Exercise 4: Decision Trees and Random Forests

For this exercise you can decide on what learning problem you want to investigate (how many samples, independent, dependent). Remember you can explore small datasets and then validate on the complete dataset.

4.1 - Decision Trees:

4.1.1 - Compute the optimal decision point for the first 5 PCAs and compute the information gain associated to it (plot 5 graphs, one for each component, and show the highest information gain).

4.1.2 - Compute a decision tree for the digit classification and visualize it.
It can be very difficult to visualize a C5.0 tree, here it can be easier to use "rpart" and "rpart.plot".

4.1.3 - Evaluate the decision tree using cross validation. Discuss the important parameters.

4.2 - Random forests:

4.2.1 - Create a Random Forest classifier and evaluate it using cross validation.
Discuss the critical parameters (e.g., number and depth of trees)