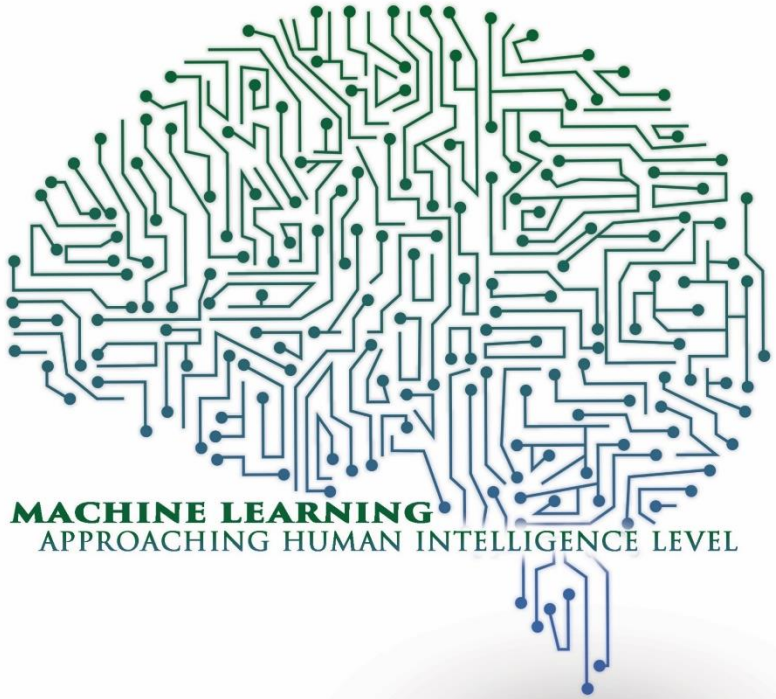
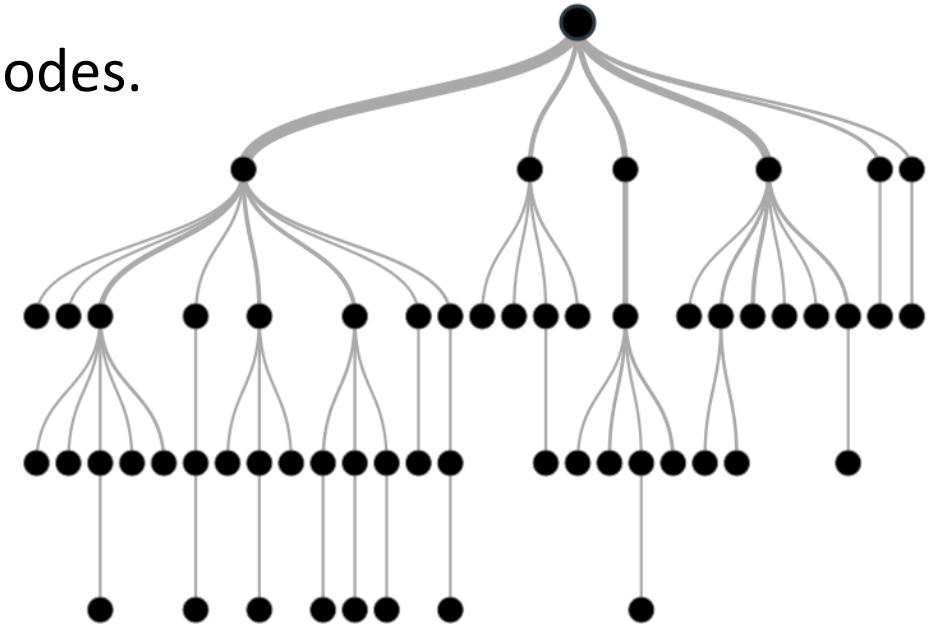


Decision Trees & Random Forest

Classification and Regression Trees

- Grow a binary tree.
- At each node, “split” the data into two “daughter” nodes.
  - Splits are chosen using a splitting criterion.



Bottom nodes are “terminal” nodes.

- For regression the predicted value at a node is the *average* response variable for all observations in the node.
- For classification the predicted class is the *most common class* in the node (majority vote).
- For classification trees, can also get estimated probability of membership in each of the classes

Splitting criterion

Regression: residual sum of squares

$$RSS = \sum_{left} (y_i - y_L^*)^2 + \sum_{right} (y_i - y_R^*)^2$$

where y_L^* = mean y-value for left node

y_R^* = mean y-value for right node

Classification: Gini criterion

$$Gini = N_L \sum_{k=1,\ldots,K} p_{kL} (1- p_{kL}) + N_R \sum_{k=1,\ldots,K} p_{kR} (1- p_{kR})$$

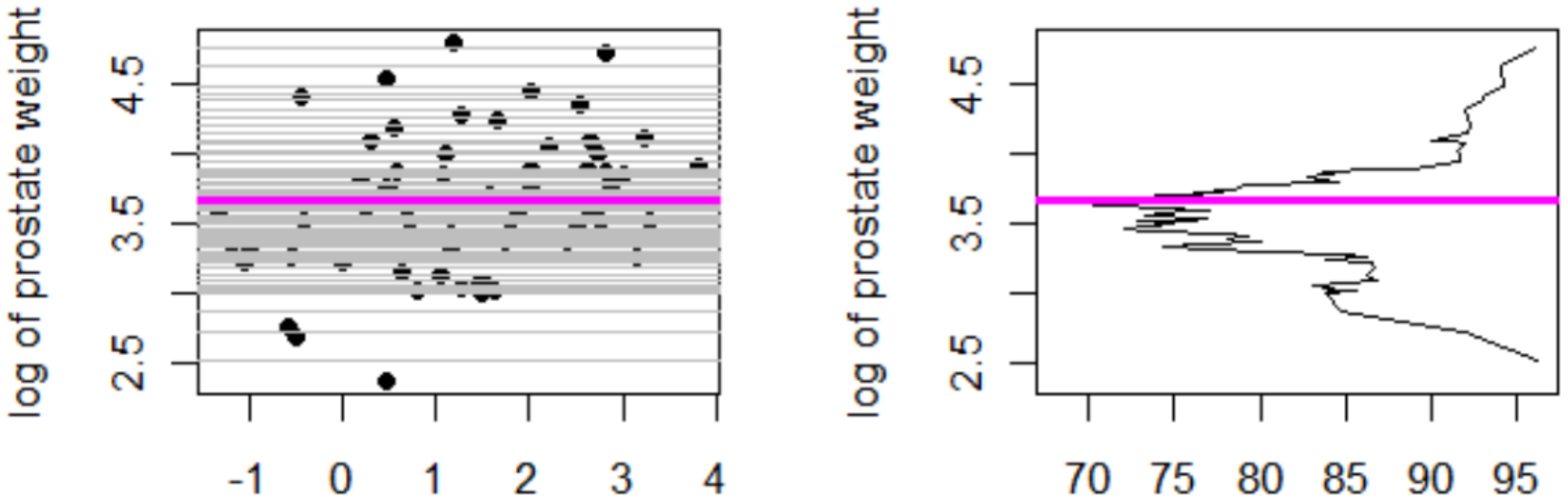
where p_{kL} = proportion of class k in left node

p_{kR} = proportion of class k in right node

Standard deviation

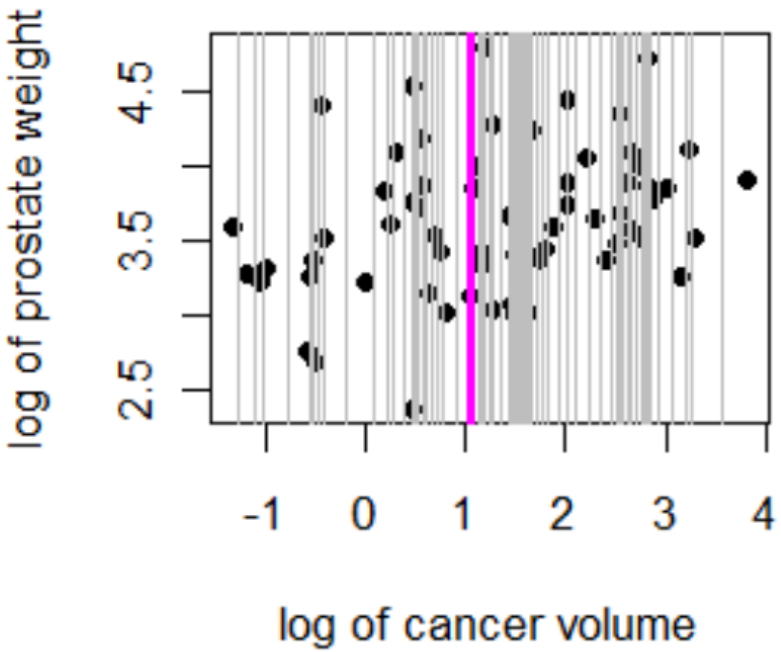
$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

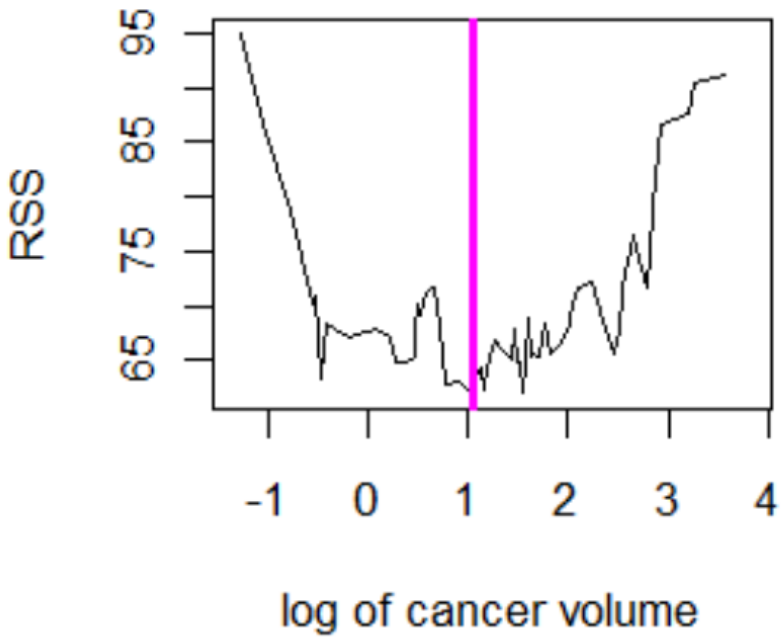
Cancer Example



Best horizontal split is at 3.67 with RSS = 68.09.

