Homework 3

For this assignment, you are free to abandon the Titanic dataset (if you want) and start exploring new datasets that are available from various resources. You can start simply by googling "machine learning datasets" or check out these two resources:

http://archive.ics.uci.edu/ml/ https://www.kaggle.com/datasets

This will also prepare you for the final project where you will work on a dataset of your choice.

Remember that you are looking for a classification (rather than regression) task. When you have selected the dataset, download it and think about whether you need to do some pre-processing (e.g. convert categorical attributes to numerical values, do the standartization/normalization). Now split it into training (70%), development (15%) and test (15%) sets. Finally, think about what the most appropriate evaluation metric is for your dataset (e.g. accuracy vs. F1 score).

Here are the specifics of what needs to be done:

1. Use either logistic regression or SVM implementations in scikit-learn to train a classifier. Use the default classifier parameters. Evaluate your classifier on the development set.
2. Now try to tweak the classifier parameters and see if you can improve your model's performance on the development set.
3. Implement your own k-nearest neighbors classifier (KNN). Now try to tune the k value to achive the best possible performance on the development set.
4. Compare your best model you built in step (2) to your best KNN model by evaluating them on the **test** set.

Each problem is worth 25 points.

The ability to summarize your findings is extremely important when doing research or working as a data scientist. Please summarize your findings in a 1-page paper. Include the details on what classifier you used, how you chose the best model parameters, what these parameters were, and how your best models (e.g. SVM vs. KNN) stacked up against each other. Please see the general guidelines for homework submission in my Box folder.