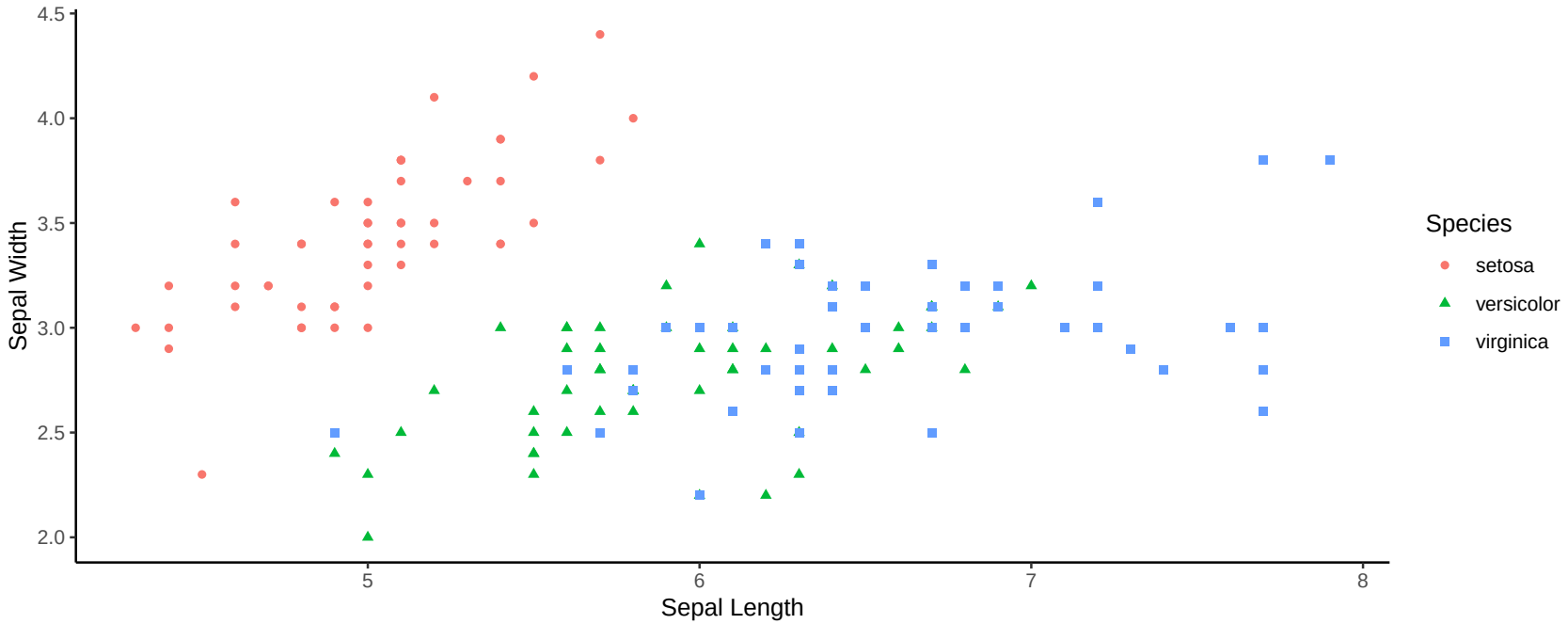
Decision Trees

Ryan Miller

Introduction

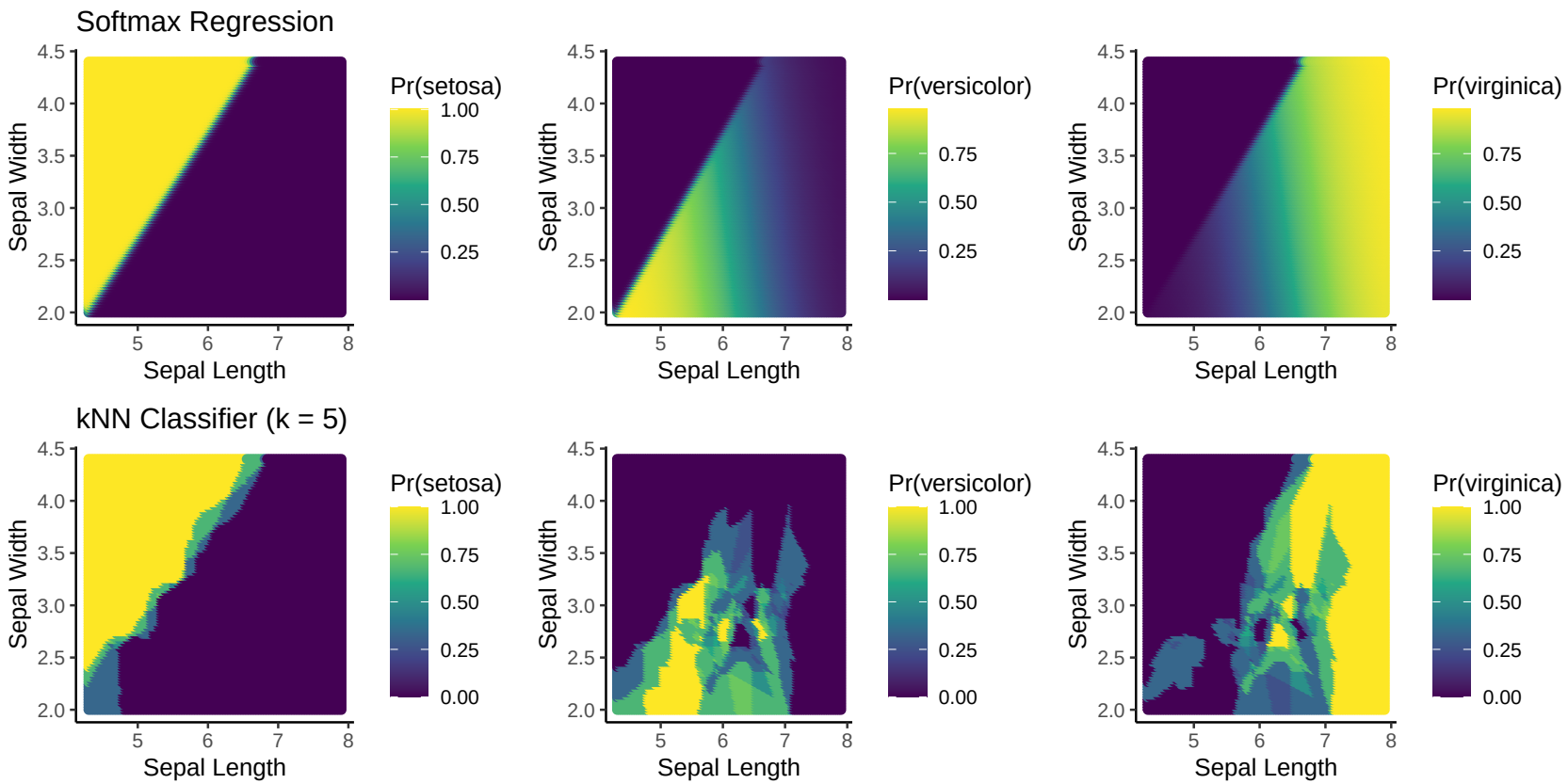
Shown below is Fisher's famous "iris" data set, which contains petal and sepal dimensions for examples of three different species of iris:



