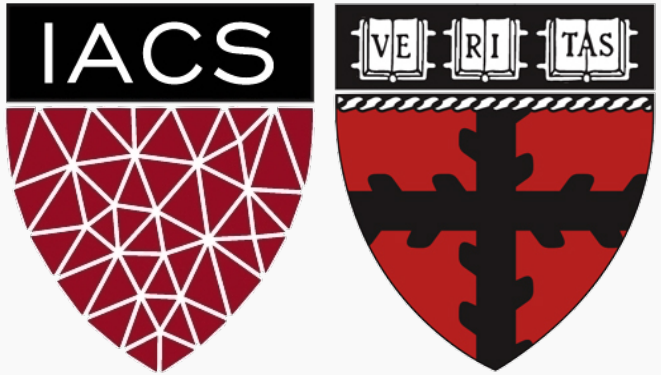
Decision Trees

CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai



Outline

- Motivation
- Decision Trees
- Classification Trees
- Splitting Criteria
- Stopping Conditions
- Regression Trees
- **Pruning**

Alternative to Using Stopping Conditions

What is the major issue with pre-specifying a stopping condition?

- you may stop too early or stop too late.

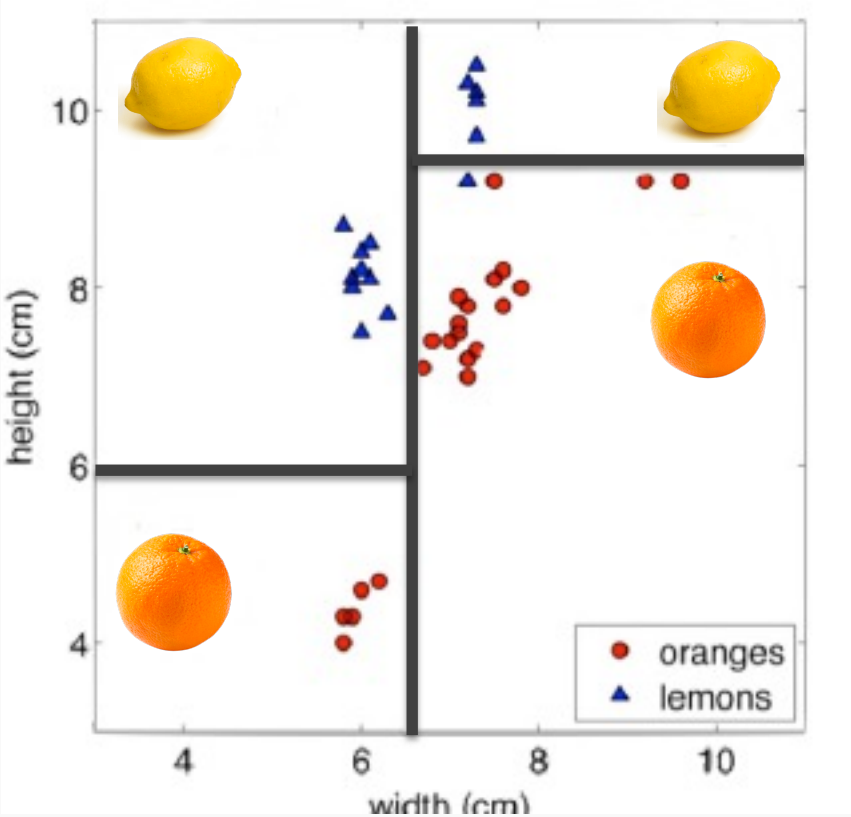
How can we fix this issue?