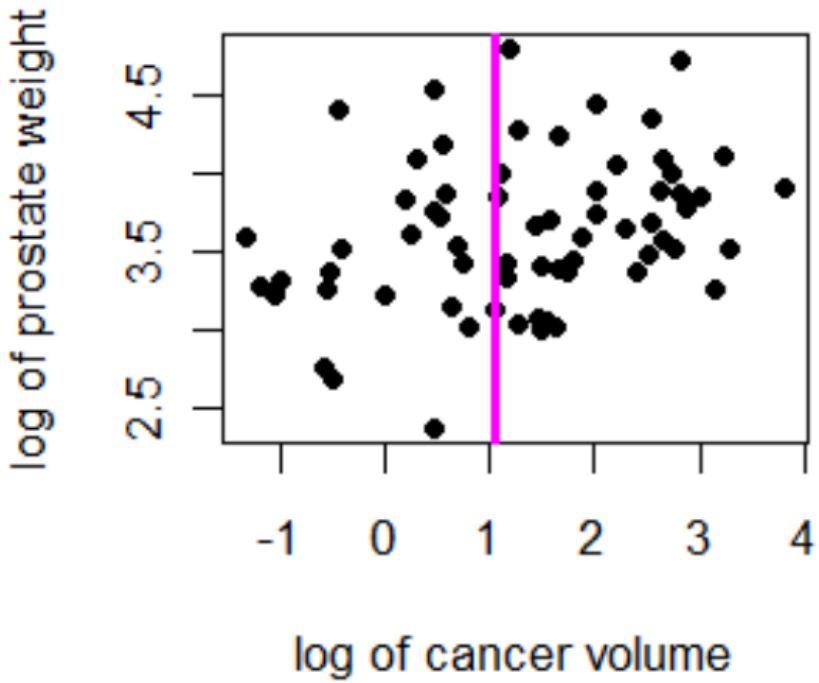
Adele Cutler
Utah State University

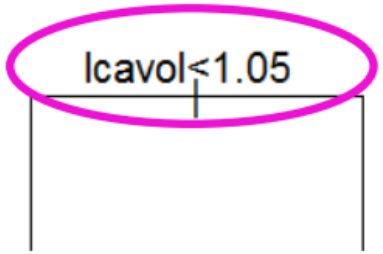




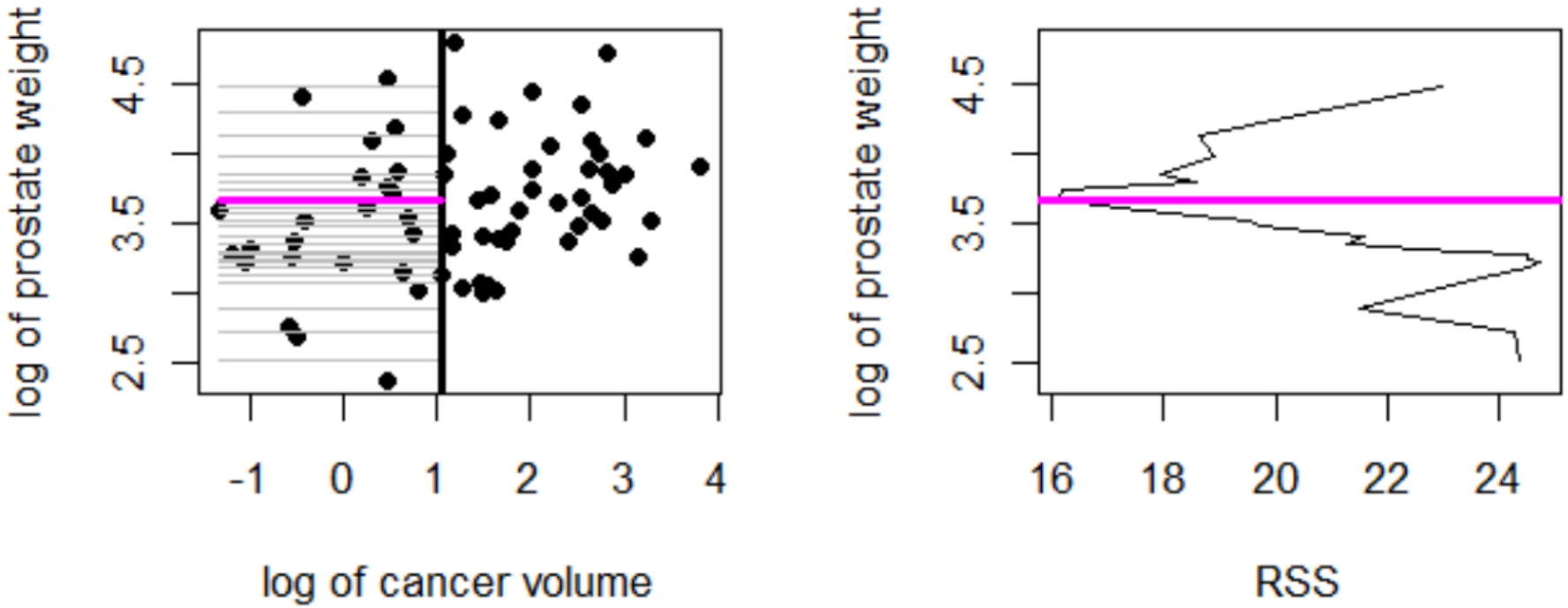
Best vertical split is at 1.05 with RSS = 61.76.

Adele Cutler
Utah State University



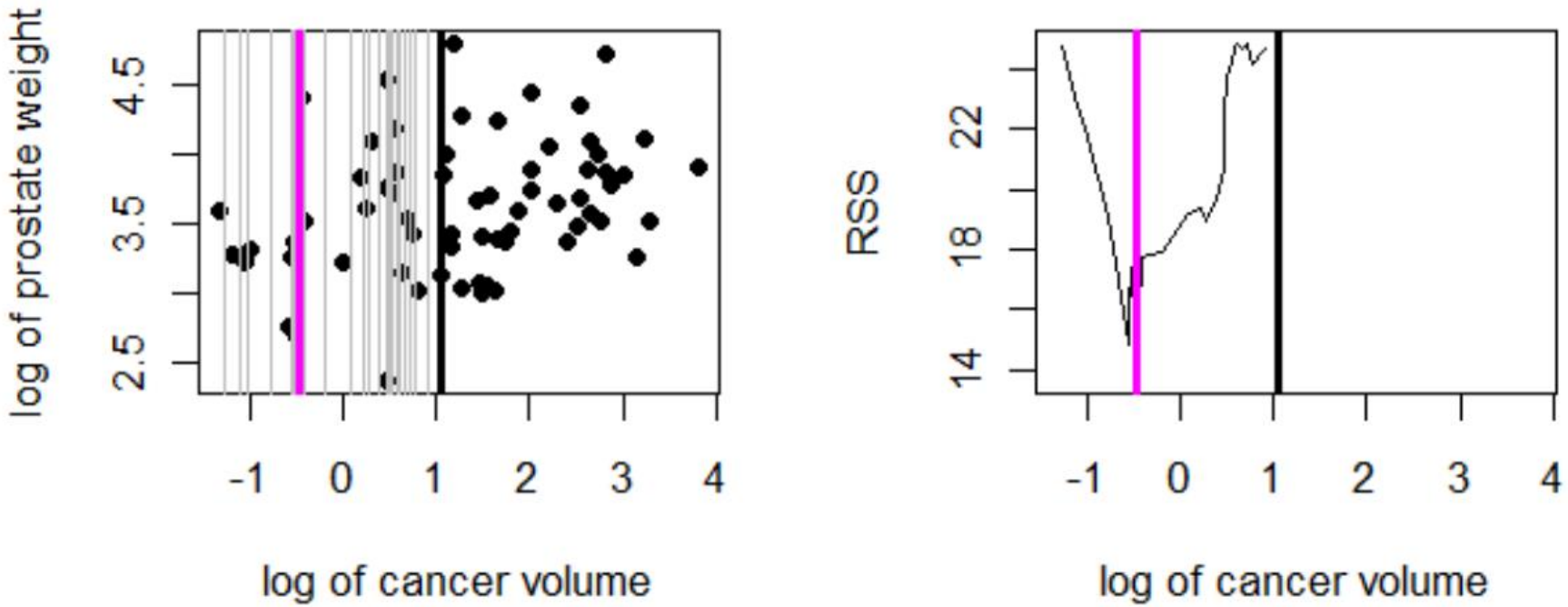


Adele Cutler
Utah State University



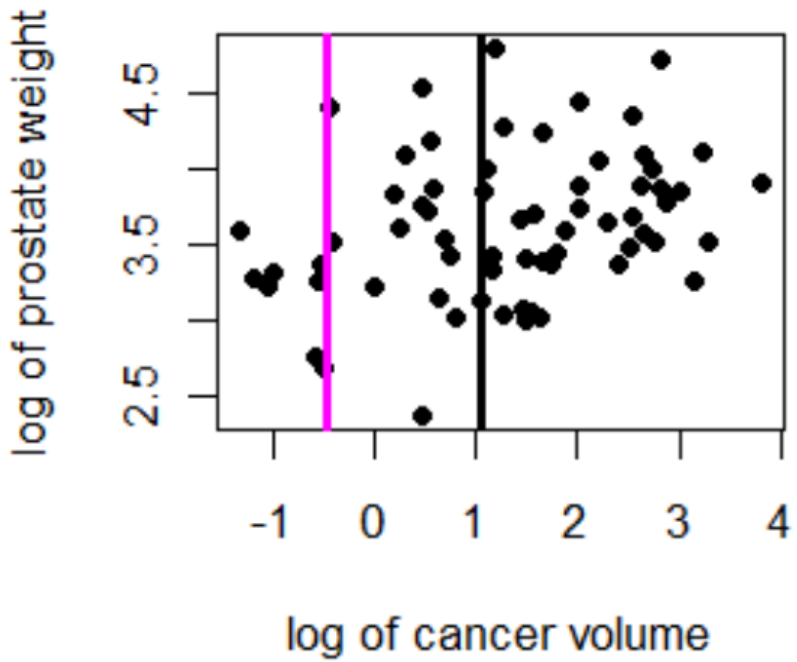
Best horizontal split is at 3.66 with RSS = 16.11.

