CS 345: Data Mining Names:_____

Assignment 4 _____

For this exercise, you will be using 3 different data sets: `vote.arff`, `colic.arff`, and `zoo.arff`. Vote.arff is already in your data directory from when you downloaded WEKA. The other two files are available for download on the course website under today's homework.

Do the following with `vote.arff` first and then repeat the steps using `colic.arff` and then `zoo.arff`.

1. Load the data set into WEKA. Go to the Classify tab. Select classifiers.bayes.NaiveBayes. Run it using 10-fold cross-validation (use 10-fold cross validation for all tests). Record the error rate (Incorrectly Classified Instances %) in the correct table under NaiveBayes, Single.
2. Now use the J48 decision tree algorithm to classify. (classifiers.trees.J48) Record the error rate in the correct table under J48, Single.
3. Bagging and AdaboostM1 are available under the "Meta" category in WEKA. Please use the following settings and record the error rate in the correct table:
   - Bagging: set `numIterations` to 30. You will run experiments with the classifier set to trees.J48 and bayes.naiveBayes.
   - AdaboostM1: set `numIterations` to 30. Set `weightThreshold` to 1000. You will run experiments with the classifier set to the same two algorithms as for Bagging.

vote

Base Learner	**Single**	**Bagging**	**Boosting**
J48			
NaiveBayes			

colic

Base Learner	**Single**	**Bagging**	**Boosting**
J48			
NaiveBayes			

zoo

Base Learner	**Single**	**Bagging**	**Boosting**
J48			
NaiveBayes			

4. Answer the following questions:

    a. Which algorithms+data sets are improved by Bagging?

    b. Which algorithms+data sets are improved by Boosting?

    c. Does Boosting ever hurt performance? If so, try to provide a possible explanation.