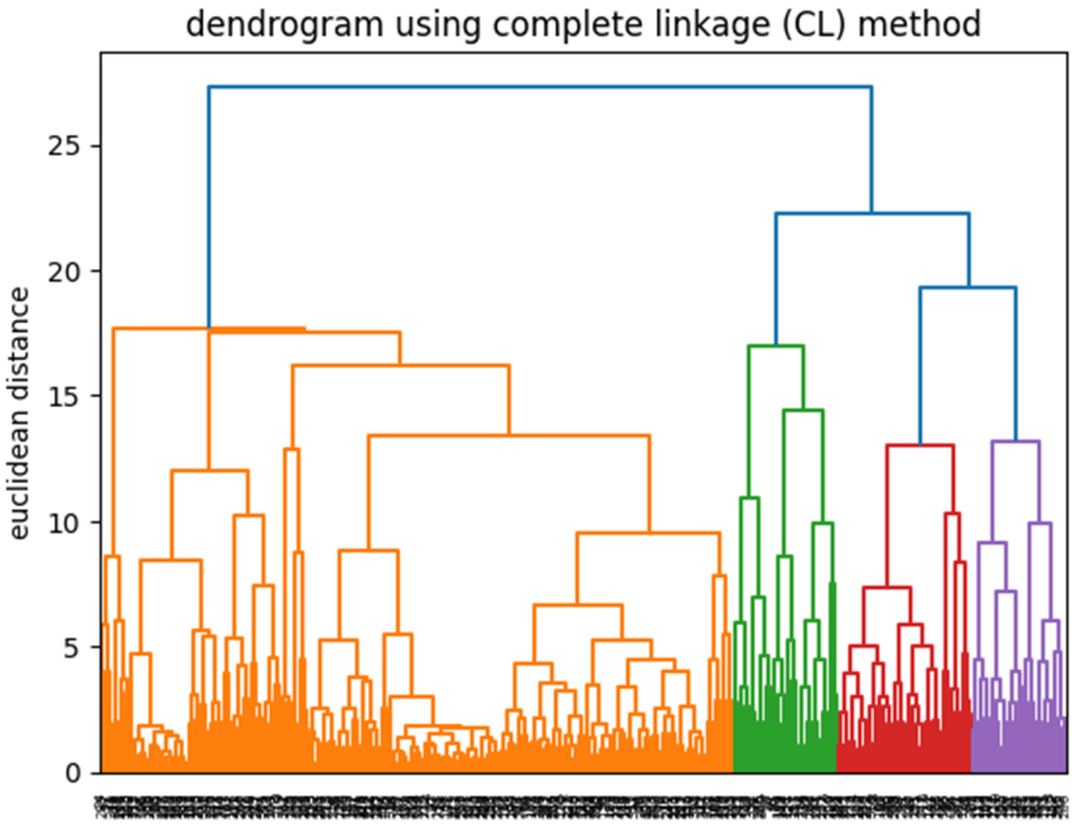


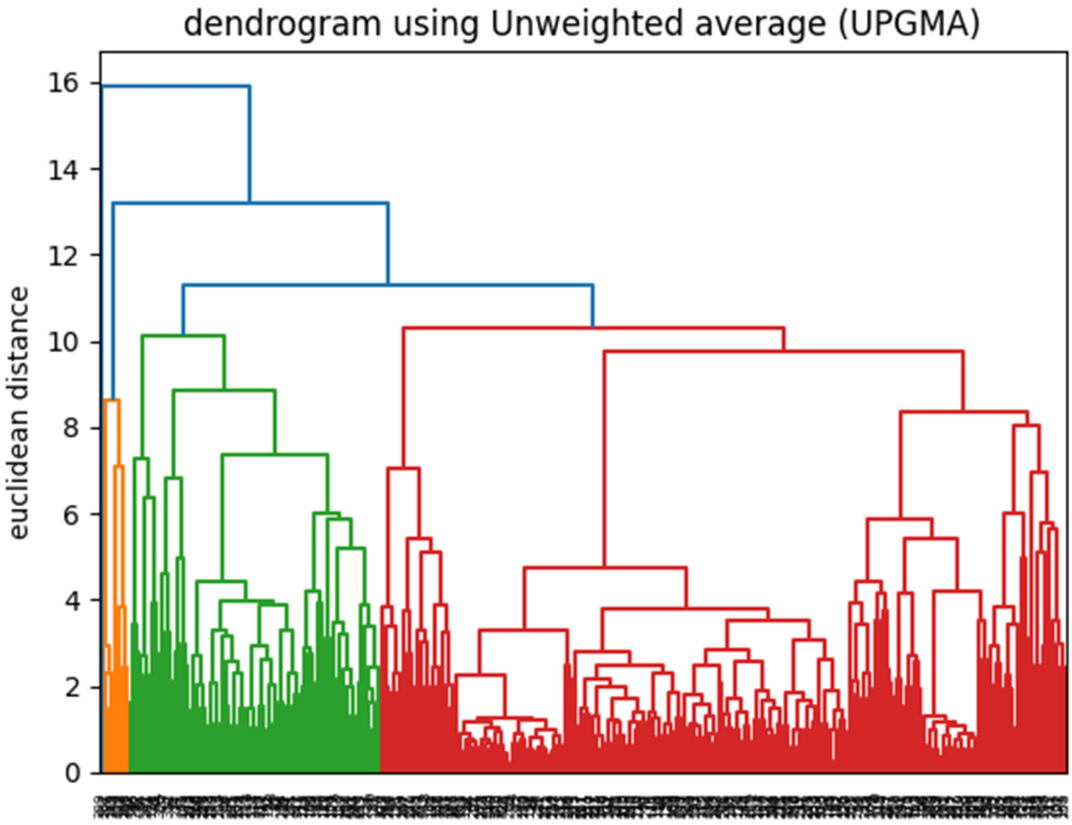


c) **k-means** analysed in **kmeans.ipynb**



d) **AHC** analysed in **AHC.ipynb**





e) SOM analyzed in **SOM.ipynb**

Obtained Transformed data of shape (n, self.n*self.m). The Euclidean distance from each item in X to each cluster center.