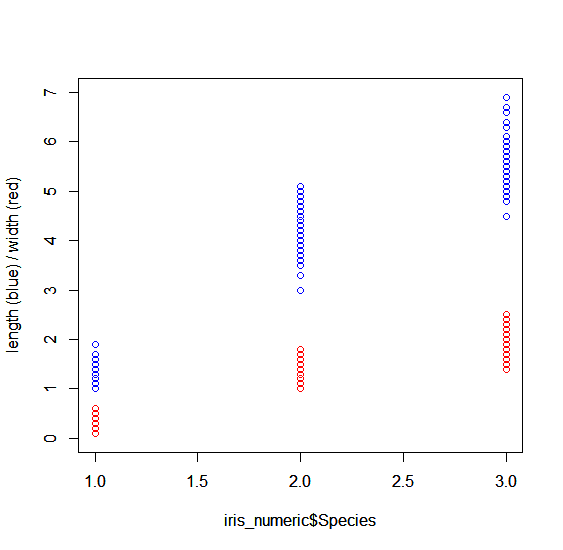
iris

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
| Sepal.Length | 1 | -0.11757 | 0.871754 | 0.817941 | 0.782561 |
| Sepal.Width | -0.11757 | 1 | -0.42844 | -0.36613 | -0.42666 |
| Petal.Length | 0.871754 | -0.42844 | 1 | 0.962865 | 0.949035 |
| Petal.Width | 0.817941 | -0.36613 | 0.962865 | 1 | 0.956547 |
| Species | 0.782561 | -0.42666 | 0.949035 | 0.956547 | 1 |

Most strongly correlated are a grouping between Petal Length, Width and Species. These mean that they will most likely be the determining factor of species.

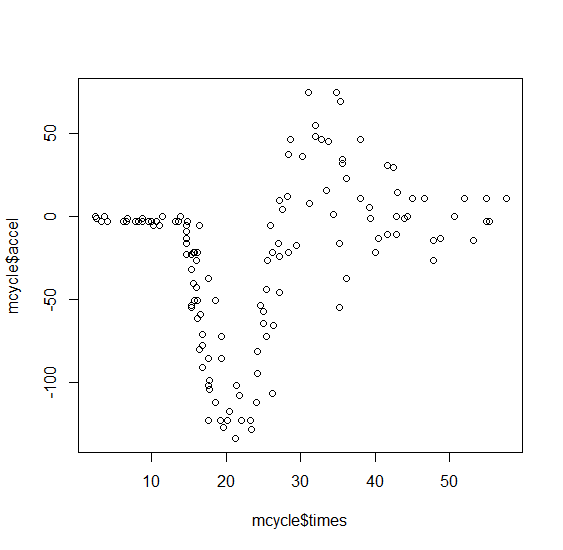
These variables are so correlated they will likely both be predictor variables.



Mcycle

|  |  |  |
| --- | --- | --- |
|  | times | accel |
| times | 1 | 0.296403 |
| accel | 0.296403 | 1 |

This set only has two variables meaning that there is only one relationship and one primary response variable.

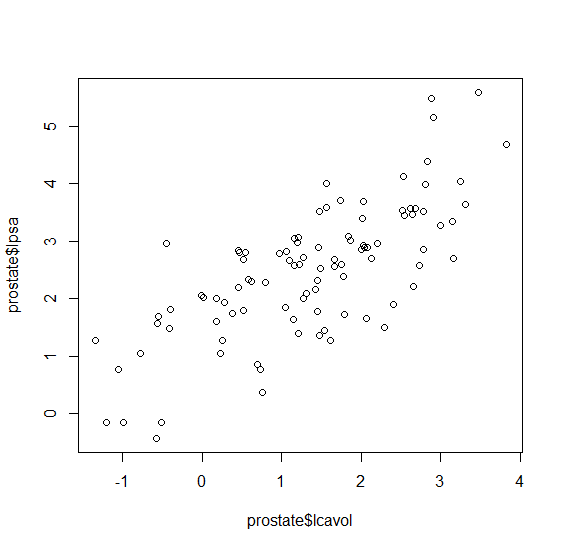


Prostate

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | lcavol | age | lbph | lcp | gleason | lpsa |
| lcavol | 1 | 0.225 | 0.02735 | 0.67531 | 0.432417 | 0.73446 |
| age | 0.225 | 1 | 0.350186 | 0.127668 | 0.268892 | 0.169593 |
| lbph | 0.02735 | 0.350186 | 1 | -0.007 | 0.07782 | 0.179809 |
| lcp | 0.67531 | 0.127668 | -0.007 | 1 | 0.51483 | 0.548813 |
| gleason | 0.432417 | 0.268892 | 0.07782 | 0.51483 | 1 | 0.368987 |
| lpsa | 0.73446 | 0.169593 | 0.179809 | 0.548813 | 0.368987 | 1 |

The two highest correlations in this data set is lcavol ~ lpsa and lcavol ~ lcp. LPSA (log of prostate specific antigen) would be highly correlated to the cancer volume as the body would be trying to fight against the cancer more (create more antigens) as the volume of cancer increases. Honestly I have no idea what capsular penetration is so I can’t weigh in on this.

