1) A. = **FALSE**

   Even though the technique leads to selection of about 63% of the observations from the original data, you are not directly going in and taking a 63% sample of the data.

   B. = **TRUE**

   (Slides 3 – 4 of Dr. Race's notes). See explanation to A above.

   C. **TRUE**

   (Slides 5 – 11 of Dr. Race's notes). Bagging (**b**ootstrap **agg**regat**ing**) is basically where you take a whole bunch of bootstrap samples of the same training data to train classifiers, thus creating many different models. These classifiers can then be ensembled together (majority voting essentially) to give you the final result.

   D. = **False**

   (Slide 12 of Dr. Race's notes). Dr. Race said in class that bagging is method used to reduce overfitting (higher variance) but only if the model was suffering from high variance to start off with. If the model was NOT suffering from high variance to start off with, then you could make the model worse by using this technique.

   E. = **True**

   (Slide 14 of Dr. Race's notes). Random forests are just basically a whole bunch of decision or regression trees ensembled together. This technique works best when the trees are each finding different patterns in the data rather than finding/working on the same patterns (correlated trees).

   F. = **False**

   (Slide 16 of Dr. Race's notes) Because random forests are a bunch of trees ensembled together, the final relationships between "the input variables" and the "target" is not easy to interpret. There is no interpretability in the final model aside from variable importance.

   G. **False**

   (Slides 18 – 22 of Dr. Race's notes). After making the first model, the technique looks to see which observations were predicted incorrectly (because they were harder to classify). It then places more emphasis on these observations that were predicted incorrectly in an attempt to better classify them. The process keeps repeating in this manner.

   H. = **TRUE**

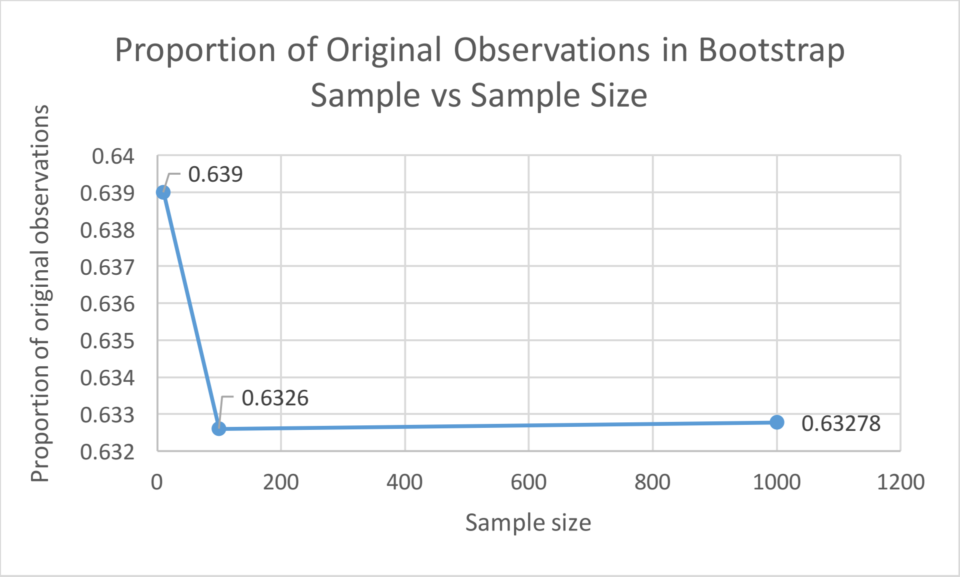
   (Slide 35 of Dr. Race's notes).

   I. = **TRUE**

   (Slides 30 – 35 of Dr. Race's notes).

2) Only a randomly selected subset of the features or variables need to be considered at each split. You are doing this because you are trying to avoid getting the trees from identifying the same patterns. You want the different trees to model different patterns in the data and by getting each of the trees to focus on different variables, you can try and get the trees to find different patterns. (Slides 14-15 of Dr. Race's notes).

3) Yes as sample size increases, the proportion of original observations in the bootstrap sample did get closer to 63%. Find chart below on one example of a simulation.


Proportion of Original Observations in Bootstrap Sample vs Sample Size
Proportion of original observations
0.64
0.639
0.638
0.637
0.636
0.635
0.634
0.633
0.632
0.639
0.6326
0.63278
0
200
400
600
800
1000
1200
Sample size