How can we use these data to classify new examples of iris?

Previous approaches

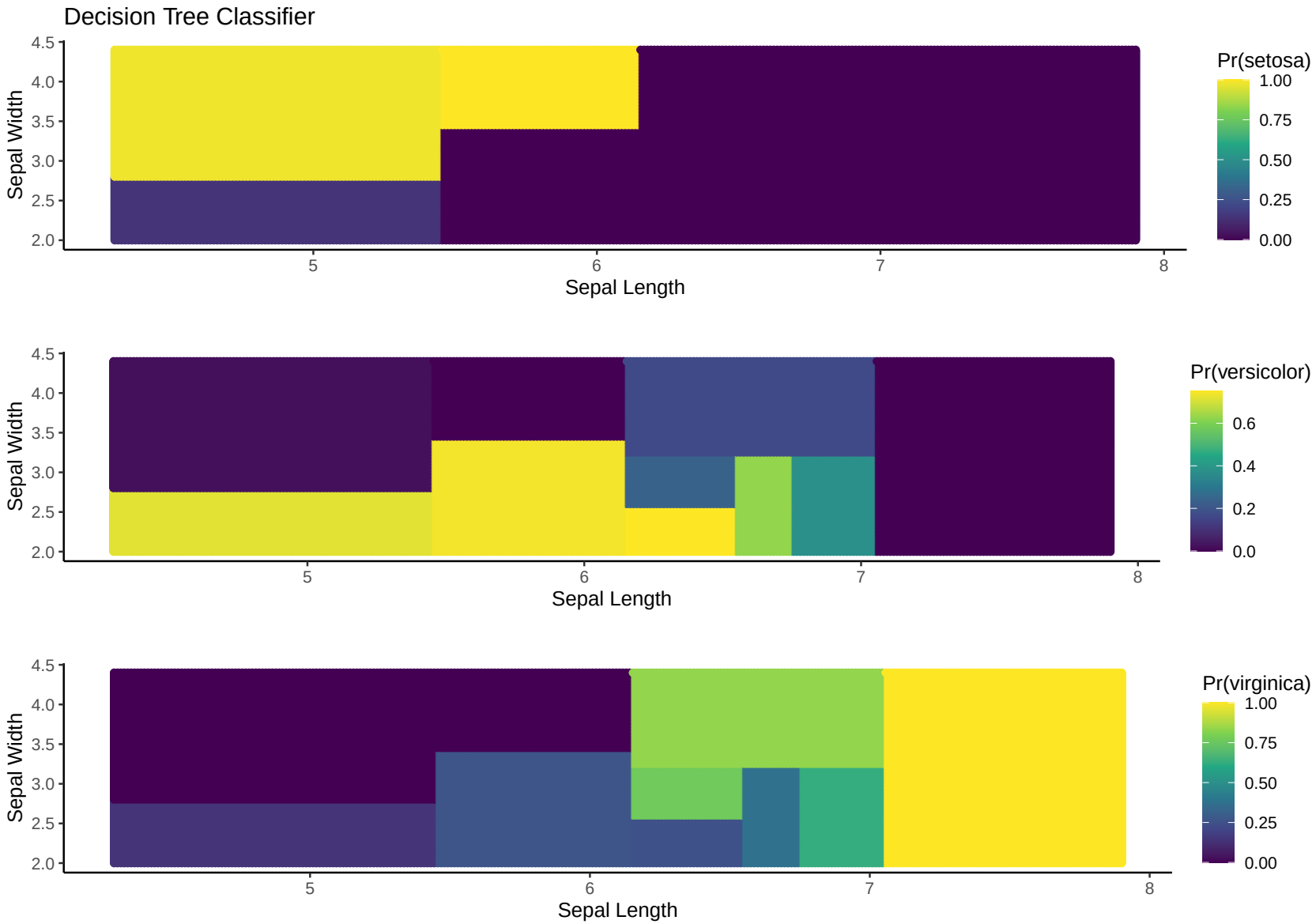
KNN and *softmax regression* are methods of multi-label classification. Shown below are the results:



Previous approaches

- The structure imposed by softmax regression might be overly biased for this application
- While the k-nearest neighbors model likely is too high in variance
- Today we will introduce tree-based models, which still impose some structure, but have the potential to be less biased than highly structured models like softmax regression

Decision trees on the iris data

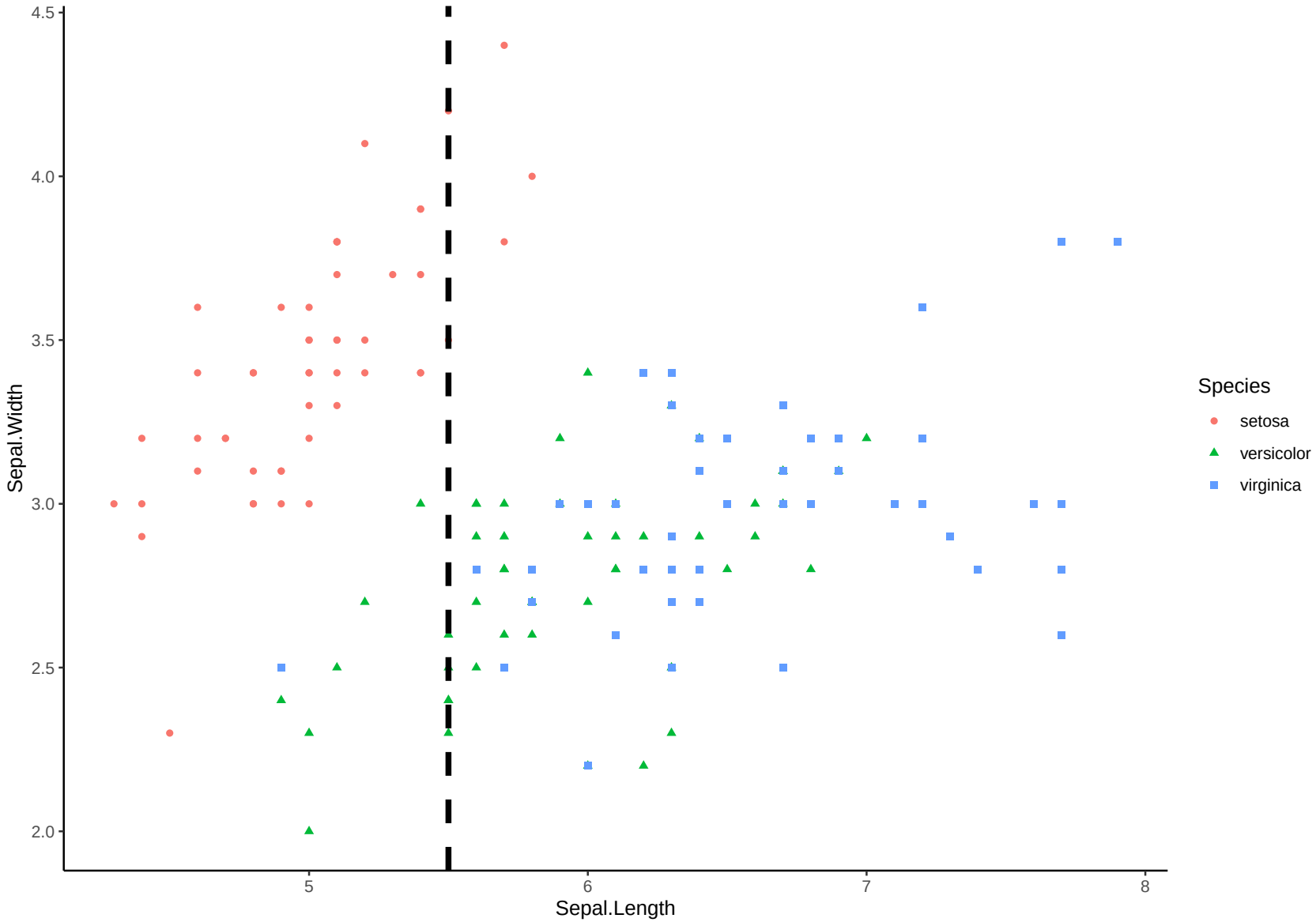


Decision trees

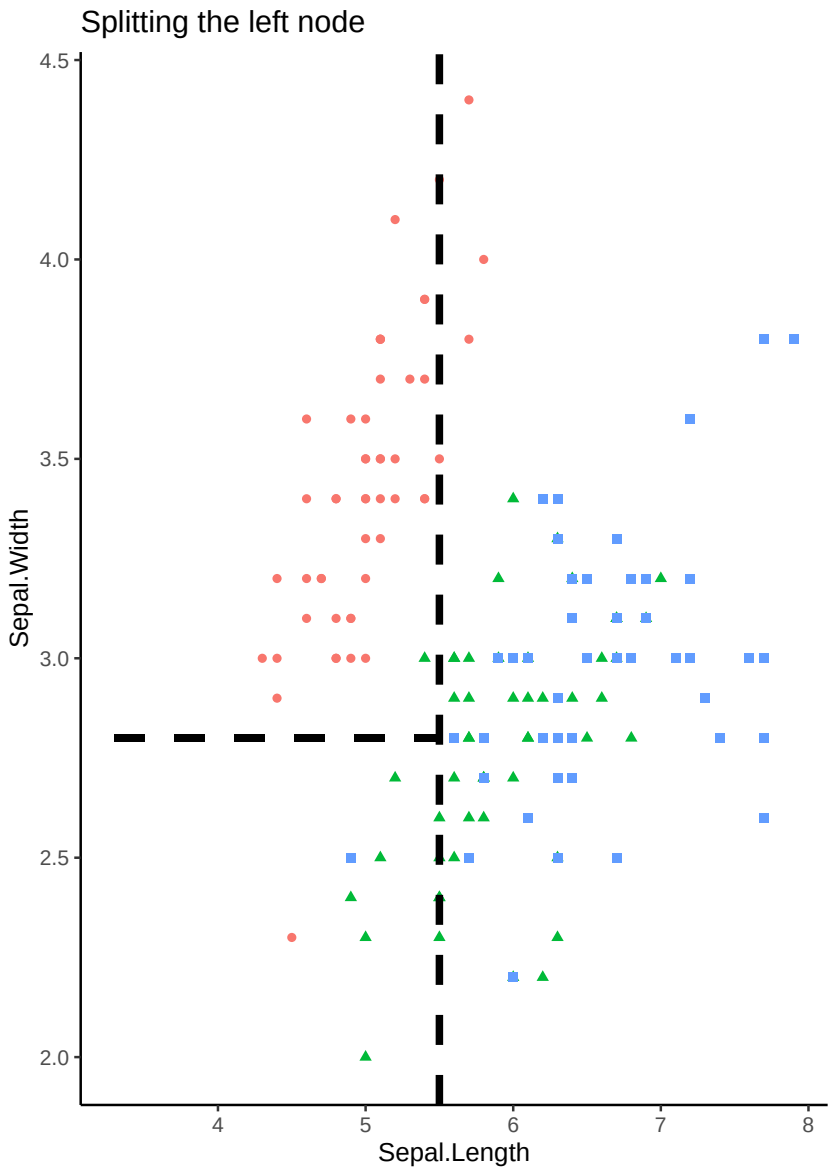
Decision trees are trained by recursively partitioning the p-dimensional space (defined by the explanatory variables) until an acceptable level of homogeneity or "purity" is achieved within each partition:

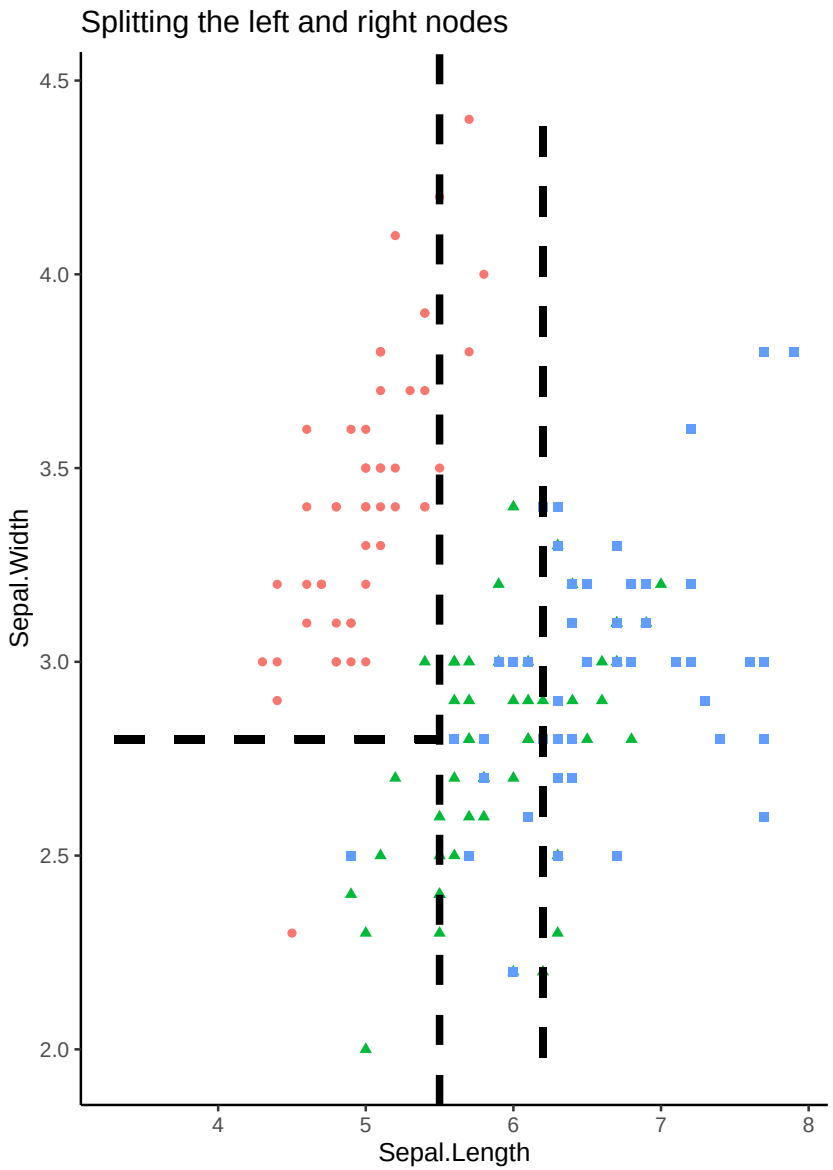
1) Starting with a "parent" node, search for a splitting rule that maximizes the *homogeneity* or *purity* of the "child" nodes
2) Next, considering each node that hasn't yet been split, find another splitting rule that maximizes *purity*
3) Repeat until a stopping criteria has been reached

