Lab 11. Hierarchical and DBSCAN clustering

The **MNIST database** is a large database of handwritten digits (0-9) that is commonly used for training various image processing systems. The original MNIST database contains 60,000 training images and 10,000 testing images. You can use the function **"`sklearn.datasets.load_digits`"** to download a subset of this data set.

1. Apply PCA and select first two directions to convert the data in to 2D.

2. Apply Agglomerative clustering with 10 clusters on the data. Plot the points in these clusters using different colors. Compare with the clustering obtained by K-means approach.
(Use **`sklearn.cluster.AgglomerativeClustering`**)

3. i) Apply DBSCAN clustering with default parameters and compare the results with K-means and Agglomerative clustering.
ii) Vary the parameter ***eps*** *(maximum distance between two samples to be considered) to 0.05, 0.5 and 0.95 and observe the results. Vary* ***min_samples*** *(The number of samples in neighborhood**)** to 1, 10, 30 and 50 and observe the results.*
(Use **`sklearn.cluster.DBSCAN`**)

4. Obtain the purity score for all the clustering methods.

```
#Purity score computation
import numpy as np
from sklearn import metrics
from scipy.optimize import linear_sum_assignment

def purity_score(y_true, y_pred):
    # compute contingency matrix (also called confusion matrix)
    contingency_matrix = metrics.cluster.contingency_matrix(y_true, y_pred)

    # Find optimal one-to-one mapping between cluster labels and true labels
    row_ind, col_ind = linear_sum_assignment(-contingency_matrix)

    # Return cluster accuracy
    return contingency_matrix[row_ind, col_ind].sum() / np.sum(contingency_matrix)
```