[100 points] In this question, you will handle a data set that contains thousands of features. Your task is to primarily perform feature selection using L1 (LASSO) regularized logistic regression models. Next, you will attempt to use only the selected features models to tune other models: Multilayer Perceptron, and K-Nearest Neighbor. It's possible the accuracy may improve.

The given data set (**big.data2_classes.txt**) contains expression levels of 34,260 proteins (columns as features) in 549 human cells (rows as samples). The response column (last column labeled "class") has 2 classes representing 2 different types of cells. Your task is to decrease the number of features to accurately differentiate the 2 different cell types. The two classes have approximately equal number of samples.

Below is a brief outline of the analytics pipeline that you fill follow. A good practice is to save the results (to files) after each step so that you do not have to re-run everything from scratch.

a. Split the data into training and testing set. [5 points]

b. **In the training set**, compute the maximum value for each feature. This represents the minimum expression level of a protein across all the cells. Proteins with low expression level can be removed since they are not biologically important. Hence, remove features with maximum value of less than 2. You will probably retain about 26,000 proteins (features).
   In the testing set, remove the same set of features**.**
   [Note: df.max() gives maximum value in each column] [10 points]

c. In the training set, compute the variance for each feature. Retain the top 5000 features with the highest variance.
   [Note: df.var() gives maximum value in each column]
   In the testing set, retain the same set of features**.** [10 points]

d. Scale training and testing data. [5 points]

e. Perform 5-fold cross validation resampling method to find the best **shrinkage parameter** for L1 (LASSO) and L2 (Ridge) logistic regression on the scaled training set. You can use 'accuracy' as the metric. Select the best L1 model based on the best mean validation accuracy score from 5-fold cross validation. [15 points]

f. From best LASSO regularized model in (e), identify the selected features in the model. [5 points]

g. Using only the features from (f), train a Multilayer Perceptron (MLP) using 5-fold cross validation. You are free to choose what parameters to tune. Also train a K-nearest neighbor (KNN) model using 5-fold cross validation. [20 points]

h. Retrain L1 best model, KNN, and MLP best model on the entire training set with selected features from (f) [10 points]

i. Retrain L2 best model on the entire training set. [10 points]

j. Report accuracies on entire training set and testing set from L1, L2, KNN, and MLP model. [10 points]