Exercise 2.1: Classification with Bagging and Random Forest

For the dataset used in the 1st assignment, build classification systems by applying:

1. The Bagging method with decision-tree as the base classifier, and
2. The Random Forest method.

In both methods, set the minimum number of examples per leaf to 5. Study how the generalization ability changes as you increase the number of classifiers in the ensemble to the following values: 25, 50, 75, 100. Use the out-of-bag error to measure the generalization ability. Compare the performance of the two methods with the classifiers you examined in the 1st assignment on the same dataset.

Exercise 2.2: Clustering

Use the following clustering methods:

- k-means
- Agglomerative clustering
- Spectral clustering with the RBF kernel

In the directory "clustering," there are 2D synthetic datasets. You can plot the examples to visually identify the actual number of clusters. Once you find a clustering solution, present it by plotting the examples with the same color for examples in the same cluster and different colors for different clusters.

For all datasets, use the actual number of clusters as observed from data visualization. Print the best clustering solution you find for each of the following 6 cases:

- k-means
- Agglomerative clustering (single link, average link)
- Spectral clustering (RBF kernel for various sigma values, e.g., 0.1, 0.5, 1).

For the 'gaussian_rings' dataset, identify a value of sigma for which the spectral method produces the correct clustering solution.

Provide observations about the comparative performance of the methods on the datasets.

Subsequently, for the datasets that do not contain rings, attempt to estimate the actual number of clusters using k-means as the clustering algorithm and silhouette as the evaluation criterion. Provide observations on the results obtained.