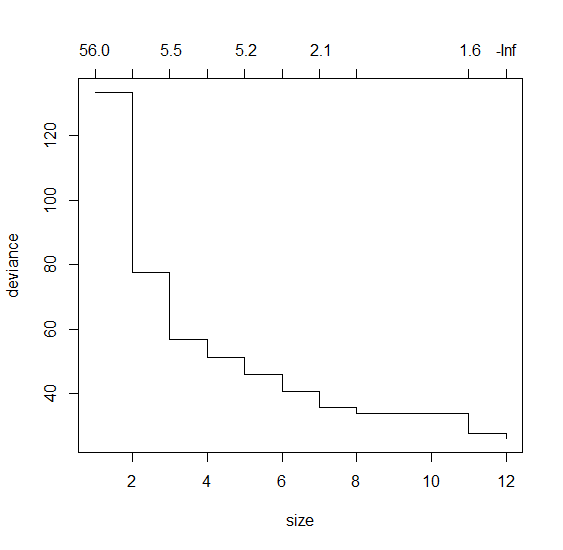
The response variable is lcavol and likely the most significant predictor would be LPSA due to the significantly higher correlation.



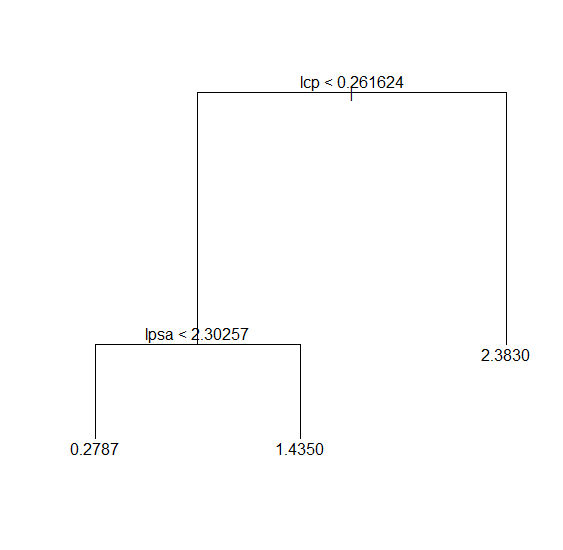


Prostate

After loading the data set and creating the tree we need to determine the optimal size of tree to create. By using the prune.tree function in R we can create a plot which will give us deviance and the size of the tree itself. Optimally we’d want a tree that is not too large without having the deviance too high.

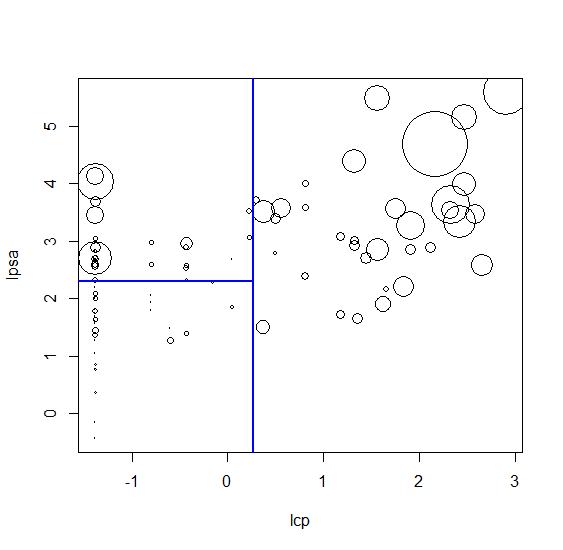


This plot shows the optimal size of the tree would be 3. This gives the smallest tree with the proportion of deviance.



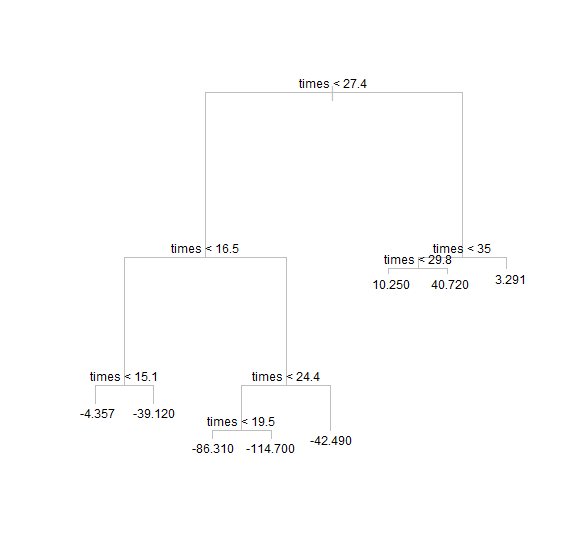
With that the following tree is created splitting the model when LCP is less than 0.261624 and again when LPSA is less than 2.30257.

