Homework 2: SNF prediction

After having a total joint surgery (either a total hip or total knee replacement), some patients must be discharged to a Skilled Nursing Facility (SNF) for recovery. These SNFs are very expensive and can be quite disruptive to patient's lives. The goal of this project is to develop a model that can predict, before the surgery, where the patient will be discharged to. This is a **binary prediction task**: will the patient go to a snf or not.

On bluehive in the folder */scratch/dsc381_2019/Homework_2_files/* I have placed the dataset you will use for this project. It is named "snf.csv". This file contains real medical data that has been de-identified. The outcome is in the column *INDEX_DISCH_DISP_NM*. Please use your best judgement to figure out what the other column names mean, they should be pretty explicit in most cases.

First **preprocess the data as discussed in lecture**. I have included several files in the folder */scratch/dsc381_2019/Homework_2_files/* that can help you preprocess the diagnosis codes. These include "mappings.csv" which maps icd9 to icd10 codes and vice-versa, "hcupccs-9.csv" which is a grouper for icd-9 codes, and "hcupccs-10.csv" which is a grouper for icd-10 codes. Remember to split the data into your 3 sets (training, validation, and testing) before you train any models. You should use a 70-15-15 split, do not use cross-validation.

As part of the preprocessing, you should **construct three new variables** based on the column *ProcName1*. The first should indicate which side the surgery was performed on (left, right, both, or unknown). The second should indicate if the surgery was on the knee or the hip. The third should indicate if the anterior approach was used. You can assume based on domain knowledge (me telling you right now) that these three factors will all be important. If you would like to construct additional variables from the information present in any of the columns feel free to do so, but it is not required. If you think any of the variables should not be used, you may remove them, but explicitly explain your reasoning in your README.

Once preprocessing is finished, please **use sklearn to run a logistic regression, an SVM, and a random forest** to predict SNF placement. Remember to use the anaconda environment created for the class. For this project you **DO NOT need to tune model hyperparameters**, you may just use the default values.

All preprocessing and modeling should be done in python, on bluehive, using the class environment. Please write up a README that includes a description of how you preprocessed (and split) the data, and run time instructions for how to train and test the models you developed. Please also include accuracy, precision, and recall for each of your models, as calculated on the testing set. Please create a zip file containing your preprocessing code, your model training/inference code, your README, and your training/validation/test datasets. Your code should all be appropriately commented. Submit this zip file using the homework submission script on bluehive. You must submit your files by 5pm on Wednesday 10/2.