Example (first split)

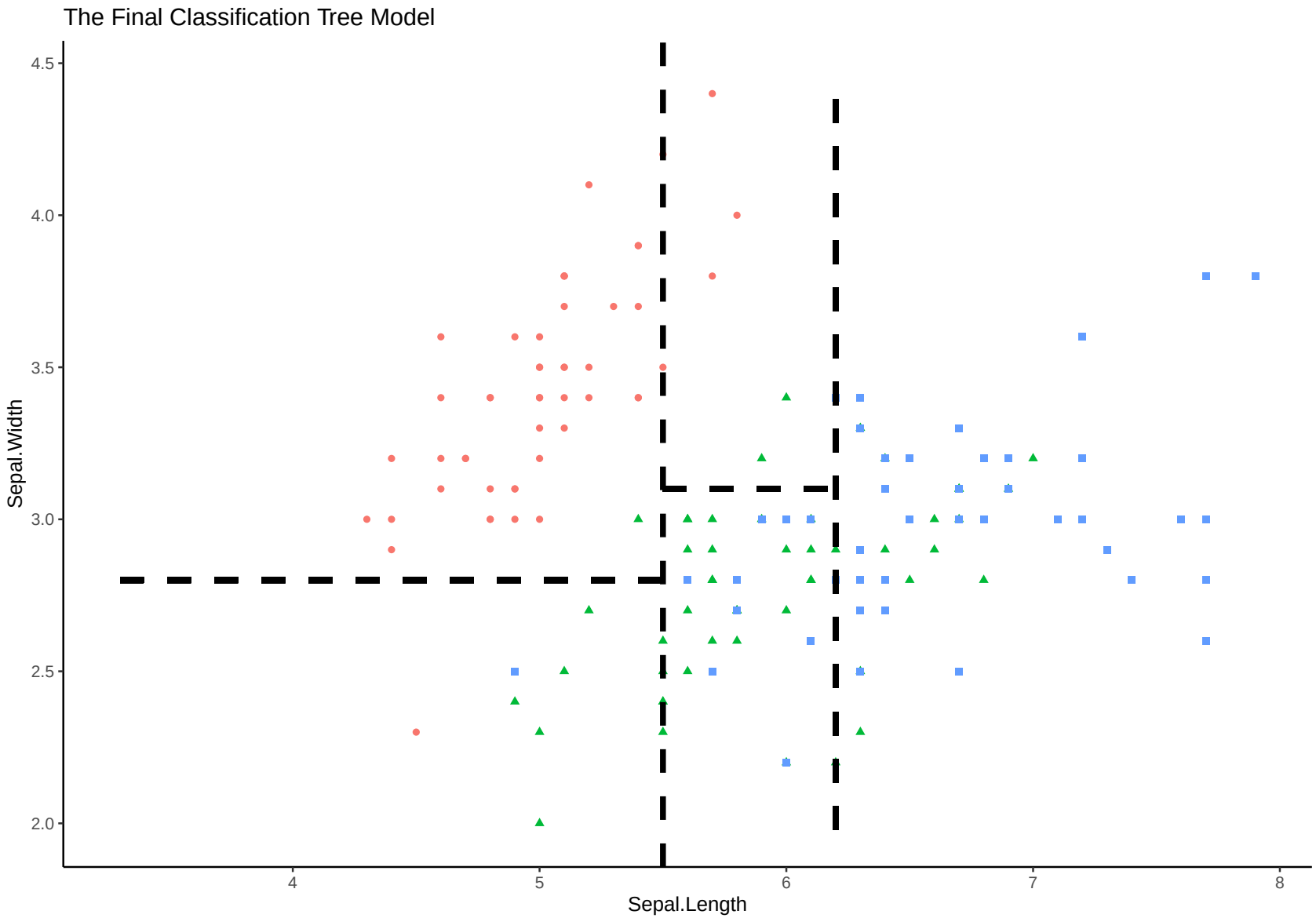


Example (second split)

