**Lab 3: Programming Assignment**

**Predicting Drug Response in Tumor Cell Lines**

**Due Monday March 11th at Midnight.**

For each of these problems, use 5-fold cross validation and display the confusion matrix, f1 scores and other evaluation metrics as appropriate in addition to the accuracy or mae achieved.

Data for this problem includes the following:

1. Transcriptomes (RNAseq), and Genome Variation (SNPs) characterizing the cell lines from various studies (GDSC, CCLE, gCSI, CTRP, NCI60)
2. Combined “dose response” data, including drug, cell lines, dose, growth values.
3. Aggregated response data, drug, cell lines, computed AUC, etc.
4. Drug related data, Dragon7 descriptors, ECFP, PFP, SMILES, etc.
5. ALMANAC study including dose response for drug pairs

They can be found in the shared **MLiC-Datasets-Lab3** directory.

For each part please turn in your code (a python notebook is a reasonable way, but a python script is also fine), turn in output from the program (text of graphics, graphics preferred). And a short (1 paragraph write up for each part and section, explaining what you did and your critique of the results, comments on problems or difficulties and possible future approaches that might do better).

**Part 1. “By Drug” Tumor Dose Response**

1. Using SciKit Learn build a machine learning classifier that predicts the dose response for tumors for each of 10 drugs selected by you. This model should be trained on data from one study and validated on the same study. Sweep through each study. Sweep through multiple ML methods.
2. Using model selection methods of your choice determine which classical ML method (try at least four methods) performs best on the by drug dose response problem.
3. Using feature selection methods of your choice determine a < 100 gene (RNA-seq) signature that can be used to predict dose response for each of the drugs. Determine how many genes in the compact signature are in common between your selected drugs.
4. Using **Keras**, build a deep learning classifier that performs the same regression task, and determine the learning curve (relationship of number of training samples to prediction accuracy) for your network, recommend using at least 10 training set sizes to estimate the learning curve.
5. Extra Credit: Use the TPOT autoML system to search for a better solution to part a.

Part 2. **Dose independent formulation.**

1. Using SciKit Learn build a machine learning regresser/classifier that predicts drug response using the aggregated dataset. Use AUC1 as the label. Threshold AUC1 (> 0.50 = 0, <= 0.50 = 1) for the prediction target in the classification case. Input data should include tumor features and drug features. (hint, use RNA-seq and Dragon7 descriptors). Try this on four studies (e.g. CTRP, GDSC, CCLE, gCSI)
2. Using model selection methods of your choice, determine which classical ML method (try at least four methods) performs best.
3. Using feature selection methods of your choice, determine the 100 most important input features that can be used to classify drug response. Determine if drug features or tumor features have more predictive power.
4. Using **Keras**, build a deep learning classifier that performs the same classification task, and determine the learning curve (relationship of number of training samples to prediction accuracy) for your network.
5. Extra Credit 1: Use each study (e.g. NCI60, CTRP, CCLE, GDSC, gCSI) in kind to train and validate on a different study. This is known as the cross study validation problem.
6. Extra Credit 2: Explore the use of different features to represent the drugs, try drug descriptors, fingerprints or some other representation. Determine which types of features are more predictive.