# Machine Learning 2018
## Classification & Performance Estimation and Regularization
## Homework
## Deadline: 03/09/2018

### Classification

Choose a dataset of your choice and predict a variable of interest.

1. Use logistic regression.
    a) Which variables seem to be more important for the classification task? How do you conclude that?
    b) Assess the performance of your model by plotting the AUC and examining the confusion matrix. Does your model have higher sensitivity of specificity?
    c) Try to fit a model using only the most important variables that you identified in step a). What difference do you see in the performance on the test set? How can you explain it?

2. Use K-Nearest neighbors. Experiment with values of k and access the performance of your model using the confusion matrix. Which is the best k you could find?

Which method performs better on your data? Can you guess why?

### Performance Estimation and Regularization

Using the dataset you chose for the linear regression assignment, try to fit a model with regularization. (You can also choose a different dataset, it's up to you)

1. Use Ridge Regression
2. Use Lasso

- For each method, use cross validation to find the best lambda and calculate the MSE for the optimal lambda on the test set.
- Do you see an improvement in the MSE compared to non-regularized linear regression? Why this may be happening?
- After using lasso regularization, which variables are still important for prediction?

### General comments

**1** You can use whichever programming language you prefer, but please add comments to your code so that I can understand what you are doing.

**2** Please, include your name in the name of the files you send me so that I can distinguish which homework belongs to whom more easily.

**3** You can email me at sofia.nomikou@nyumc.org if you have any questions. Enjoy!