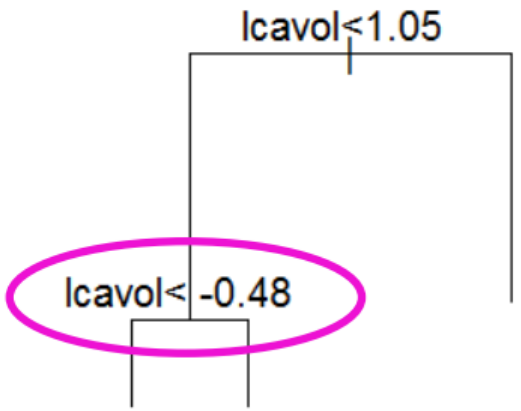
Adele Cutler
Utah State University



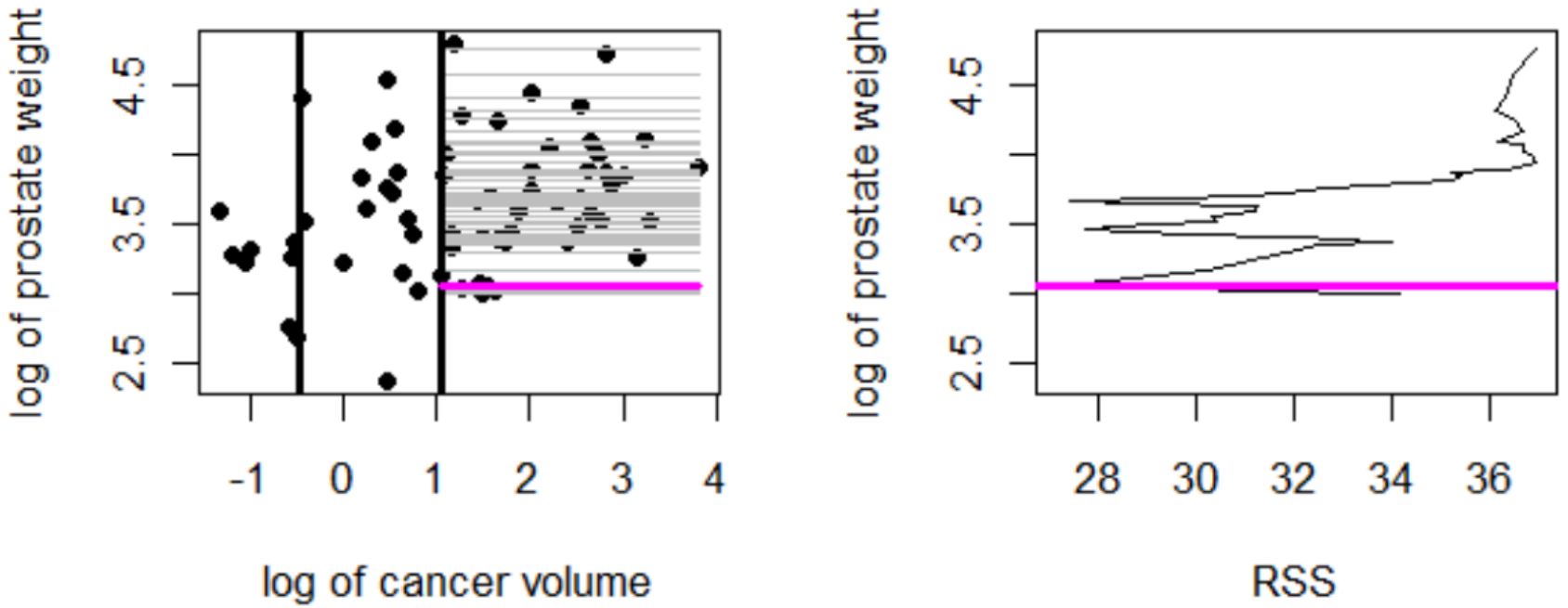
Best vertical split is at -.48 with RSS = 13.61.

Adele Cutler
Utah State University



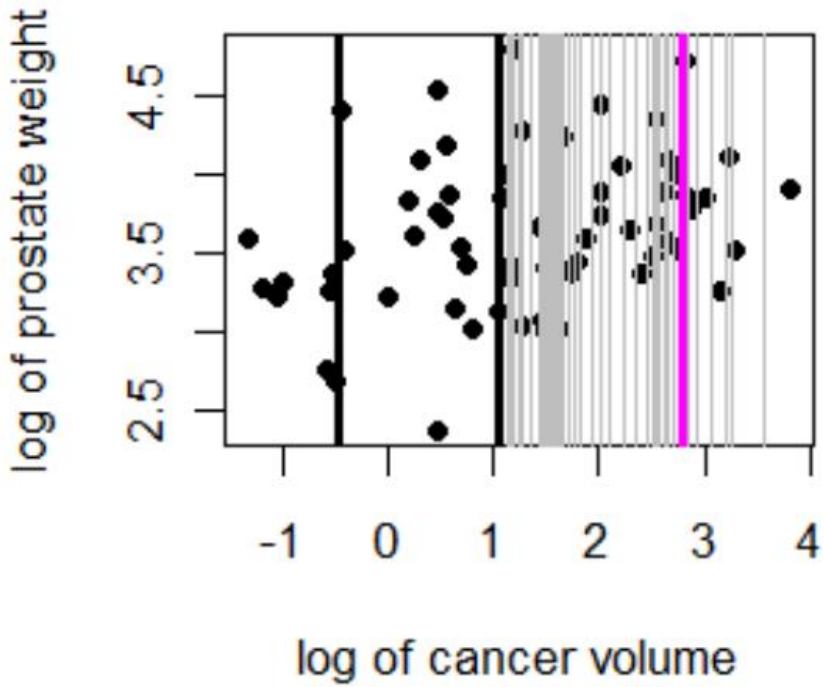


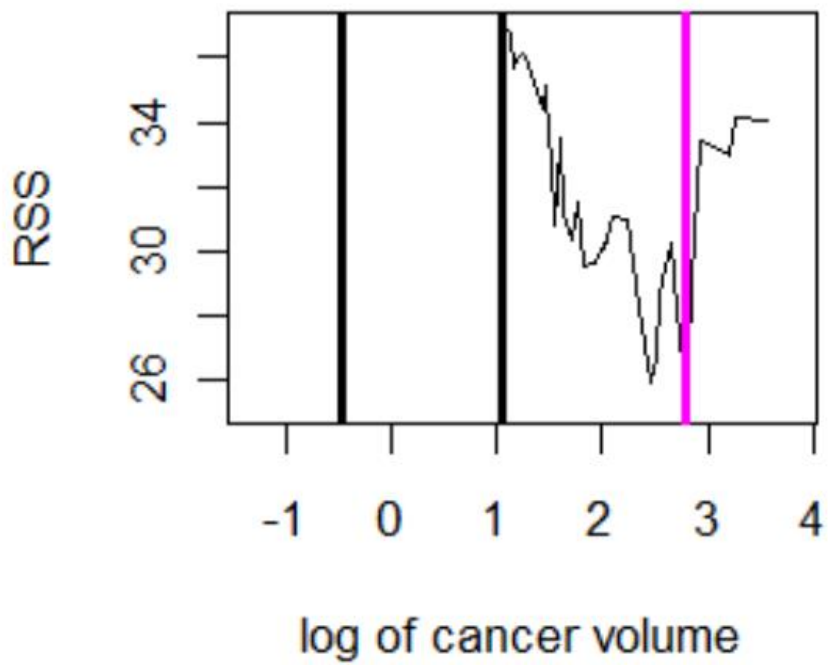
Adele Cutler
Utah State University



Best horizontal split is at 3.07 with RSS = 27.15.

