

## Announcements

Make sure to sign in on the [google form](#) (I send a list of which section questions are useful for which pset questions afterwards)



Final project due 12/12.

## Introductions

- One question or thought related to lecture last week (decision trees, bagging, random forests)

## The Entire Arboretum

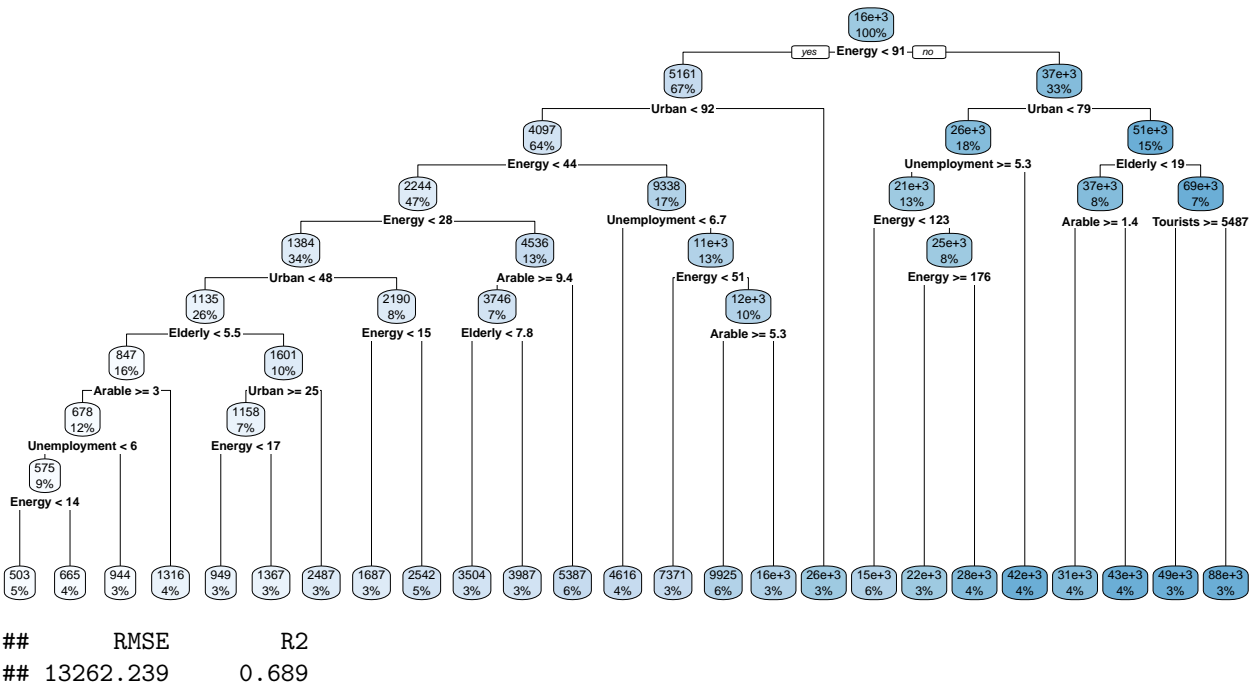
The packages `rpart` and `randomForest` will contain most of the functions for decision trees and random forests. The most important functions for us will be:

- `rpart(formula, data, control)` with `control` as `list(minsplit=1, cp=0, maxdepth=20)`, where `minsplit` says we want the node to be split if it has at least 1 observation; `cp` (complexity parameter) says we will accept any improvement in fit; and `maxdepth` is the maximum node depth.
- `prune(tree, cp)` where `cp` is the complexity parameter.
- `randomForest(formula, data, maxnodes, mtry, ntree)` where `mtry` is the number of variables randomly sampled as candidates at each split, and `ntree` is the number of trees to grow.

These problems will deal with a dataset of country-level statistics from [UNdata](#), [Varieties of Democracy](#), and the [World Bank](#).

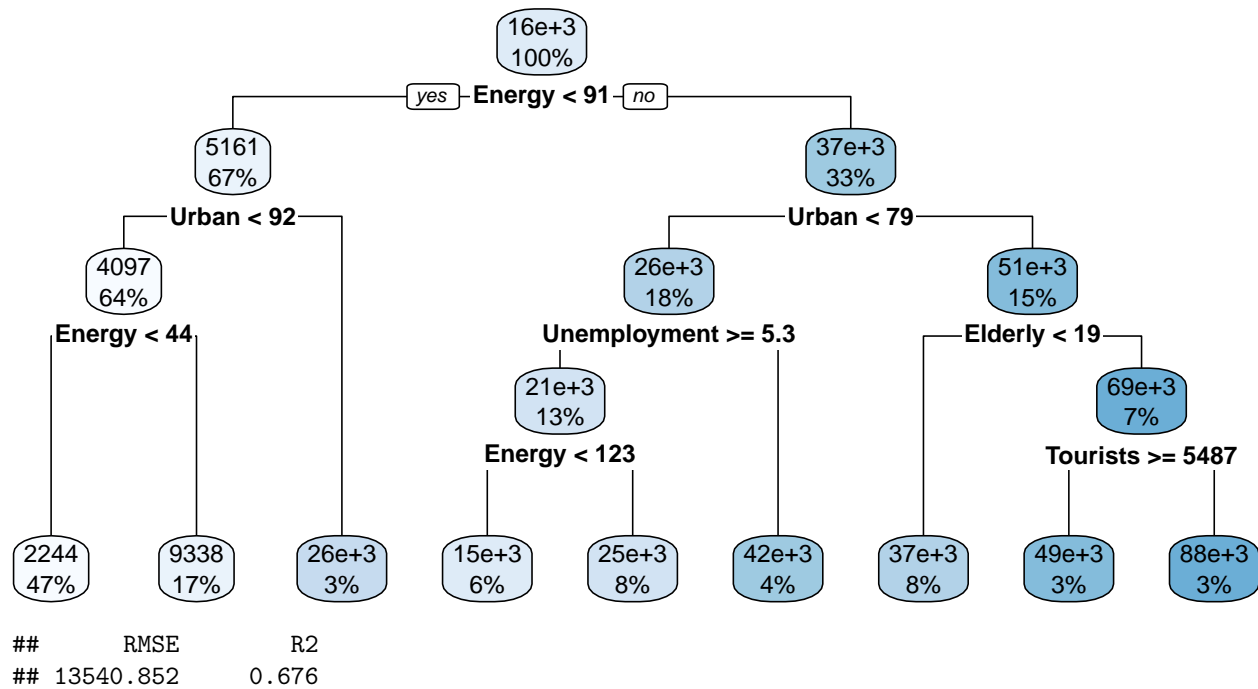
1. The following is a fit decision tree to predict GDP per capita along with RMSE and  $R^2$  statistics for the tree's predictions relative to the true values. The decision tree uses `minsplit=1`, `cp=0`, `maxdepth=20` on the variables:
  - **Urban**: Percent of people living in urban areas
  - **Elderly**: Percent of people over the age of 60
  - **Arable**: Percent of total land area that is farmable
  - **Energy**: Gigajoules of energy produced per person
  - **Unemployment**: Unemployment rate as a percent
  - **Tourists**: Thousands of tourist/visitor arrivals

The 2010 US values for these variables are shown below. What is the US's estimated GDP per capita?

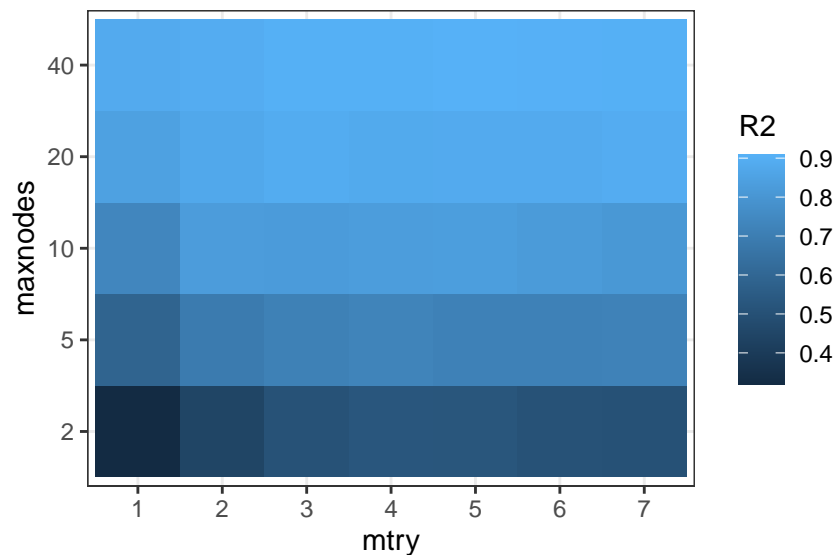


Urban	Elderly	Arable	Energy	Unemployment	Tourists
80.8	18.3	17.2	301	9.6	60010

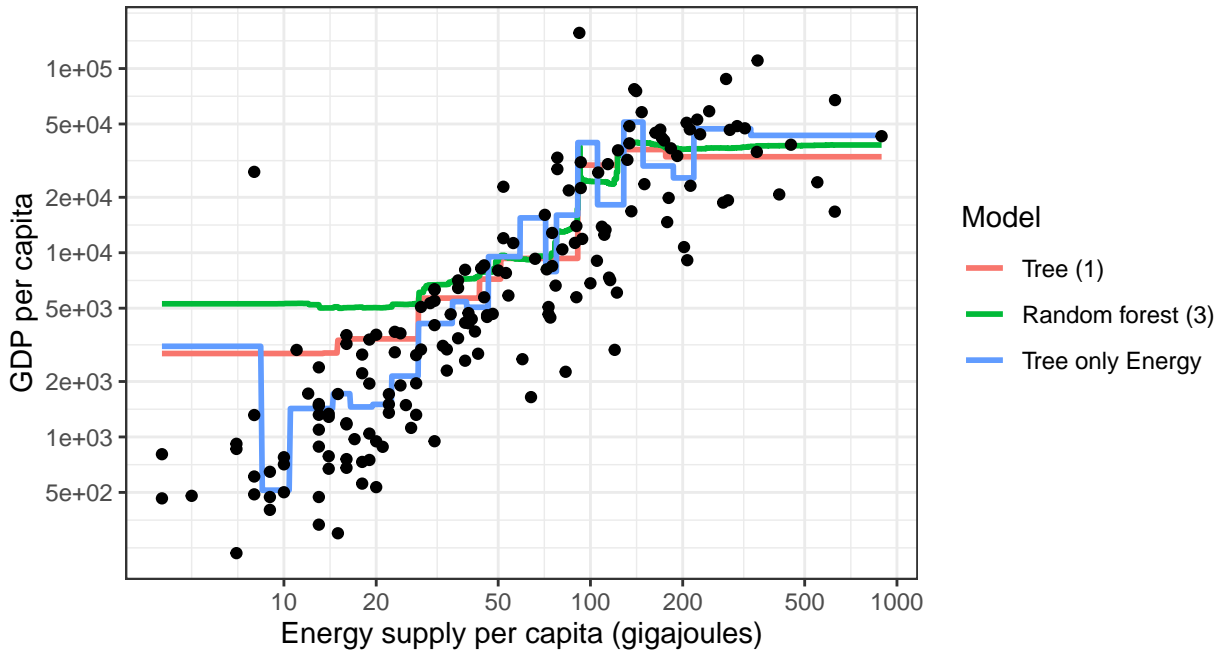
2. The following is a well-pruned regression tree to predict GDP per capita from the predictors above. The model starts with `minsplit=1`, `cp=0`, `maxdepth=20` and uses 10-fold cross-validation to prune based on the complexity parameter `cp`. Interpret the differences between the previous tree and the pruned one. What is the US's estimated GDP per capita?



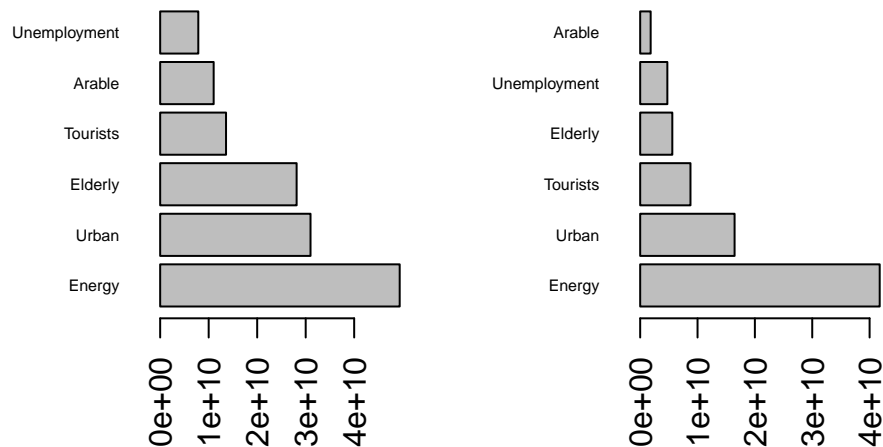
3. The following is a set of random forest models to predict GDP per capita from the above predictors considering `mtry` 1 through 7 and `maxnodes` 2, 5, 10, 20, and 40 with `ntree=200`. The plot shows the  $R^2$  of the out-of-bag predictions versus the true values for each parameter combination. Choose the best `mtry` and `maxnodes`.



4. The following plot shows how GDP per capita changes with gigajoules of energy produced per person holding the other variables constant. We estimate just this effect by, for each country, replacing the **Energy** term with the range of energies in consideration and then averaging over the countries for each value in our range. We are comparing our decision tree from question 1, our best random forest from question 3, and a decision tree fit only on **Energy**. Describe what you see in a few sentences.



5. Which predictors are most important in the first tree (left) and the best random forest (right)? How do they compare in relative importance?



## Categorical decision trees

In this class, we mainly look at decision trees as a non-parametric way of making a prediction about a continuous variable of interest. However, another (probably more common) use of decision trees is in predicting categorical variables. While looking at the sum of squared errors is a reasonable way to build decision trees for continuous predictions, other methods can be more useful for categorical predictions. This question will look at two such options.

1. A first method involves calculating the entropy of the parent and child nodes. Entropy is defined as:

$$E = - \sum_{i=1}^k p_i \log p_i$$

where  $p_i$  is the proportion of the items in the node from class  $i$ . (If  $p_i = 0$ , we treat its  $p_i \log_2 p_i$  term as 0 since  $p_i \log_2 p_i \rightarrow 0$  as  $p_i \rightarrow 0$ .) When splitting using this metric, the parent node's entropy  $E_{\text{Parent}}$  is calculated, and the weighted average of the children's entropies are calculated as

$$\frac{1}{n_{\text{Left}}} E_{\text{Left}} + \frac{1}{n_{\text{Right}}} E_{\text{Right}}$$

The split that yields the lowest entropy is chosen. Show that the entropy of a node is minimized when the node is “pure” (there are only items of one class in the node).

2. Show that the entropy of the node is maximized when there are equal proportions of items from each class in the node. Hint: Consider the function  $g(p) = p \log p$ . Also, you should use Jensen's inequality for sums:

$$\frac{1}{k} \sum_{i=1}^k g(p_i) \geq g\left(\frac{1}{k} \sum_{i=1}^k p_i\right)$$

when  $g$  is convex.

3. Explain why these results match our intuition of when a node should be split or not.

4. Because calculating logs can be computationally expensive, another method is to maximize the Gini value:

$$G = \sum_{i=1}^k p_i^2$$

where  $p_i$  is the proportion of the items in the node from class  $i$ . Explain why the Gini value can be interpreted as the probability that a randomly chosen item in the node would be assigned its correct class when assigning classes randomly according to the proportions in the node.

5. Find when the Gini value is maximized.

6. Show that the Gini value is minimized when all classes have equal proportions in the node.

7. Do we want to split a node with a high or low Gini value?

## Bagging and counting

In bagging, data points are bootstrapped and used to build overfit decision trees. By taking the average of these decision trees, we reduce the variance of predictions relative to a single tree built on all the data. The reason the trees are different is that they incorporate different data points; trees built from the same data points will be the same. Therefore, knowing how many data points overlap between bootstrap samples gives a rough measure of how different the trees will be.

1. Suppose we bootstrap  $n$  data points from our original  $n$  data points. Find the expected number of unique bootstrapped data points. Use the fact that  $(1 - \frac{1}{n})^n \approx e^{-1}$  for large  $n$  to approximate the expected number. (Hint: Think of the number of unique data points as a sum of indicators for each data point being unique. Also, recall that for  $r < 1$ ,  $\sum_{j=1}^n r^{j-1} = \frac{1-r^n}{1-r}$ )
2. What is the expected proportion of data points overlapping between two bootstrap samples of  $n$  items each (i.e., the proportion of data points in one bootstrap sample that are also in the other)?