# Math 5364 Homework 3

1. What are the variables in the `kyphosis` data set in R, and what do they represent? (Hint: Use the `help` command.)

2. Use the `rpart` function to construct a decision tree for predicting kyphosis based on the other variables in the data set.

   (a) Plot the tree.

   (b) Obtain the confusion matrix.

   (c) Calculate the accuracy and error rate for the tree.

3. Based on these results, what are the two most important variables for predicting kyphosis? Create a scatterplot of these two variables, coloring the points in the plot different colors according to the presence/absence of kyphosis.

4. Repeat problem 2 using the `ctree` command, and use the "simple" option for plotting the tree.

5. Compare/contrast the resulting trees, confusion matrices, and accuracy/error rates using the `rpart` and `ctree` commands. Does either method appear to be much better than the other?

6. Calculate the entropy for the kyphosis data overall, and then calculate the weighted entropy for the trees obtained using the `rpart` and `ctree` commands.

7. Page 28 of the lecture notes shows a tree obtained from `rpart` for the iris data set, and the split point for petal length is 2.4. Then, page 38 has a tree obtained using `ctree` with a petal length split at 1.9. Is this a serious difference between these two trees? To address this question, calculate the largest petal length for a setosa flower and the smallest petal length for a versicolor flower. You should be able to do this using R commands without visually inspecting the data. In other words, your solution should work just as well for a data set containing a million rows.