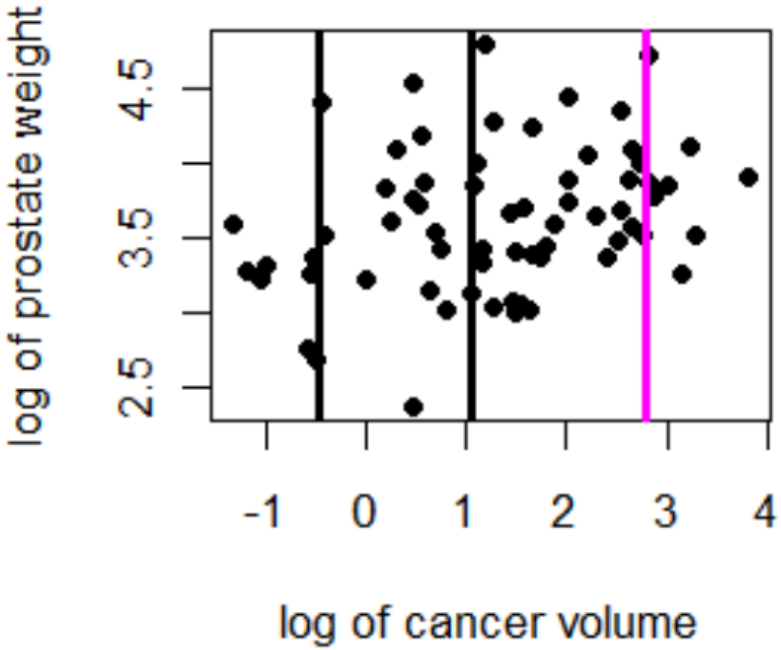
Adele Cutler
Utah State University

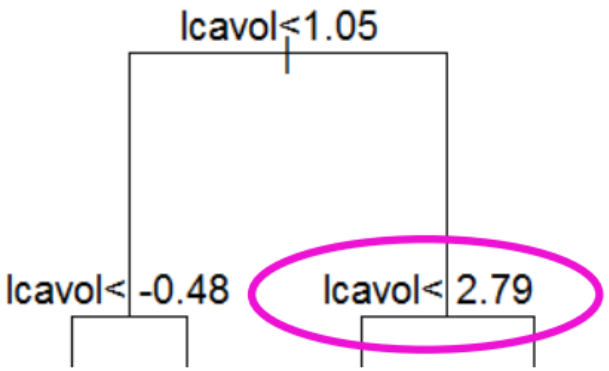




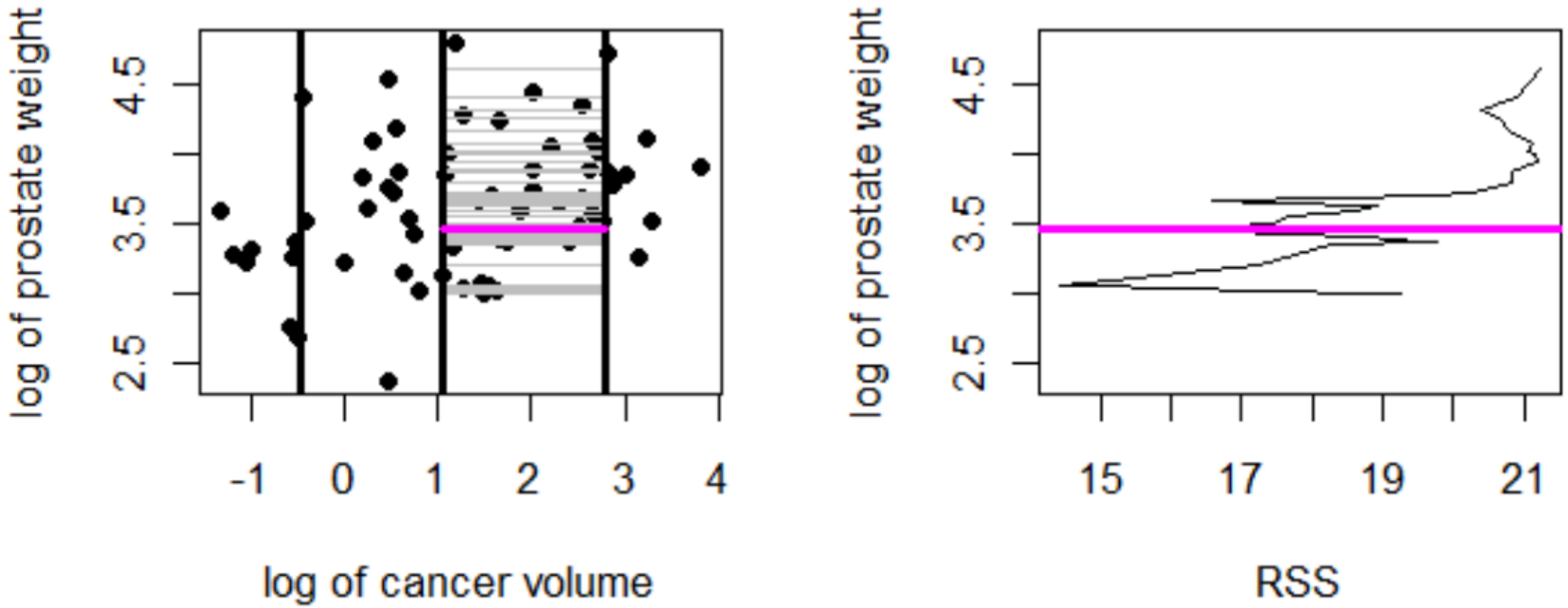
Best vertical split is at 2.79 with RSS = 25.11.

Adele Cutler
Utah State University



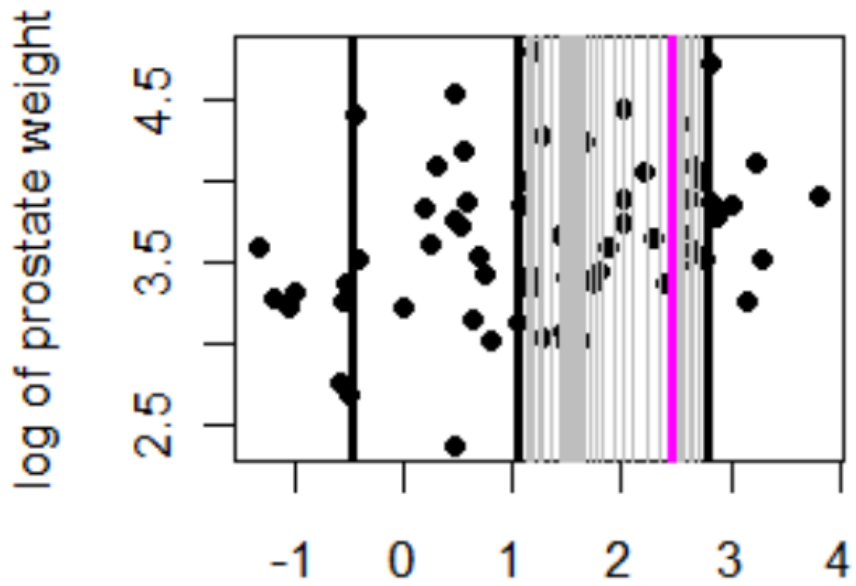


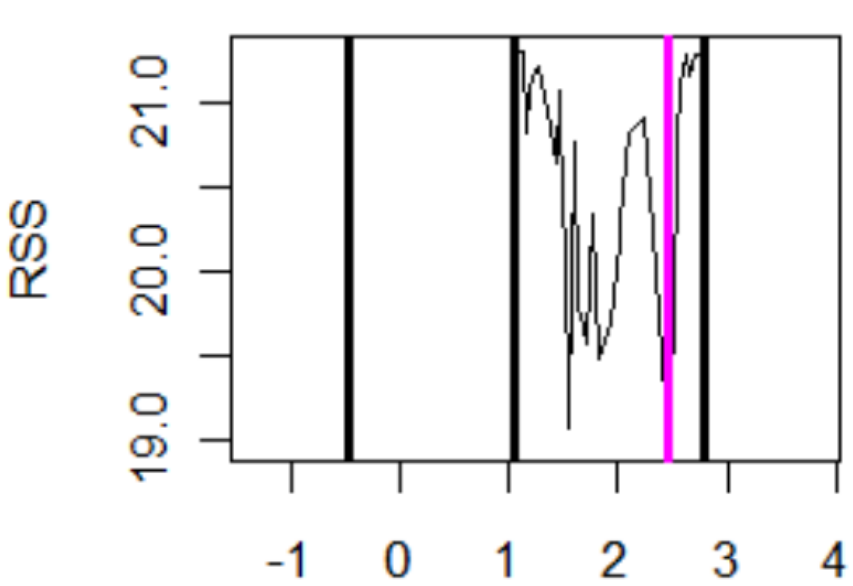
Adele Cutler
Utah State University



Best horizontal split is at 3.07 with RSS = 14.42, but this is too close to the edge. Use 3.46 with RSS = 16.14.

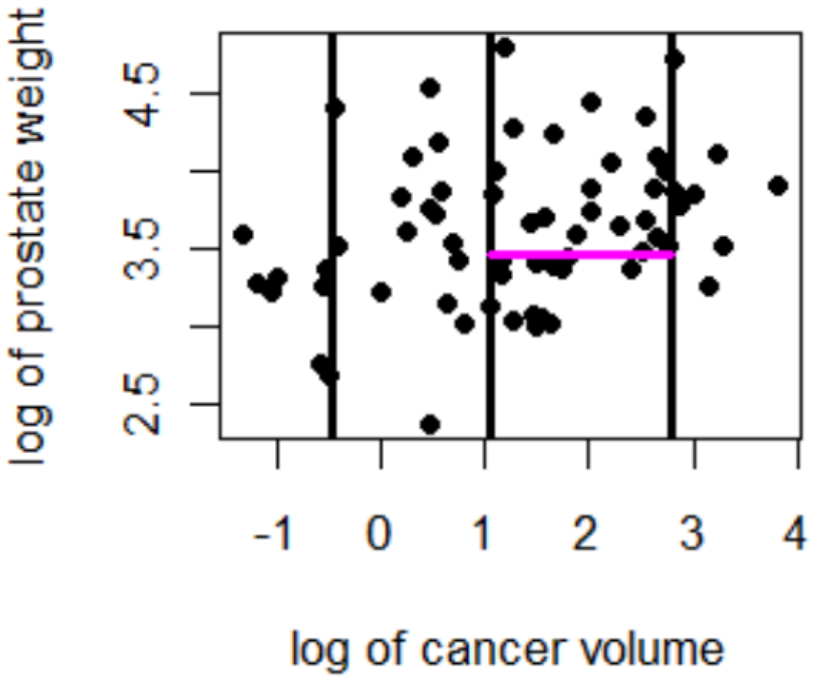
Adele Cutler
Utah State University

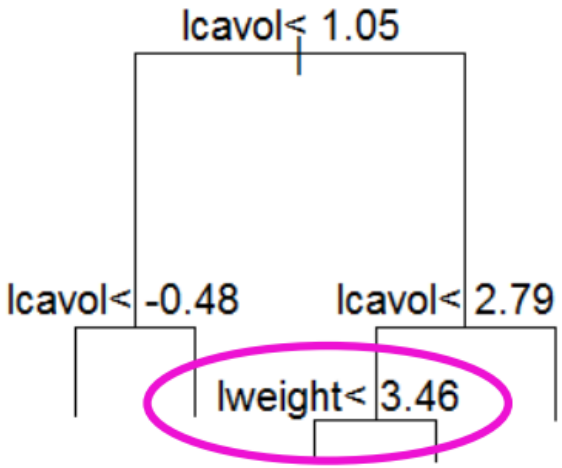




Best vertical split is at 2.46 with RSS = 18.97.

