Problem Set #6

MACS 30100, Dr. Evans
Due Monday, Feb. 26 at 11:30am

1. **Decision trees (10 points).** Joe Biden was the 47th Vice President of the United States. He was the subject of many memes, attracted the attention of Leslie Knope (Parks and Recreation, TV sitcom), and experienced a brief surge in attention due to photos from his youth. The data file `biden.csv` contains a selection of variables from the 2008 American National Election Studies survey that allow you to test competing factors that may influence attitudes towards Joe Biden. The variables are coded as follows:

   - `biden`: feeling thermometer ranging from 0 to 100. Feeling thermometers are a common metric in survey research used to gauge attitudes or feelings of warmth towards individuals and institutions. They range from 0-100, with 0 indicating extreme coldness and 100 indicating extreme warmth.
   - `female`: =1 if respondent is female, =0 if respondent is male
   - `age`: age of respondent in years, range from 18 to 93
   - `dem`: =1 if respondent is a Democrat, =0 otherwise
   - `rep`: =1 if respondent is a Republican, =0 otherwise
   - `educ`: number of years of formal education completed by respondent, range from 0 to 17 with 17+ representing the first year of grad school and up.

   (a) Split the data into a training set (70%) and a test set (30%). Be sure to set your seed prior to this part of your code to guarantee reproducibility of results. Use recursive binary splitting to fit a decision tree to the training data, with `biden` as the response variable and the other variables as predictors. Plot the tree and interpret the results. What is the test MSE?

   (b) Leave the control options for tree() at their default values. Now fit another tree to the training data with the following control options: tree(control = tree.control(nobs = # number of rows in the training set, mindev = 0)) Use cross-validation to determine the optimal level of tree complexity, plot the optimal tree, and interpret the results. Does pruning the tree improve the test MSE?

   (c) Use the bagging approach to estimate a tree to create a model for predicting `biden`. What test MSE do you obtain? Obtain variable importance measures and interpret the results.

   (d) Use the random forest approach to estimate a tree to create a model for predicting `biden`. Do this for $m = 1$, $m = 2$, and $m = 3$ (the number of variables). What test MSE do you obtain in each case? Obtain variable importance measures and interpret the results. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.