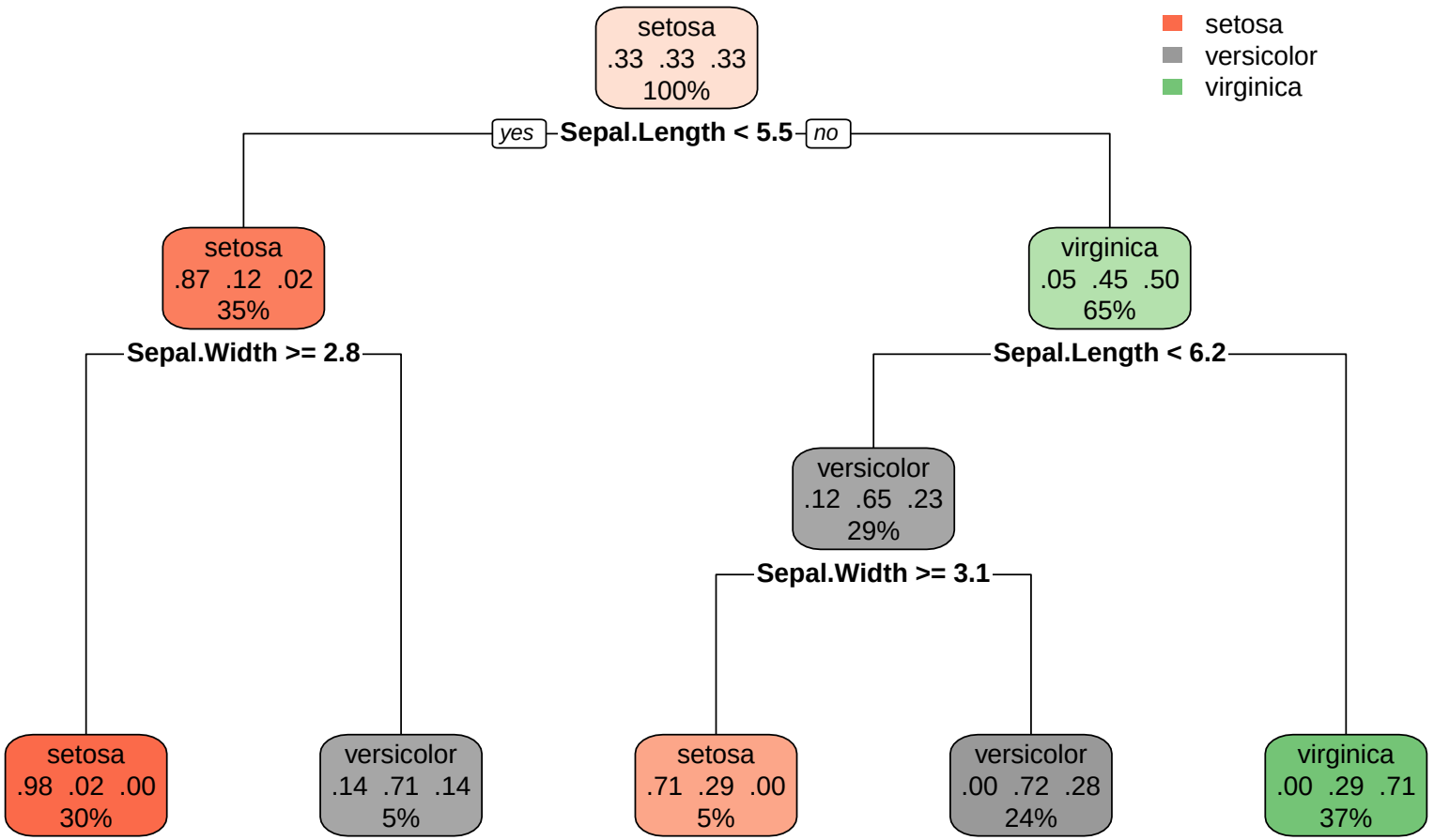




Example (final model)



Example (full tree)



Splitting criteria

Decision trees must learn their splits using an objective criteria, the most common criteria is *Gini impurity*:

$$Gini = \sum_{j=1}^{k} p_j(1 - p_j) = 1 - \sum_{j=1}^{k} p_j^2$$

- For binary classification, this reduces to $p_1(1 - p_1) + p_2(1 - p_2)$

Splitting criteria

Gini Gain

- Gini impurity gives us a measure of the impurity of a tree at given depth
  - In our example tree, the Gini impurity of the starting node was $1 - (\frac{1}{3}^2 + \frac{1}{3}^2 + \frac{1}{3}^2) = \frac{2}{3}$

Gini Gain

- Gini impurity gives us a measure of the impurity of a tree at given depth
  - In our example tree, the Gini impurity of the starting node was $1 - (\frac{1}{3}^2 + \frac{1}{3}^2 + \frac{1}{3}^2) = \frac{2}{3}$
- The best split is the one that produces the largest decrease (or *Gini gain*) in the resulting child nodes
  - In our example, the Gini impurity after the first split was:

$$0.35 * [1 - (0.87^2 + 0.12^2 + 0.02^2)] +$$
$$0.65 * [1 - (0.05^2 + 0.45^2 + 0.50^2)] = 0.434$$

- Thus, the Gini gain was 0.233

Decision Tree Learning

Building a tree involves iterating between two steps:

1. Finding the optimal split point within each variable
2. Selecting the variable to split on

Even if an exhaustive approach were taken, this can be done fairly quickly as there are $n - 1$ possible split points per variable and p variables

- Modern tree solving algorithms are beyond the scope of this course, but as you might image they do not check every possible split

Numeric Outcomes

Applying the decision tree algorithm to data with a numeric outcome requires the following changes:

- Predicted outcomes are the average value in node
- Mean squared error is used instead of Gini impurity as a measure of impurity

Other splitting criteria, such as mean absolute error or Poisson deviance are available in `sklearn`