This splits the model into three sets which are visualized as:

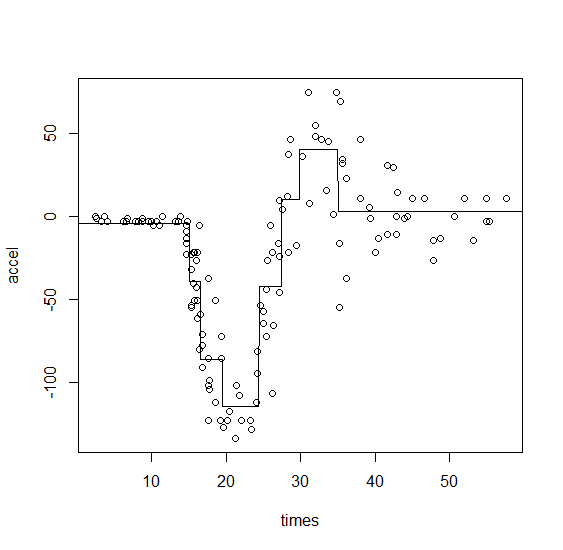


Mcycle

This data set is interesting in that it only has a response variable and one predictor variable. This means to create an accurate model we need to create a tree which splits multiple times off that single variable:

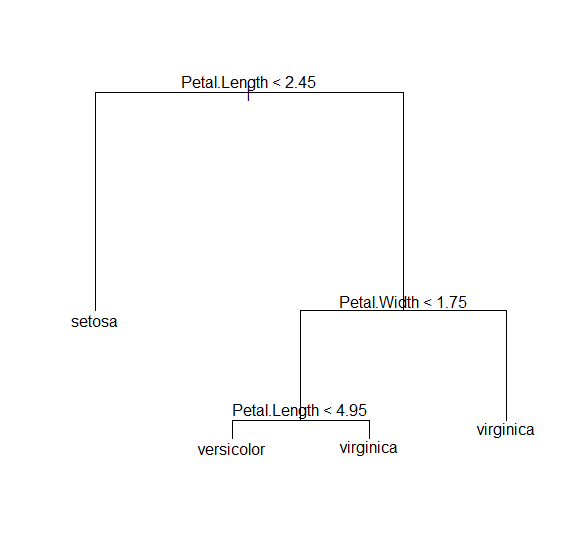


This model creates a model which divides the data in “half” by deviating a line through the set to split it into two regions:



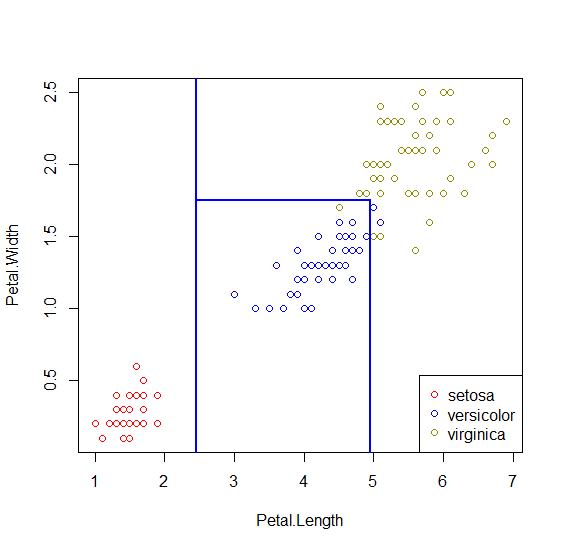
Iris

This data set needed the least attention as it is well fitted for a decision tree model. After trimming the tree it produces the following model.

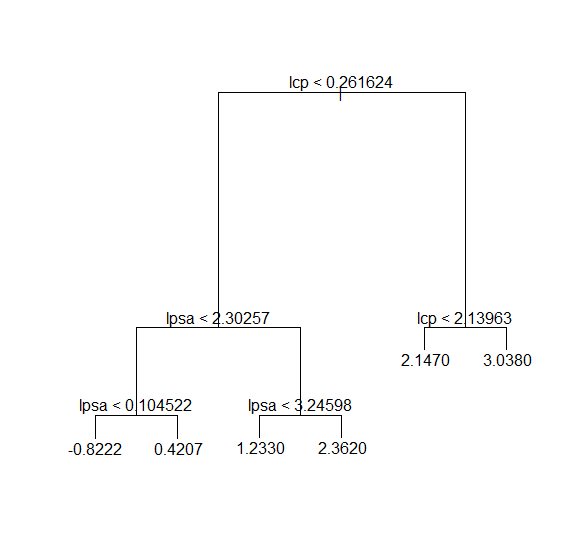


This model deems that setosa has a petal length < 2.45, if the iris has petal length of > 2.45 and a petal width of > 1.75 it must be a virginica. If the petal length is < 1.75 and has a petal length > 4.95 then it must again be a virginica. Otherwise it must be a versicolor.

1. I found that I wanted a more in-depth visual on the iris tree more like what was created on the other two data sets:



1. After setting the alpha to 5 the following tree is produced



This tree has 6 leaves which is expected based off the below graph of deviance and tree size:

