Online Applied Machine Learning
Final Project
(Due April 30, 2017 @ 11:59PM)

For the final project, you will write a report (**maximum 3 pages**) wherein which you pose an empirical question, and then take steps to answer it. The question you ask should be based on the work you have done for programming assignment #1 and for programming assignment #2. Make sure to clearly state the question you are posing.

At minimum, you much include an analysis that discusses and evaluates the performance of each of the seven algorithms on the Adult data set from the UCI repository (http://archive.ics.uci.edu/ml/datasets/Adult). The prediction task associated with this data set is to predict whether or not a person makes more than $50K a year using census data. The seven algorithms are: Naïve Bayes, Logistic Regression, Decision Tree, k-NN, SVM, Random Forest, and Adaboost.

As in programming assignments #1 and #2, you will need to report any preprocessing you did on the data (ie. did you discretize the features? If so, how?), the performance of the learned model on the training set and on the test set (this is how we can check for overfitting), and why you chose to evaluate the model's performance the way you did. Conduct a more in-depth analysis comparing why and how the classifiers performed the way they did. Make sure to talk about each classifier with respect each of the other classifiers.

Additional steps you can take include:

1. Run and evaluate a subset of the algorithms on another data set of your choosing;
2. Evaluate a larger range of pre-preprocessing techniques to gain a more in-depth understanding of the different techniques and how they impact the performance of each classifier;
3. Make a modification to one of the learning algorithms that should improve the performance of the algorithm;
4. Compare the performance of the Weka algorithms with the SKLearn algorithms;
5. Implement your own version of a subset of the classifiers and compare its performance to that of either the Weka or the SKLean implementations;
6. Etc….

This is not a complete list. If you have another approach you would like to take, please contact me and we can discuss whether or not it would satisfy the needs of this project.

You can download the data set from http://archive.ics.uci.edu/ml/machine-learning-databases/adult/ and find the ReadMe for the data set at http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names . In the ReadMe you can find current performance results of a number of different classification algorithms. This should give you a point of comparison for your own investigation or some ideas to get your creative juices going.