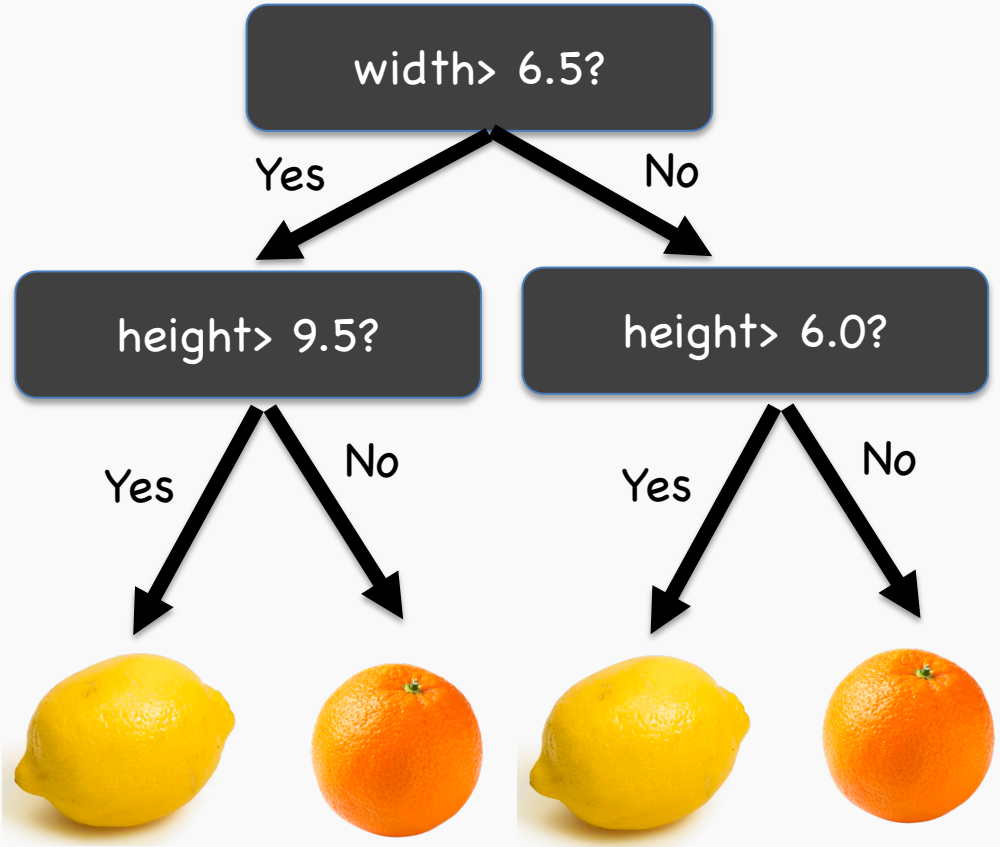
- choose several stopping criterion (set minimal Gain(R) at various levels) and cross-validate which is the best.

What is an alternative approach to this issue?