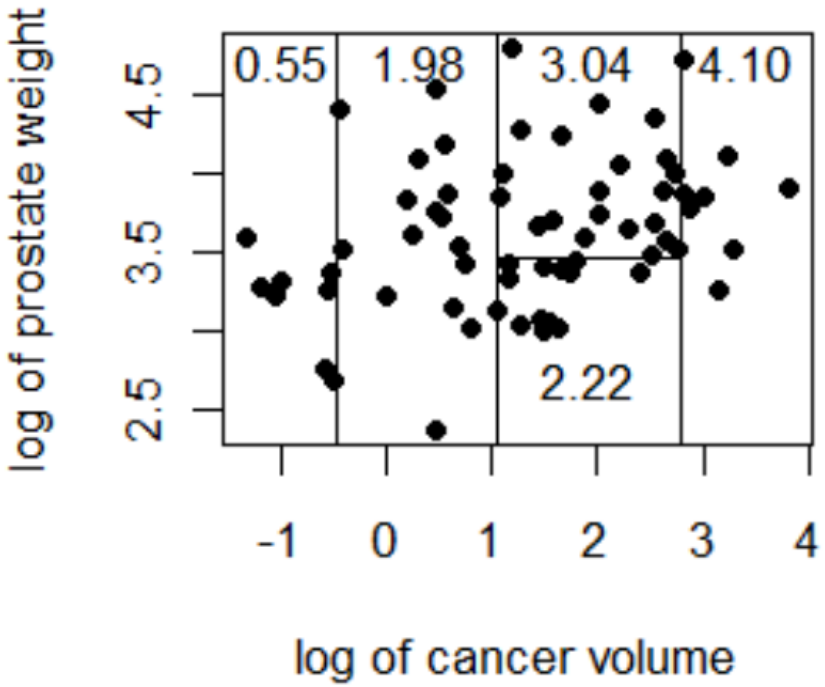
Adele Cutler
Utah State University

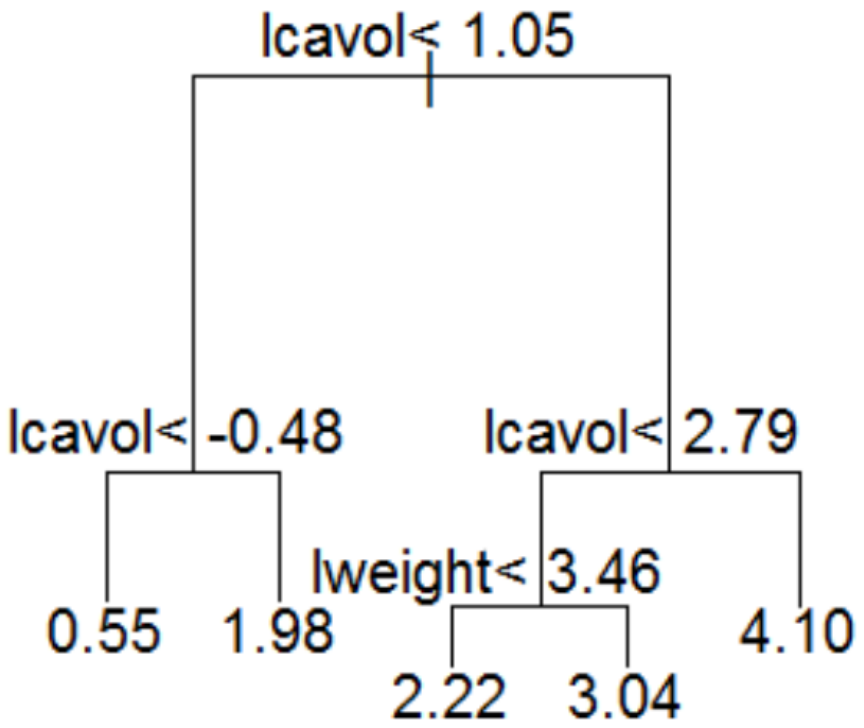




Adele Cutler
Utah State University

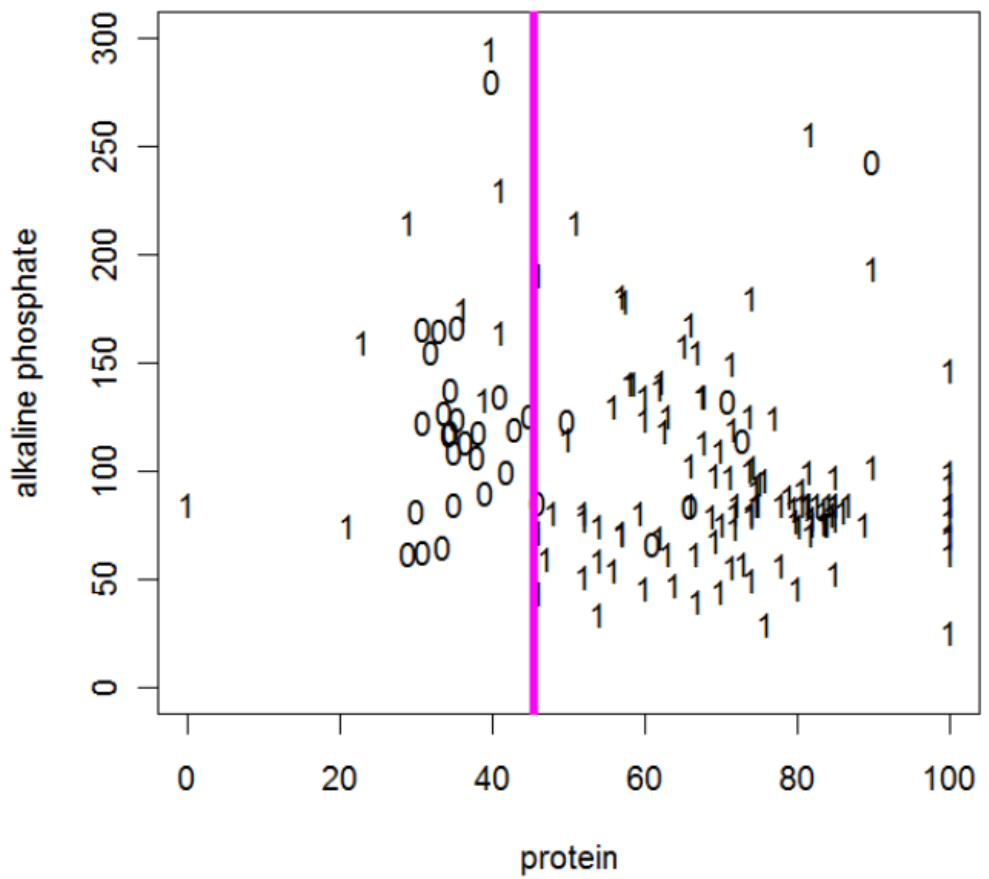
Final Tree

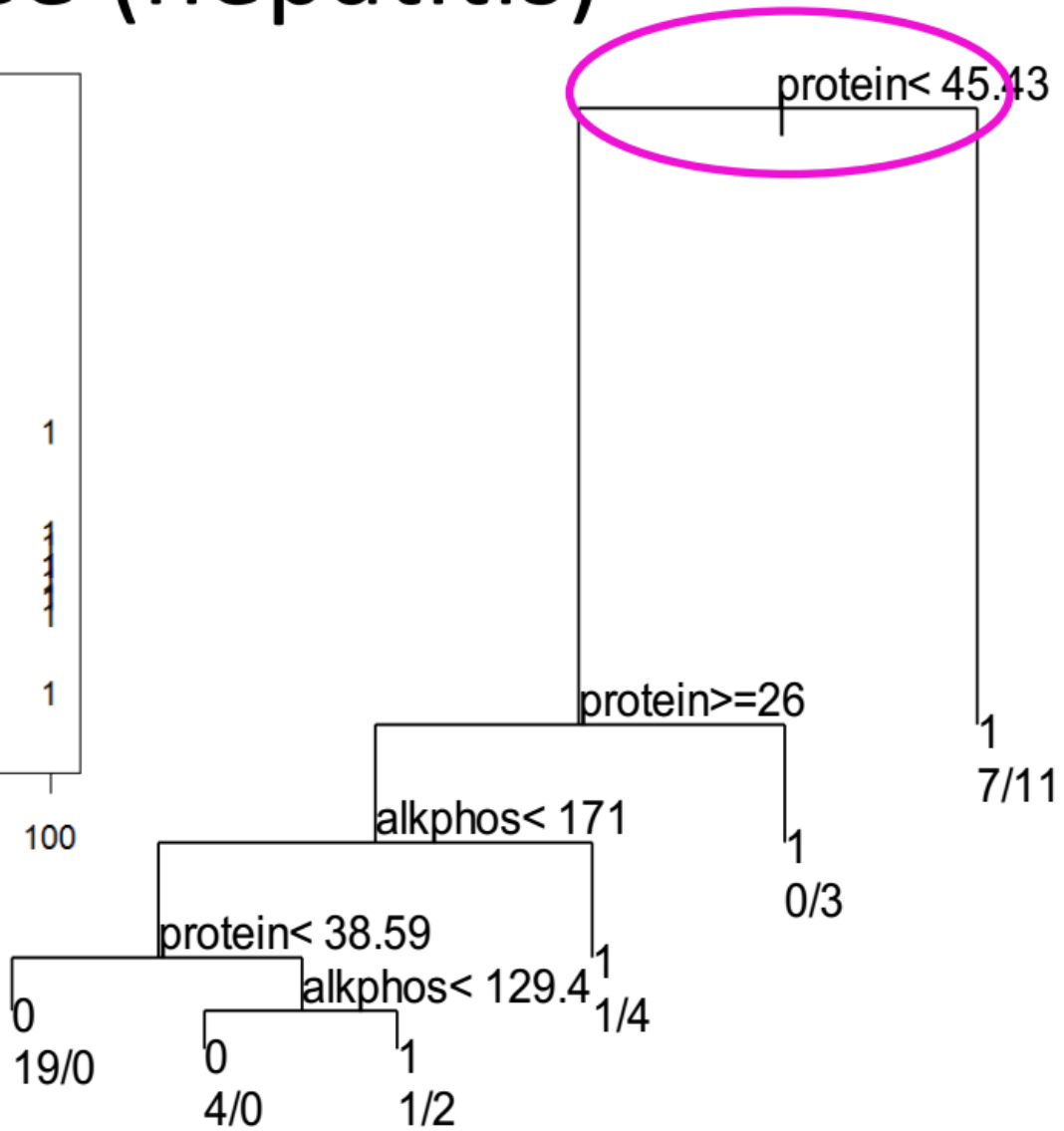




Adele Cutler
Utah State University

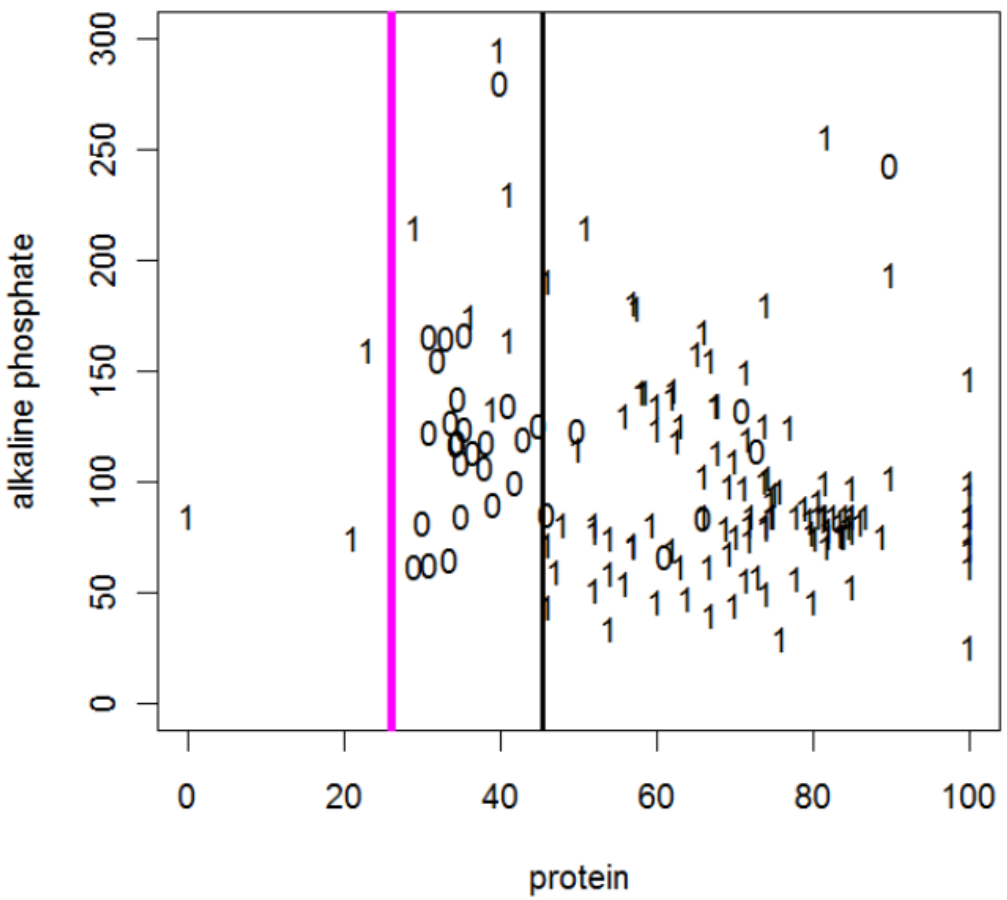
Classification tree (hepatitis)

