# CPS 844 Lab 5: Metrics, k-Fold Cross-Validation, and Nearest Neighbors Classification

This lab experiments with nearest-neighbour classification, and it makes use of the breast cancer sample data from lab2.
You are invited to 'recycle' the code from lab2 in order to preprocess the data. The solution was posted on D2L, and was titled 'lab2 - solutionV2.py'.

Write a Python script that performs the tasks described below. Submit the .py file on D2L. Please note that if you submit your file in some other format besides .py or (.txt should you meet an issue), then your mark will at most be 60%.

Most of the evaluation of the nearest neighbors classifier in this lab are done with ten-fold cross-validation. Indeed, k = 10 is one of the most popular value to evaluate models using k-Fold Cross-Validation.

Part 1:

1. (0 points) Download the dataset:

   https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data

2. For your own understanding of the dataset, read the description file:

   https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names

3. (10 points) 'Recycle' the code from lab2 to pre-process the data: assign new headers to the DataFrame, drop the 'Sample code number' attribute, convert the '?' to NaN, discard the data points that contain missing values, convert the values for the 'Bare Nuclei' attribute to numerical values, drop row duplicates.
4. (15 points) Continue with the preprocessing: separate the features from the target class, standardize the features. Modify the target values: from the description, the malignant class labels are indicated with the value 4. The 'benign' labels are indicated with the value '2'. Replace (or 'map') the values such as the label '4' becomes the integer '1', and the label '2' becomes the integer '0'.

Part 2:

5. (5 points) Use the sklearn library to construct a NearestNeighbors classifier. Keep the default value for the number of neighbors (which is 5).
6. (40 points) Compute and print out the averages of the accuracies, f1-scores, precisions and recall measurements of the nearest neighbor classifier, using 10-fold cross validation. Hint: start by looking into the documentation for the function 'cross_val_score' of

sklearn.model_selection. You will need to look at different pages until you find all the information you will need.

7. (30 points) The goal here is to create and display one confusion matrix. For that, you can create a training and test set from your pre-processed data, train the nearest neighbor classifier on the training set, and predict the labels of the test set using the trained classifier. Then summarize the prediction results using a confusion matrix. As a reminder, the confusion matrix will summarize the number of correct and incorrect predictions, broken down by each class.

   Hints:

   - There are several methods available for creating and displaying the confusion matrix, and you may need to upgrade the version of sklearn depending on the one you choose.
   - Choose as you wish the proportion of the dataset to include in the test split.