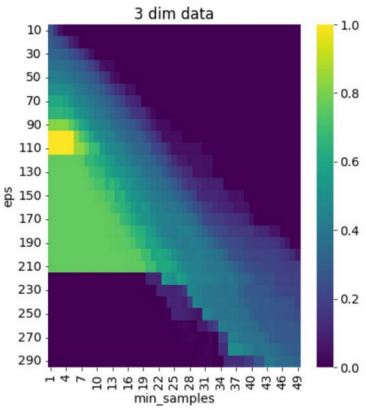
LCPB 23-24 Exercise 2, data visualization and clustering

Exercise 4A

Visualize and clusterize the data in the file **x_12d.dat** (N=600 samples, L=12 dimensions), which has also labels for checking the performances (y_12d.dat).

1. ***"eps" (ε) and "minPts" (m_P) in DBSCAN algorithm for clustering***
   Refine the grid with more values of "eps" and "minPts" and show a heat-map of the normalized mutual information (NMI) between true and predicted clusters is varying.
   The result might look like this one on the right.

   Is there a correlation between these two parameters in providing a high NMI?

   Note: in the lesson we have looked at the typical distance between a point and its closest neighbor, but this does not say what is the typical distance from the 2nd, 3rd, …, m_P-neighbor.
   The plots of ranked distances to the i-th neighbor might also help choose the ε for a given $i=m_P$.

2. **Understanding the 12-dimensional data**
   Use the PCA to visualize the first components of the data. Does it help to understand its structure?

3. **Compare different clustering methods**
   a) Perform a k-means clustering of the data, with k=3. Does it work better than DBSCAN? Why?
   b) Perform a hierarchical clustering of the data. Does it work better than DBSCAN?

4. **OPTIONAL: Visualize the data** with other methods from the scikit package