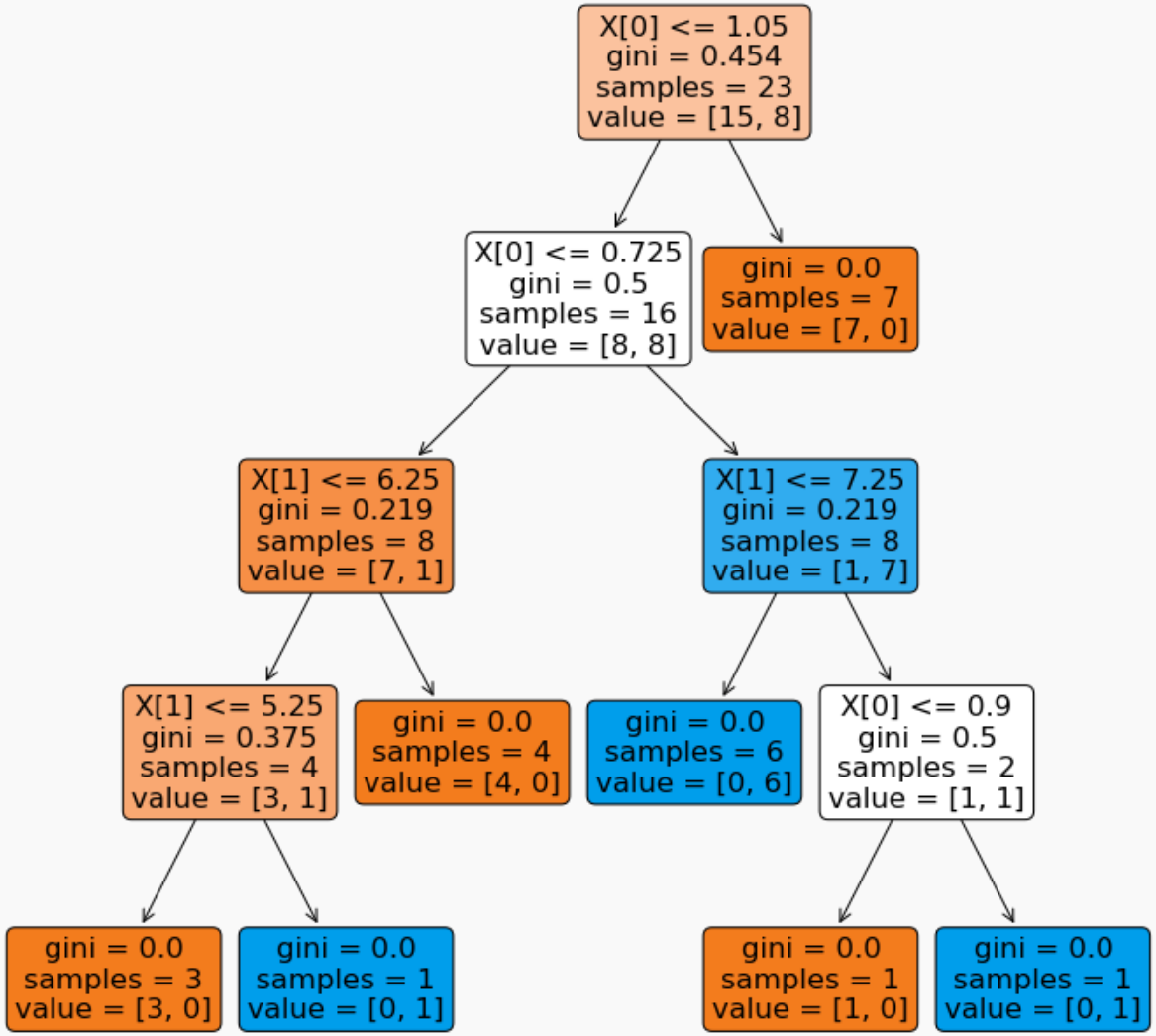
- Don't stop. Instead prune back!

Lemons or Oranges

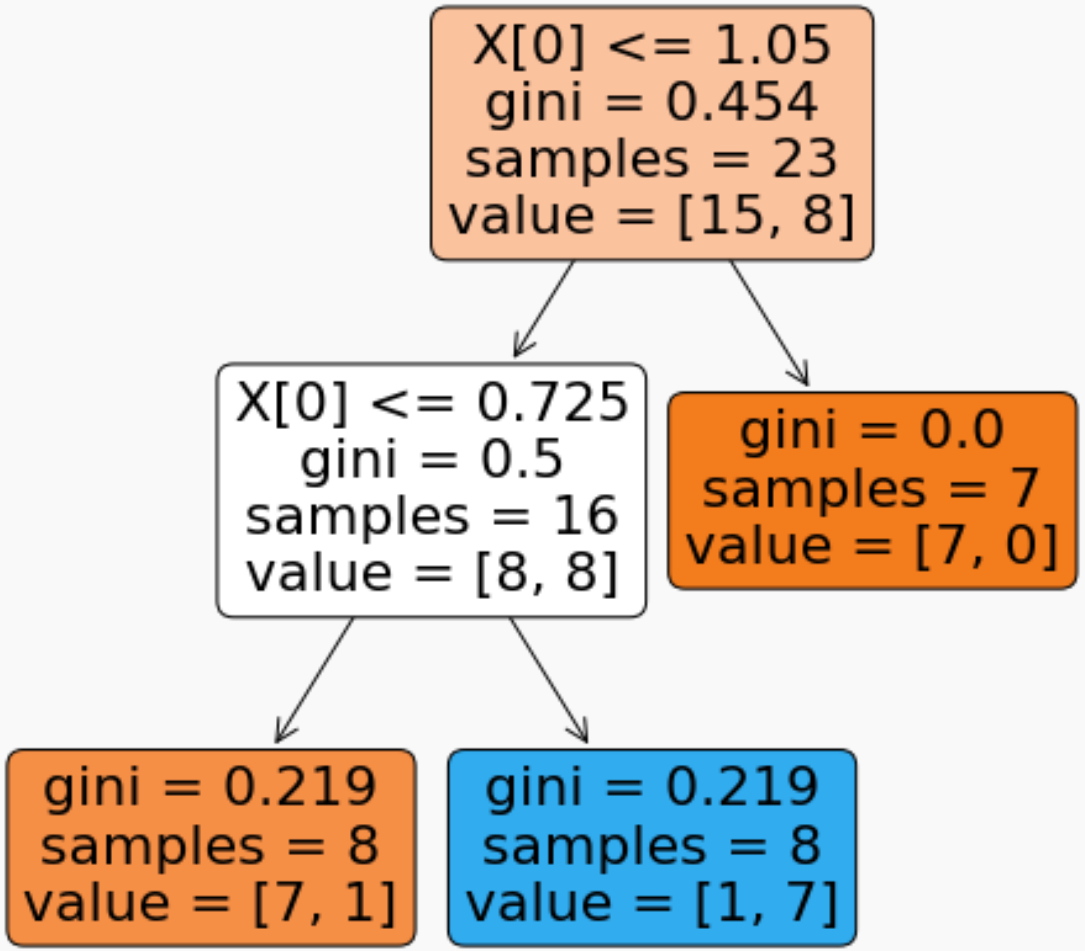




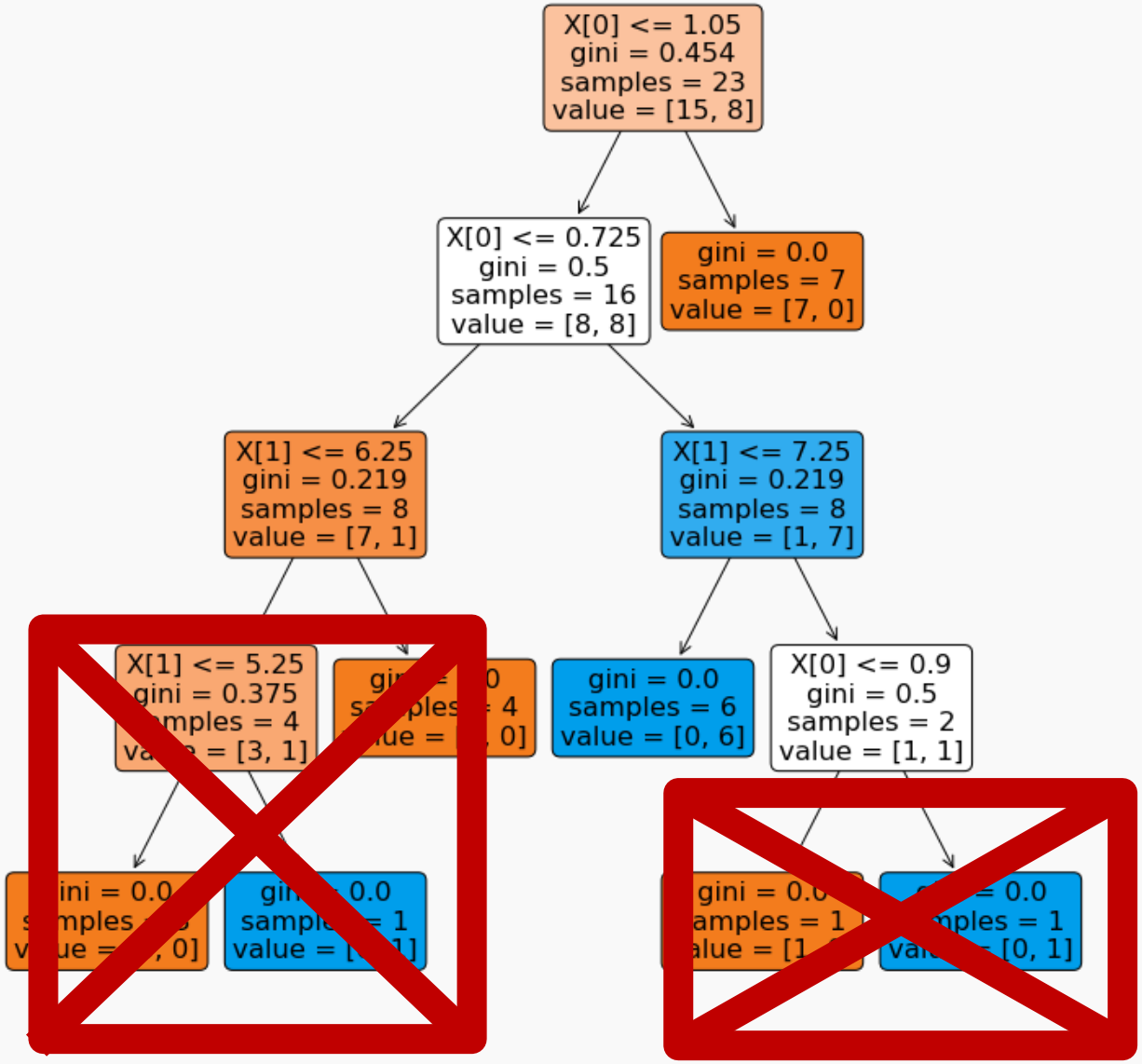
Motivation for Pruning



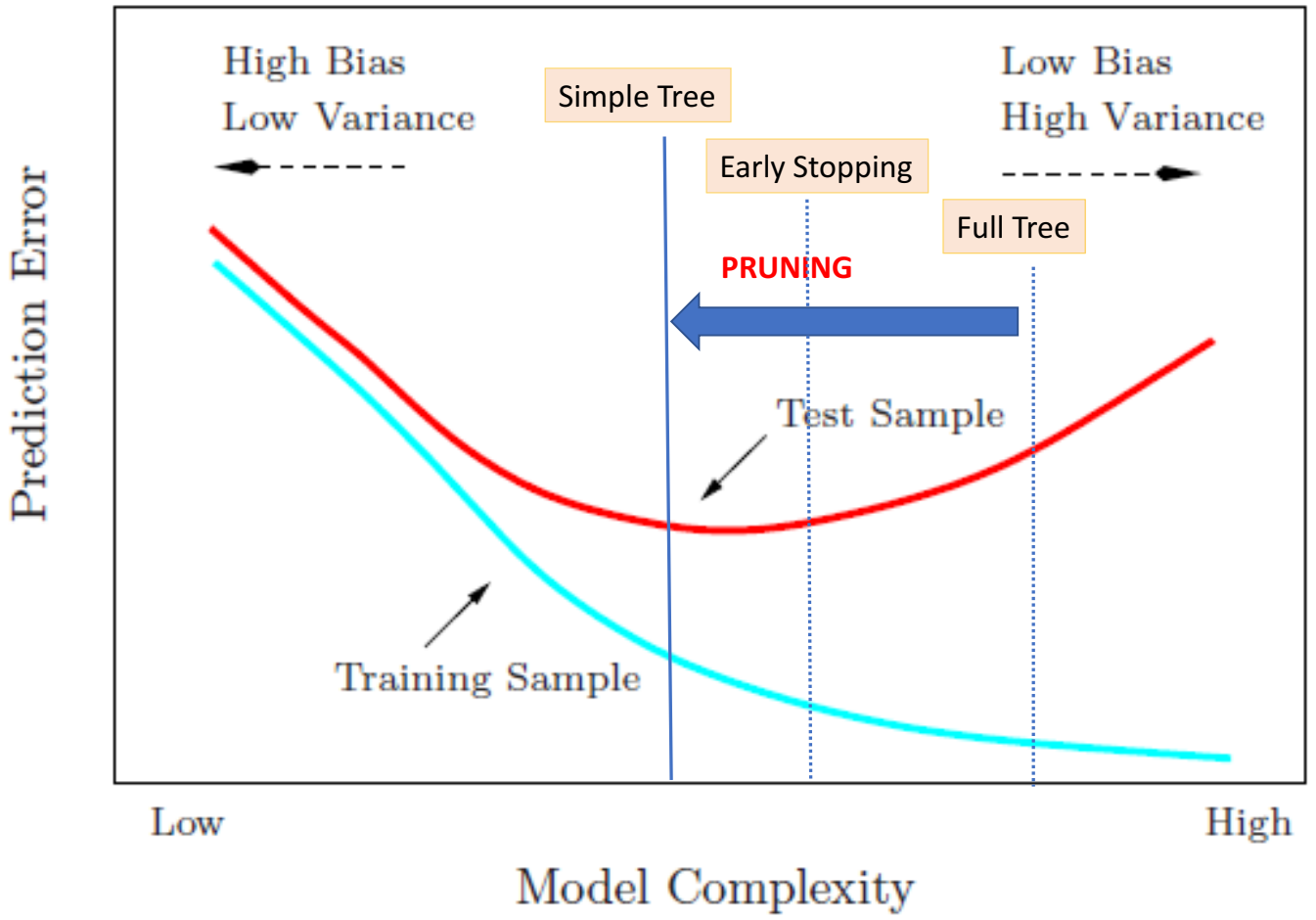
Motivation for Pruning



Motivation for Pruning



Motivation for Pruning



Pruning

Rather than preventing a complex tree from growing, we can obtain a simpler tree by 'pruning' a complex one.

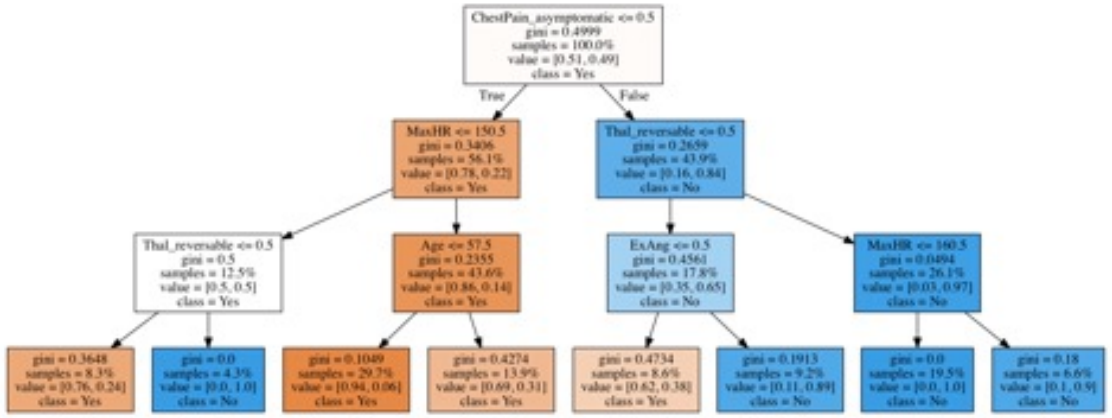
There are many method of pruning, a common one is ***cost complexity pruning***, where by we select from a array of smaller subtrees of the full model that optimizes a balance of performance and efficiency.

That is, we measure

$$C(T) = Error(T) + \alpha|T|$$

where T is a decision tree, $|T|$ is the number of leaves in the tree and α is the parameter for penalizing model complexity.

