Problem 1) We will create a kfold CV pipeline for the heart disease dataset you worked with during the midterm exam.

Problem 1a) Read in the data and drop the rows with missing values. Remember that the dataset does not contain the feature names! Separate out the feature matrix (X) and the target variable (y). What is the balance of this dataset (the baseline accuracy)? (2 points)

In [ ]:

Problem 1b) ML pipeline with logistic regression

Split your data in a stratified manner into `other` and `test` (20% in `test`) and then split `other` into 5 stratified folds. 4 of those folds will be used for training, the last fold will be CV. You'll need to loop through the 5 options the CV fold can be selected. (4 points)

Preprocess the data. Apply the OneHotEncoder and the StandardScaler to the appropriate columns. Make sure to fit_transform only train (4 out of the 5 folds). The CV and test sets should be transformed based on the preprocessor fitted to train. (2 points)

Train a logistic regression model with l1 regularization and tune the appropriate parameter. (4 points)

Repeat the procedure 10 times with 10 different random states and print the mean and std of the test accuracy score. Make sure to print the best parameters and check that the best values are not at the edge of your parameter space if possible. Check that your code is reproducable! That is, if you rerun the cell, you get back the exact same result. (3 points)

In [ ]:

Problem 1d) Train a random forest classifier and tune the appropriate parameters. Make sure to print the best parameters and check that the best values are not at the edge of your parameter space if possible. Repeat the procedure 10 times with 10 different random states and print the mean and std of the test accuracy score. Check that your code is reproducable! That is, if you rerun the cell, you get back the exact same result. (5 points)

In [ ]:

Problem 1e) Train an SVC and tune the appropriate parameters. Make sure to print the best parameters and check that the best values are not at the edge of your parameter space if possible. Repeat the procedure 10 times with 10 different random states and print the mean and std of the test accuracy score. Check that your code is reproducable! That is, if you rerun the cell, you get back the exact same result. (4 points)

In [ ]:

Problem 1f) Compare the means and standard deviations of the three techniques. How many standard deviations above the baseline accuracy are the three models? How would you rank them with respect of accuracy? (3 points)

In [ ]: