Preparation for Coding Assignment I

- Prepare your own personal computer (laptop, desktop, whatever)
- Have python 3.7x installed
- Have the following packages installed:
  - Numpy
  - Pandas
  - Scikit-learn
- Download the data set from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

Coding Assignment I

- **Task:** Use k-NN to classify whether the patients' tumors are benign (เนื้องอกปกติ) or malignant (เนื้องอกร้ายแรง)
- **Instruction:**
  - For this very first assignment, you are recommended to follow the tutorial from course year 2020 at https://www.youtube.com/watch?v=w5-fwTLzNcI.
  - For the assignment, you are asked to implement function "train_test_split" and module "neighbors" on your own. This means you will have to make the code working without any modification in cells 2-6 in the attached file coding-assignment-1-NN.ipynb.
- **Expected accuracy:** > 94%

Coding Assignment I: Data Set Information

- Number of Instances: 699
- Missing attribute values: 16
- Class distribution:
  - Benign: 458 (65.5%)
  - Malignant: 241 (34.5%)

Coding Assignment I: Attribute information

- Clump Thickness (1-10)
- Uniformity of Cell Size (1 - 10)
- Uniformity of Cell Shape (1 - 10)
- Marginal Adhesion (1 - 10)
- Single Epithelial Cell Size (1 - 10)
- Bare Nuclei (1 - 10)
- Bland Chromatin (1 - 10)
- Normal Nucleoli (1 - 10)
- Mitoses (1 - 10)
- Class: 2 (benign) / 4 (malignant)

Summary of Supervised Learning

1. Formulate your supervised learning problem.
    1. Define your feature space.
    2. Define your label space.
2. Collect your data (forming your data set) and split it into train and test parts.
3. Choose the right ML algorithm and train it with your training data.
4. Evaluate the prediction results due to the trained program via test data (evaluate the testing loss)
5. The testing loss indicates the actual accuracy of the trained program in the real use.