**Fundamentals of Data Science**  
Classification

**Introduction**

The classification task is to assign to an element a label of one of the predefined classes with known representative elements. The classification can be carried out using statistical methods as well as computational intelligence.

In the first case, the most common are the algorithm of k-nearest neighbors. The point is assigned to the class that is the most frequently represented in the set of its k nearest neighbors.

To execute the classification in the MATLAB package using this method, we use the *knnclassify* script. In Python we could use sklearn.neighbors.KNeighborsClassifier:

<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

The cross-validation method is often used for testing. It consists in the division of the set into K equal parts, and then treating one of them as testing - and the others as teaching (iteratively K times - then we get K different divisions of the set into a teaching / testing set).

In MATLAB cross-validation can be performed as follows (for K = 5):

[m,n] = size(dataset);

c = cvpartition(m,'kfold',5);

for i=1:5

index\_train=training(c,i);

index\_test=test(c,i);

data\_train=dataset(index\_train,1:n-1);

data\_test=dataset(index\_test,1:n-1);

label\_train=dataset(index\_train, n);

label\_test=dataset(index\_test, n);

**here we classify testing set and calculate accuracy**end

**here we calculate mean accuracy from all folds**

The same can be achieved in Python using sklearn kFold method:

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html#sklearn-model-selection-kfold>

**To do during the lab:**

1. Please import *seeds* dataset and verify the quality of the classification using the k-nearest neighbors algorithm (using cross-validation with K=5).
2. Does the average accuracy of 5 validation stages change with alternation in the number of neighbors in the k-nearest neighbors algorithm?
3. Does changing the metric, for example Manhattan (cityblock) or cosine (cosine) influence the quality of classification?
4. Does the algorithm work better for data reduced to two dimensions?
5. For a single division into a test and teaching set, construct a classification tree. Hints: in Matlab functions to fit a tree (*fitctree*), predit using a tree (*predict*) and plot a tree (*view*) are already implemented, in Python you can use methods of *tree.DecisionTreeClassifier()*, together with *plot\_tree* to plot a tree.