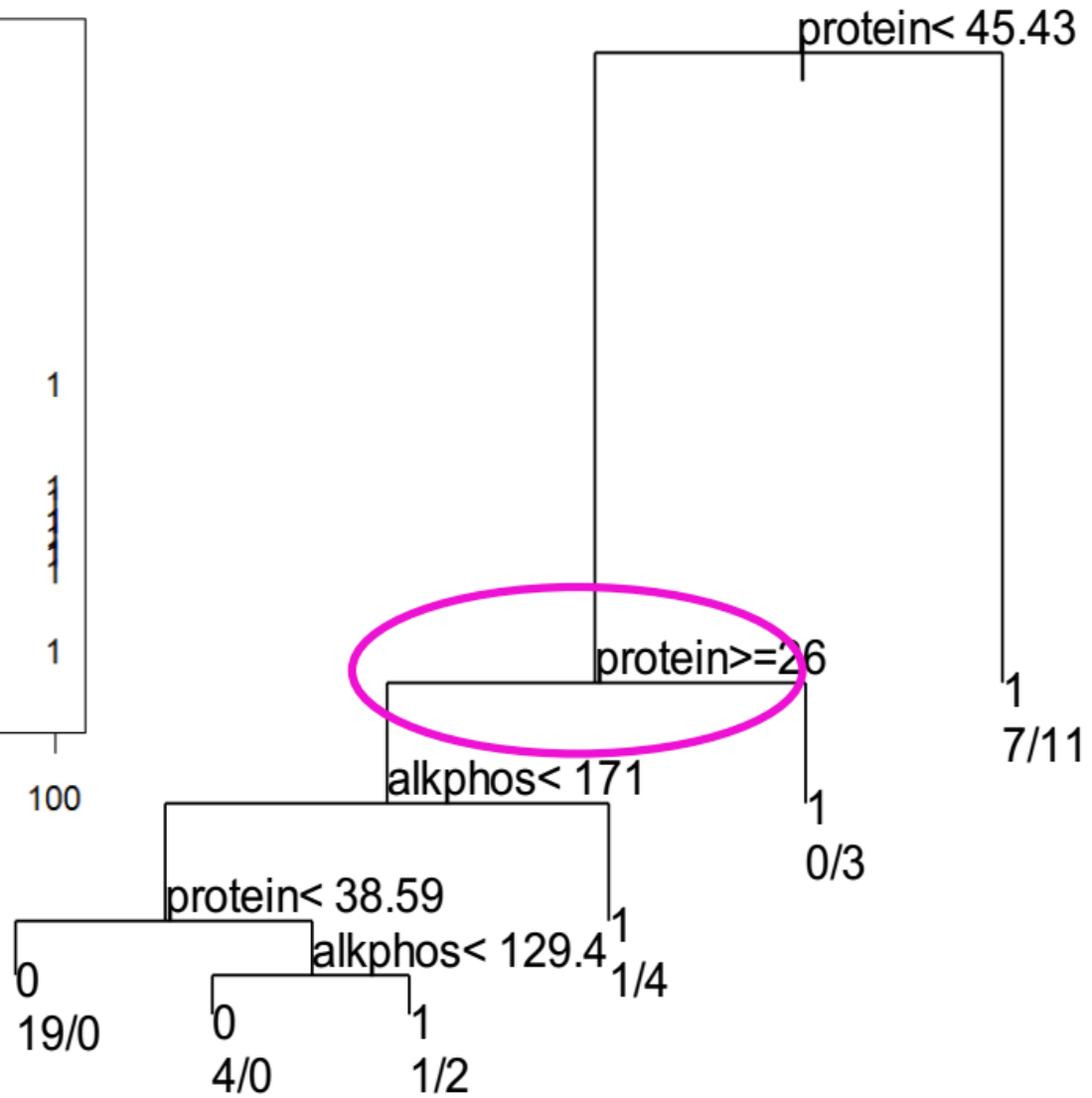




Adele Cutler
Utah State University

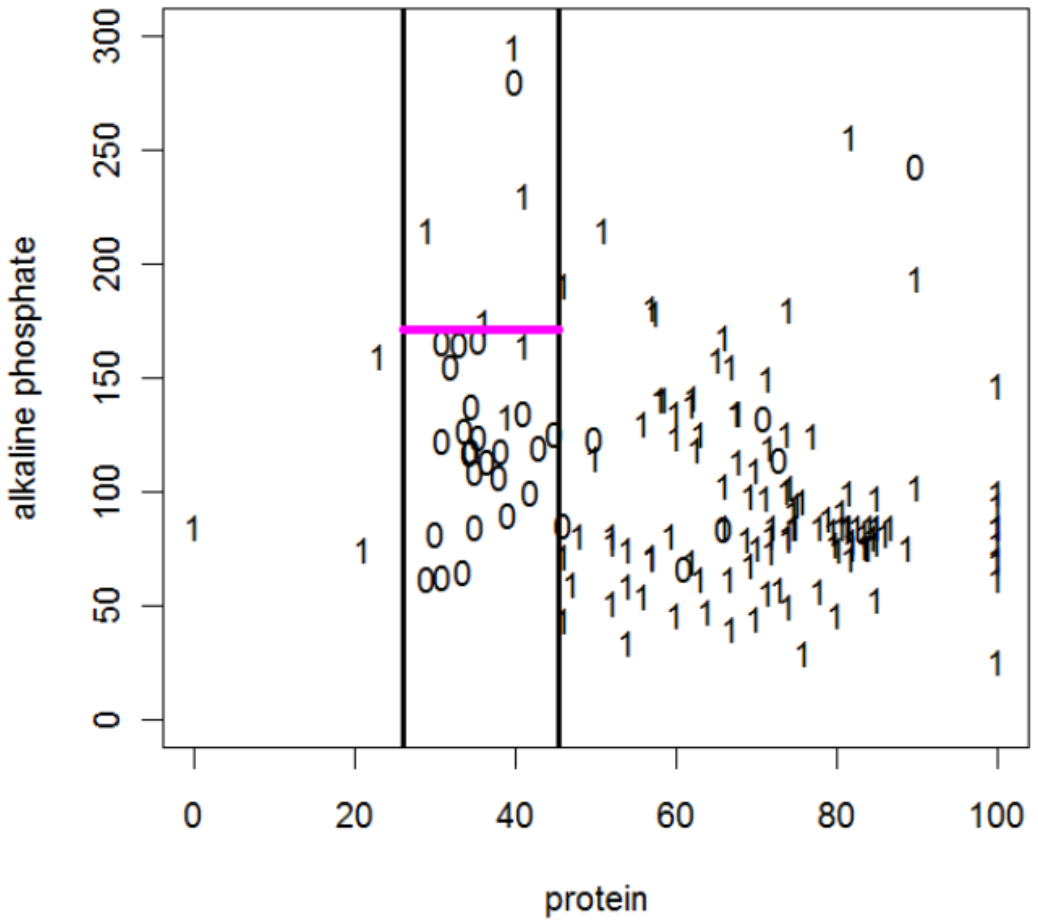
Classification tree (hepatitis)

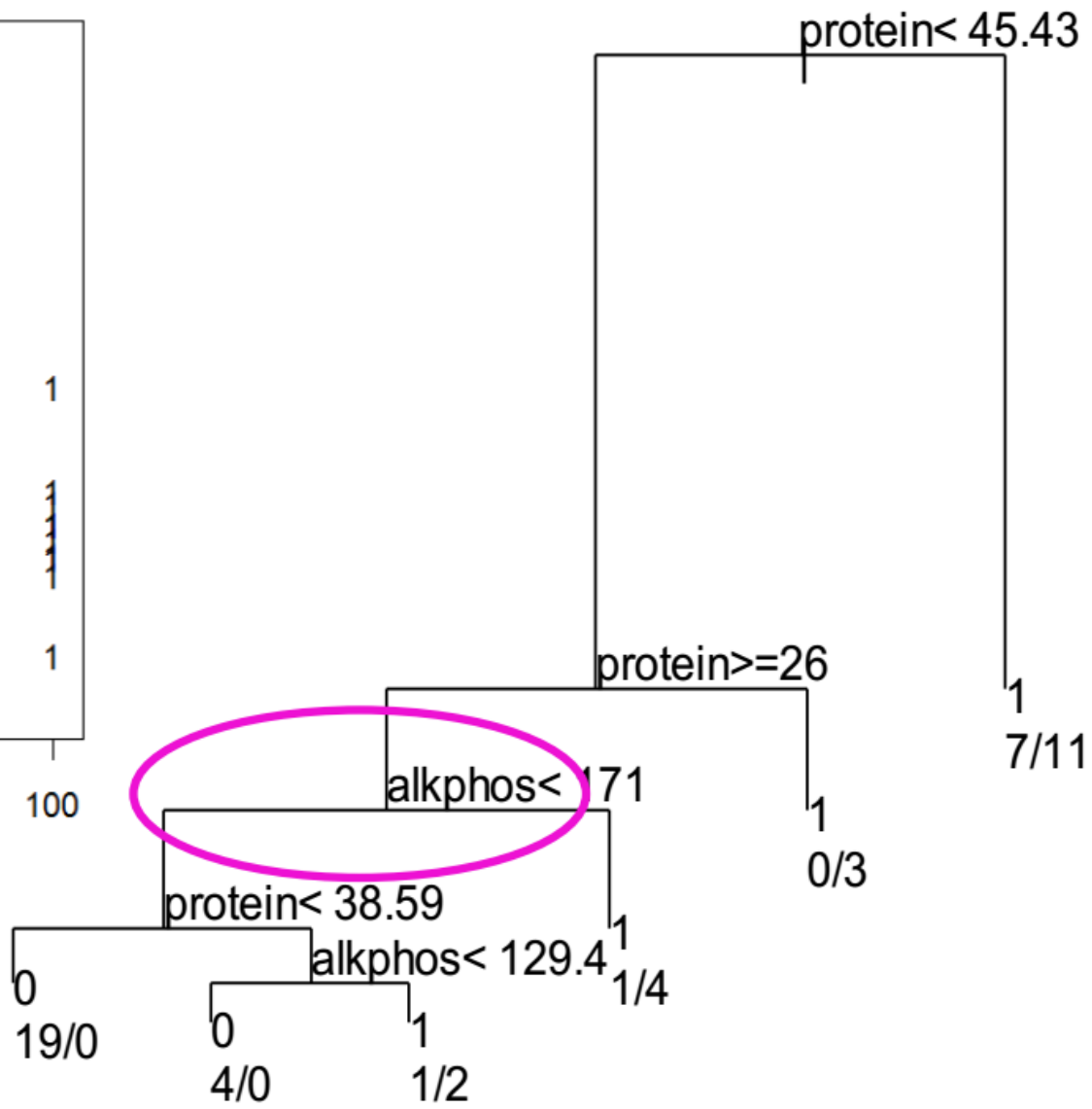




Adele Cutler
Utah State University

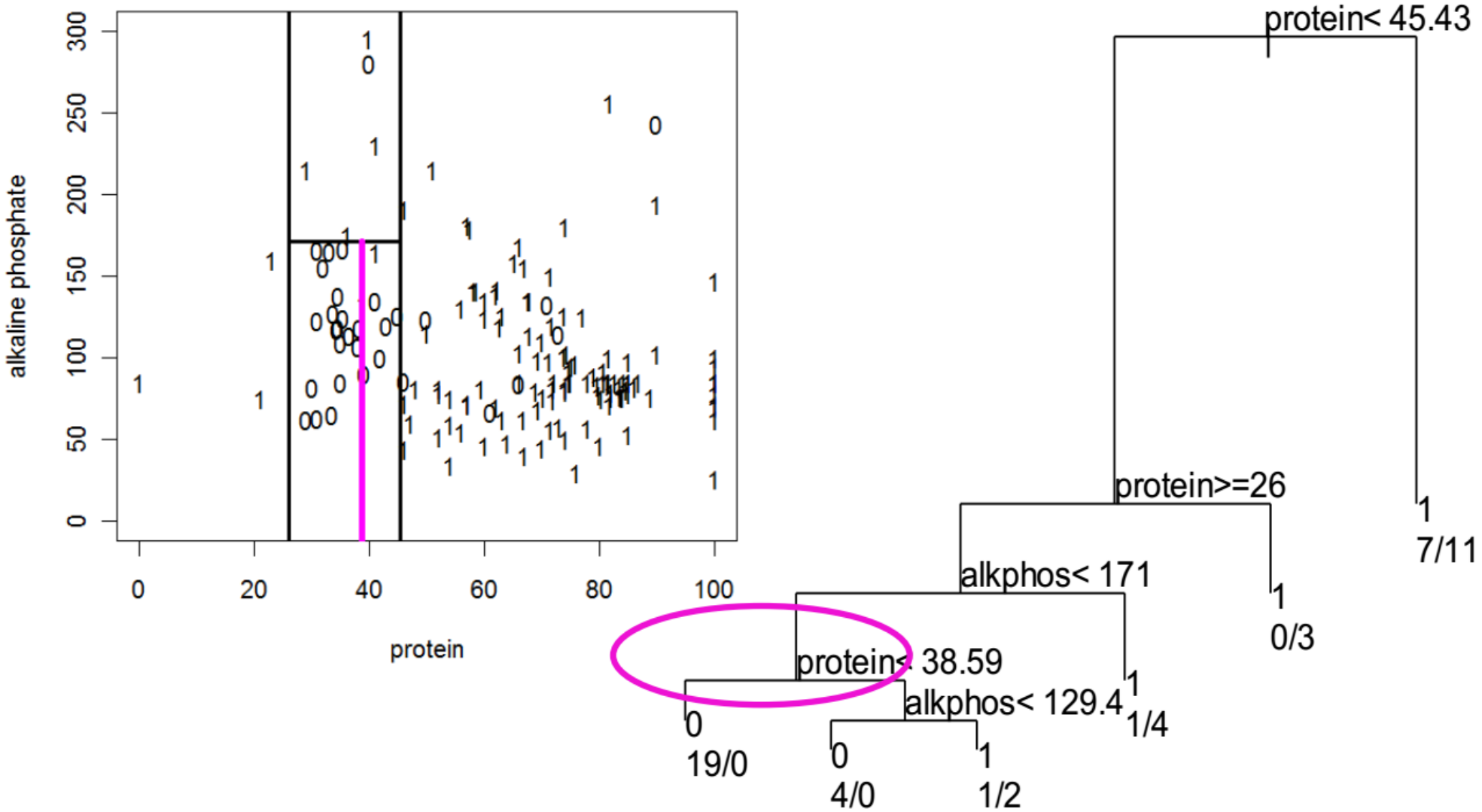
Classification tree (hepatitis)





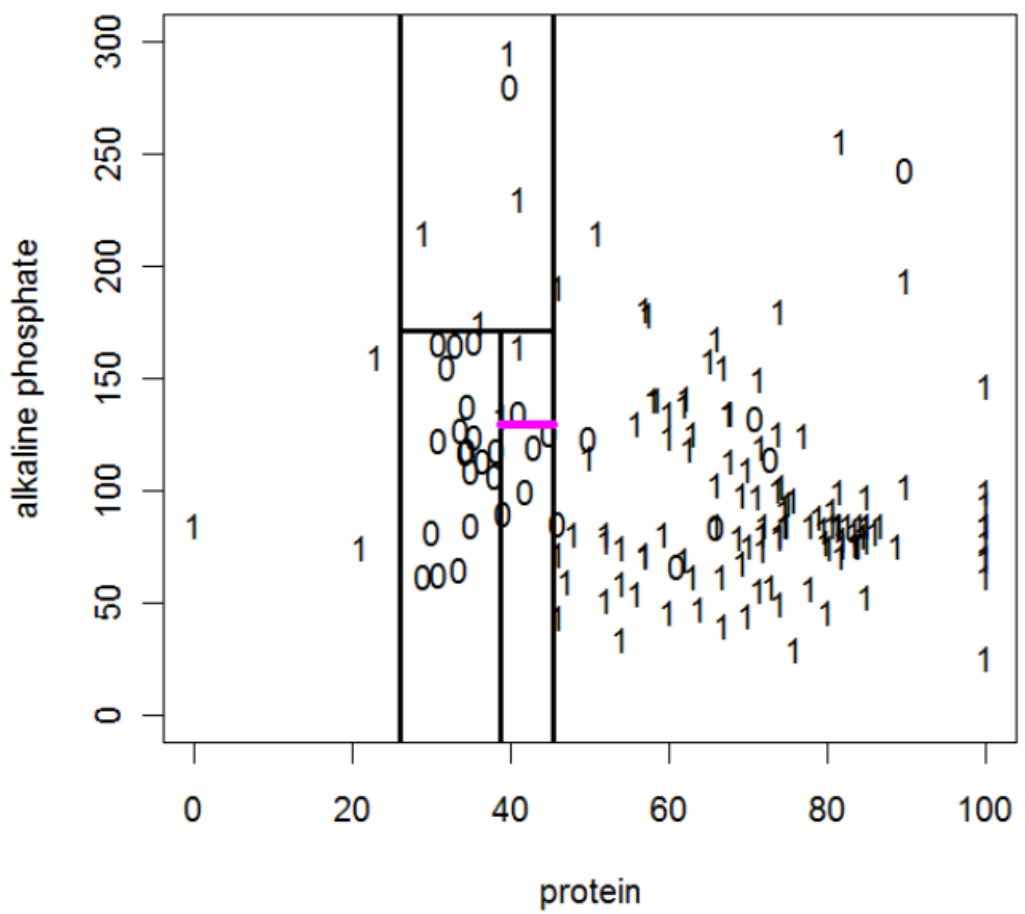
Adele Cutler
Utah State University

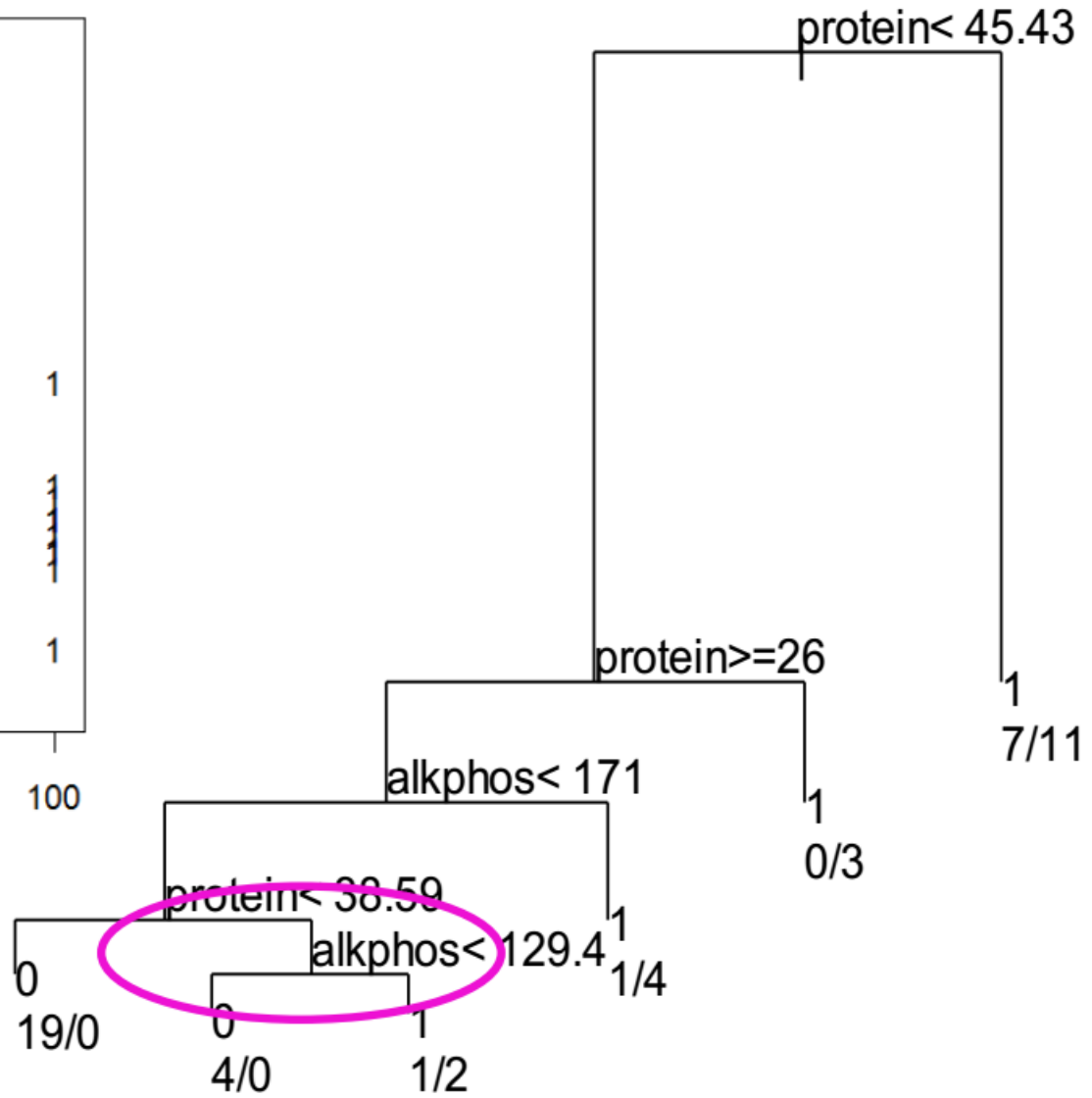
Classification tree (hepatitis)



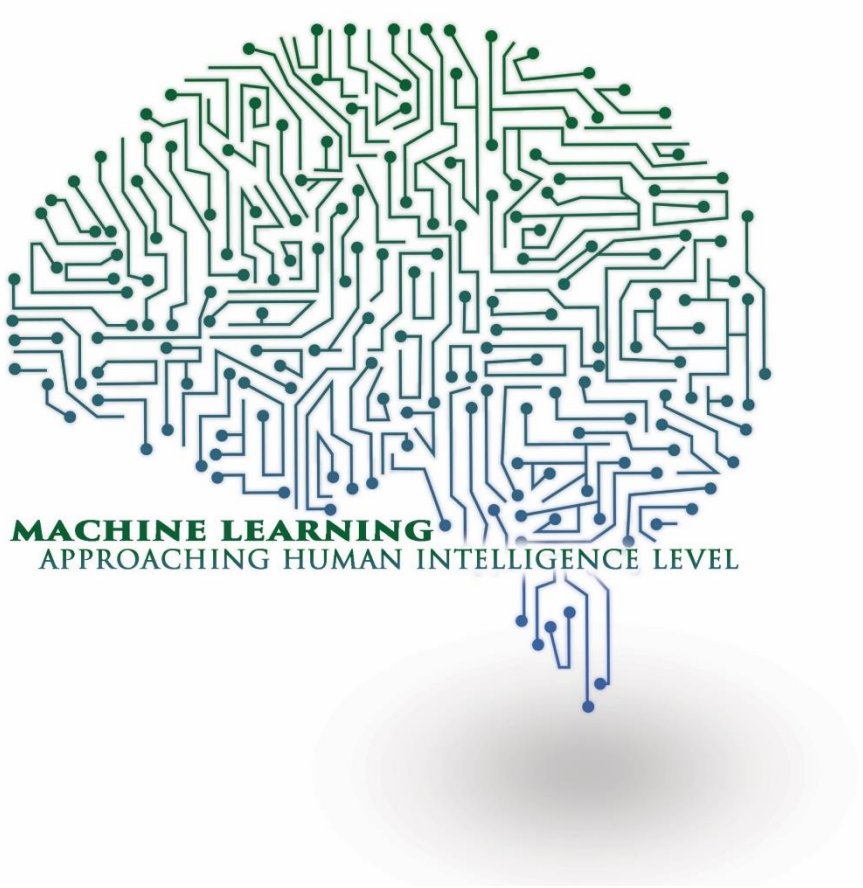
Adele Cutler
Utah State University

Classification tree (hepatitis)





Adele Cutler
Utah State University



Random forest

Random forest