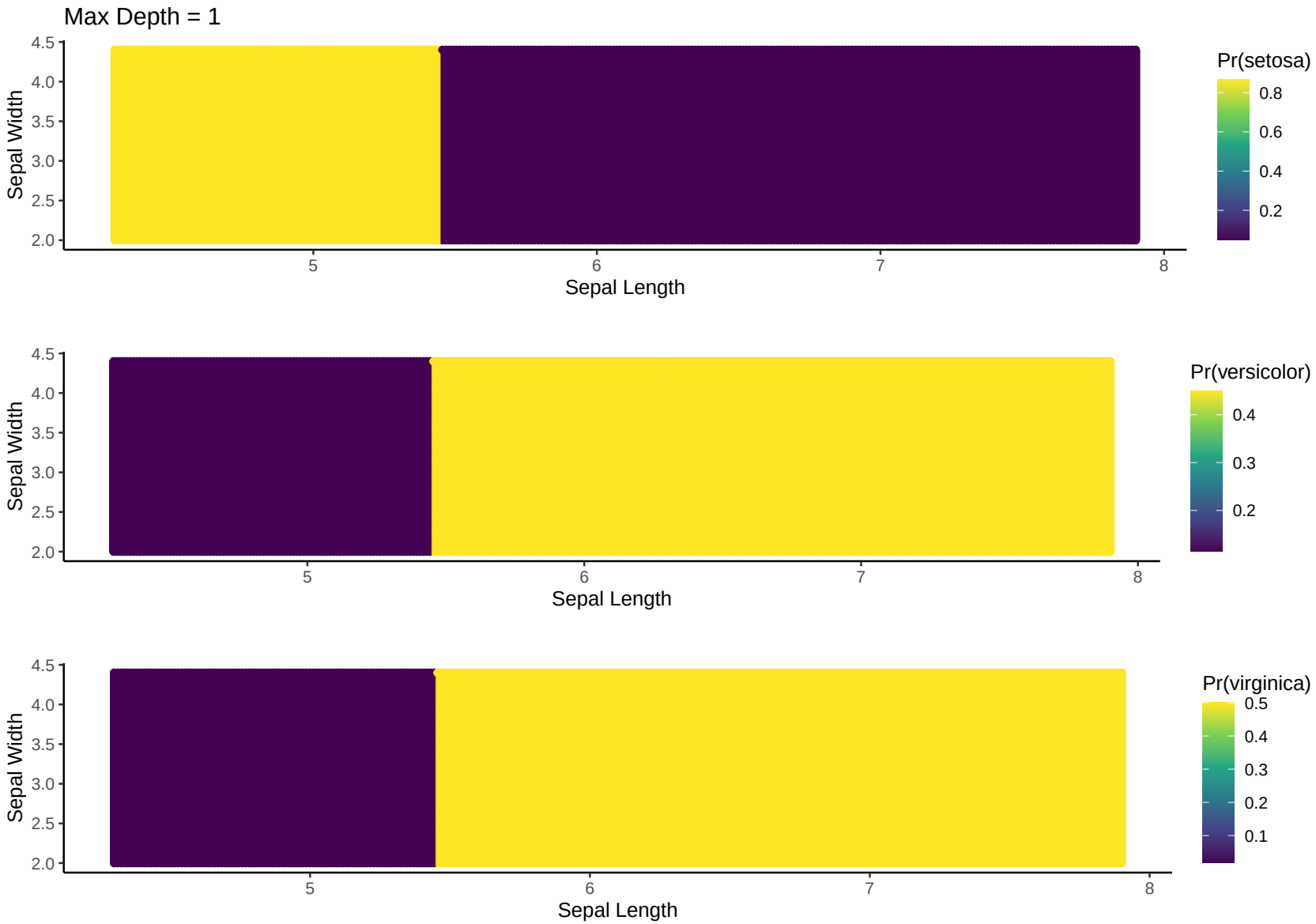
Stopping the algorithm

Decision trees can be grown until every terminal node is perfectly pure; however, such trees will be very overfit to the training data. We can manipulate the bias-variance trade-off in a fitted tree in the following ways:

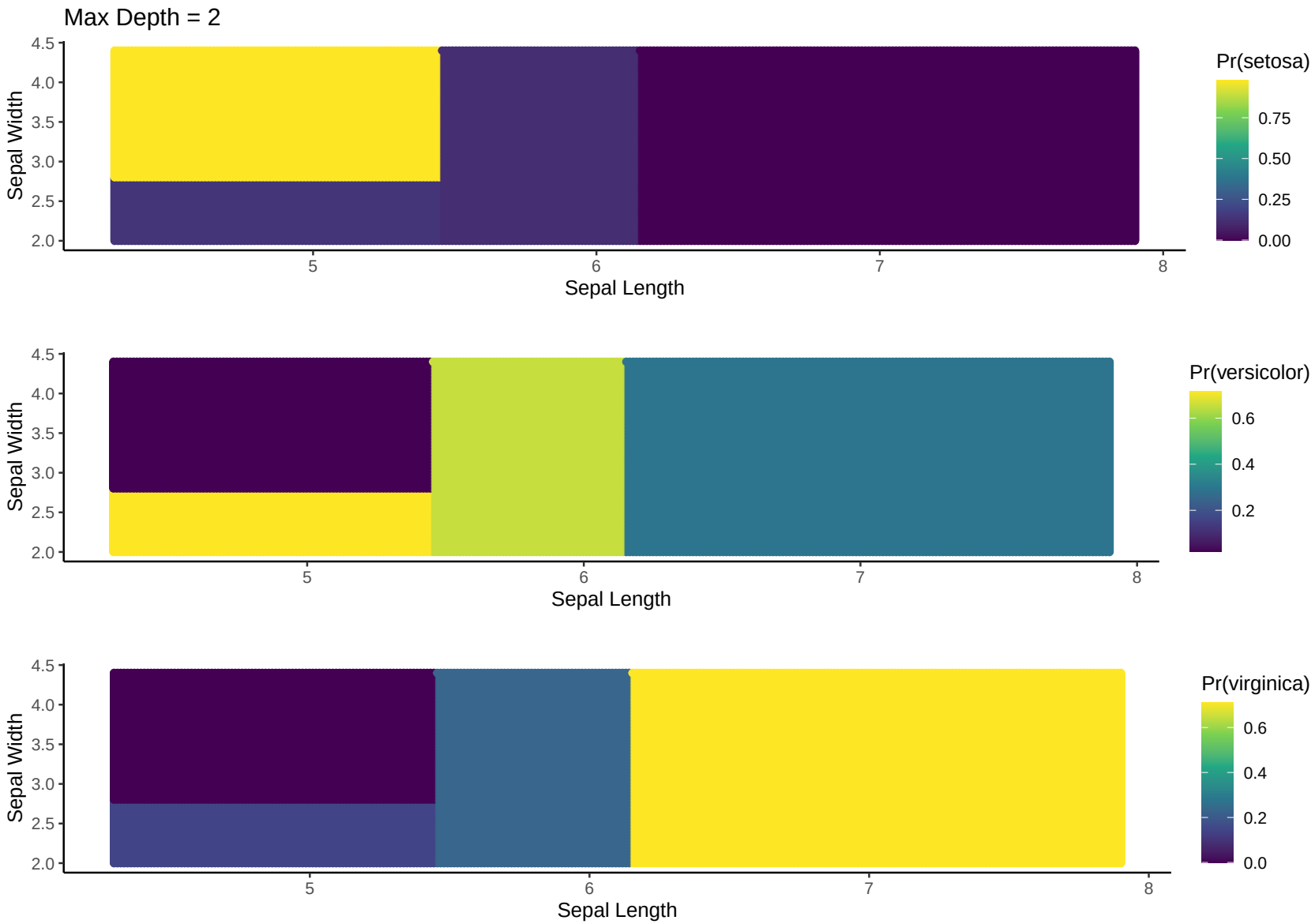
1. Restricting the maximum depth of the tree (ie: the number of sequential rules)
2. Allowing only nodes of sufficient size be eligible for splitting
3. Requiring a certain improvement in purity for a split to occur