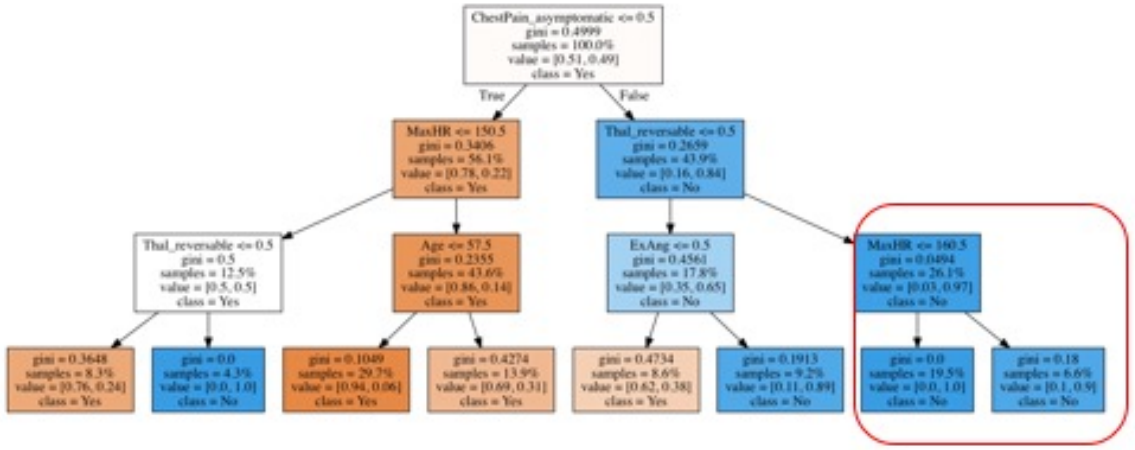
Pruning



$\alpha = 0.2$

Tree	Error	Num Leaves	Total

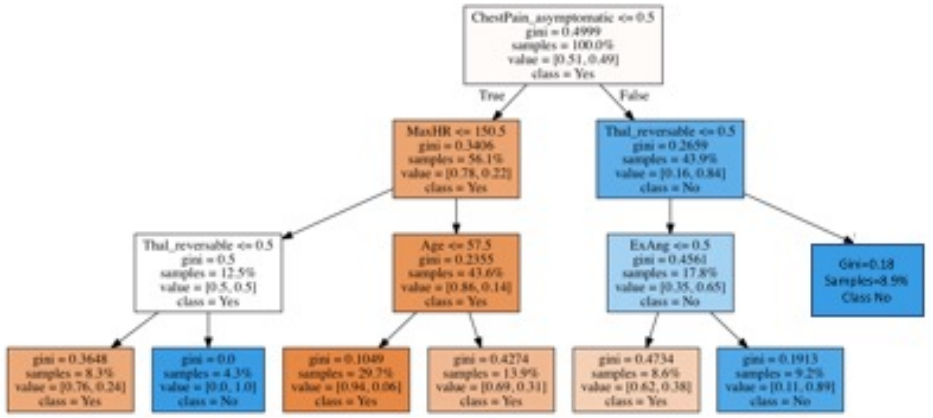
Pruning



$\alpha = 0.2$

Tree	Error	Num Leaves	Total
T	0.32	8	1.92

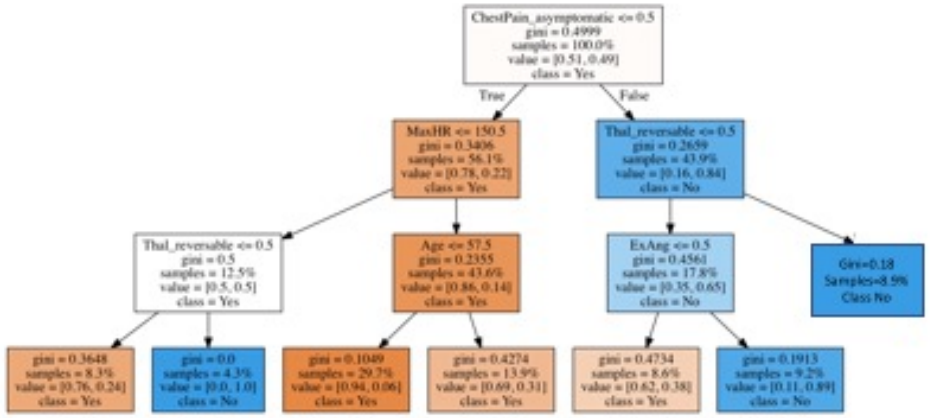
Pruning



$\alpha = 0.2$

Tree	Error	Num Leaves	Total
T	0.32	8	1.92
Tsmall	0.33	7	1.73

Pruning



$\alpha = 0.2$

Tree	Error	Num Leaves	Total
T	0.32	8	1.92
Tsmall	0.33	7	1.73

Smaller tree has larger error but less cost complexity score

Pruning

$$C(T) = Error(T) + \alpha|T|$$

1. Fix α.
2. Find best tree for a given α and based on cost complexity C.
3. Find best α using CV (what should be the error measure?)

Pruning

The pruning algorithm:

1. Start with a full tree T_0 (each leaf node is pure)
2. Replace a subtree in T_0 with a leaf node to obtain a pruned tree T_1. This subtree should be selected to minimize

$$\frac{Error(T_0) - Error(T_1)}{|T_0| - |T_1|}$$

3. Iterate this pruning process to obtain $T_0, T_1, \ldots, T_L$ where T_L is the tree containing just the root of T_0
4. Select the optimal tree T_i by cross validation.

Note: you might wonder where we are computing the cost-complexity $C(T_l)$. One can prove that this process is equivalent to explicitly optimizing C at each step.