Assignment 1

Machine Learning Workshop

1. [20 pts] Define each of the following machine learning terms:
   - dataset
   - training, testing, validation dataset
   - ground truth, label
   - pre-processing, feature, numerical, nominal
   - decision surface
   - model validation, accuracy, cross-validation

2. [20 pts] Pick **two** of the Scikit-learn datasets which are already included in the library (i.e. the ones with `datasets_load_`) and find out the following:
   - the number of data points
   - the number of features and their types
   - the number and name of categories (i.e. the `target` field)
   - the mean (or mode if nominal) of the first two features

3. [40 pts] Implement a correlation program from scratch to look at the correlations between the features of `Admission_Predict.csv` dataset file (not provided, you have to download it by yourself by following the instructions in the module Jupyter notebook). Display the correlation matrix where each row and column are the features, which should be an 8 by 8 matrix (should we use `'Serial no'`?). You can use pandas `DataFrame.corr()` to verify correctness of yours.

   Observe that the diagonal of this matrix should have all 1's and explain why? Since the last column can be used as the target (dependent) variable, what do you think about the correlations between all the variables? Which variable should be the most important for prediction of `'Chance of Admit'`?

4. [20 pts] Classification of mushrooms, edible or poisonous. Download the `assignment01_mushroom_dataset.csv` dataset file from the module content. Load the data set in your model development framework, examine the features to see they are all nominal features. The first column is the class which represents the mushroom is poisonous or not. Apply necessary pre-processing such as nominal to numerical conversions. Make sure sanity check the pipeline and perhaps run your favorite baseline classifier first.

   Report the performance of your classifier.