Assignment 2 (due October 15) (4 points)

In this assignment you are given a multi-attribute census dataset that contains both categorical and numeric attributes. You will explore a few issues: (i) whether more features would necessarily give better test accuracy; and (ii) whether more training data would necessarily give better test accuracy.

1- Load the Census training dataset.
2- Create a decision tree, called F5, using the first five attributes, i.e., from *age* to *education-number*. The two classes are individuals who make more or less than $50K.
3- Create a decision tree, called F10, using the first ten attributes, i.e., from *age* to *sex*.
4- Create a third decision tree, called F14, using all the fourteen attributes.
5- Now repeat steps 2 to 4 above by using only half of the training data, i.e., the first 50% of the training data. Call the respective decision trees H5, H10 and H14 ("H" as in "Half").
6- Load the Census test set.
7- Compute the sensitivity and specificity of F5, F10, F14, H5, H10 and H14 based on the test set.
8- **Hand in:** For each of the 6 trees, show the root split point, the sensitivity and specificity on the test set. Comment on the accuracy of the 6 trees based on their sensitivity and specificity. Comment on whether more features and/or more training data would necessarily give better test accuracy.