Generally, maximum depth is the most important factor to consider as it directly relates to the complexity of a tree

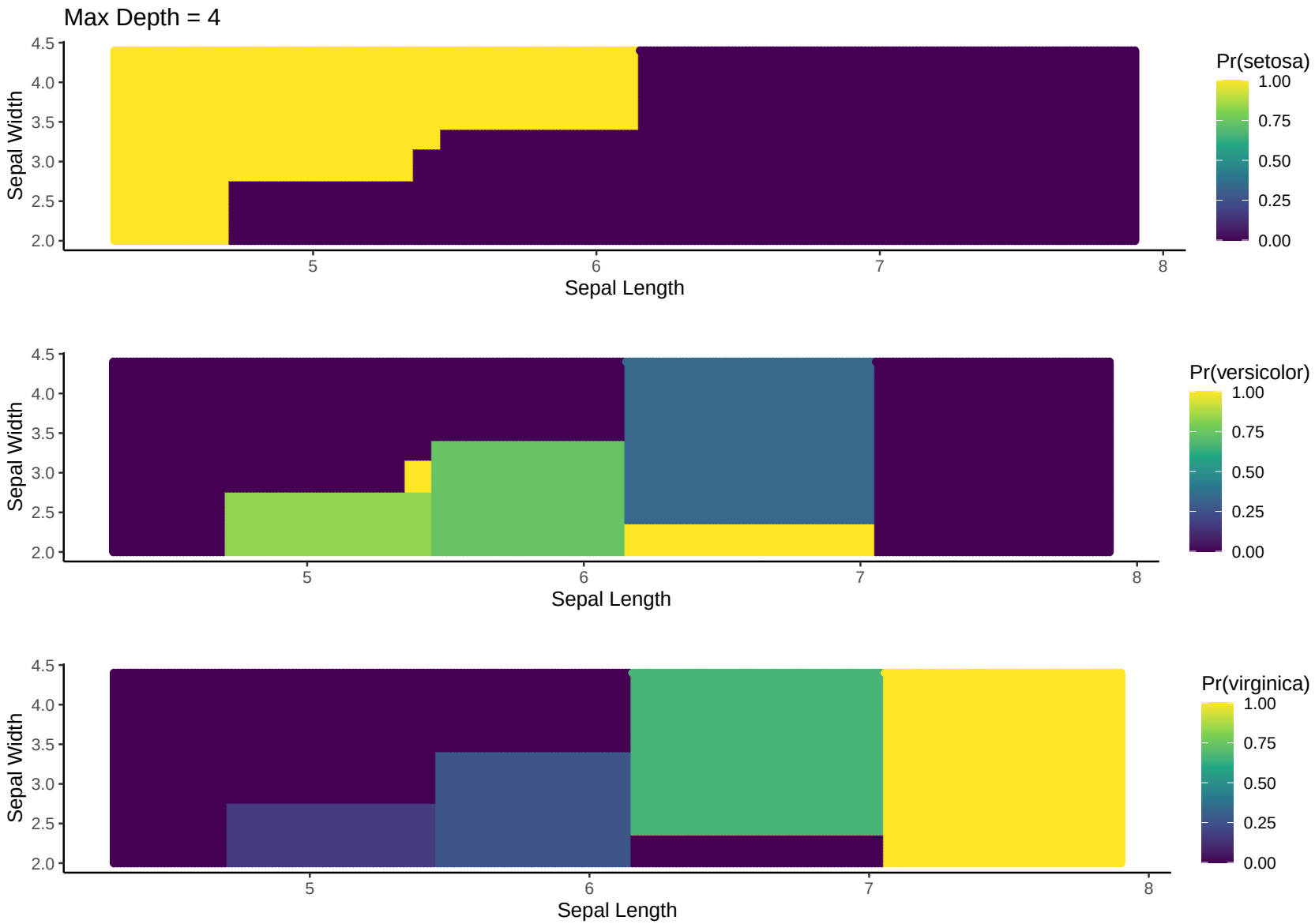
Examples (bias-variance tradeoff)



Examples (bias-variance tradeoff)



Examples (bias-variance tradeoff)



Examples (bias-variance tradeoff)

Resting metabolism dataset: