Principles of Statistical Data Mining
Introduction to Classification Trees

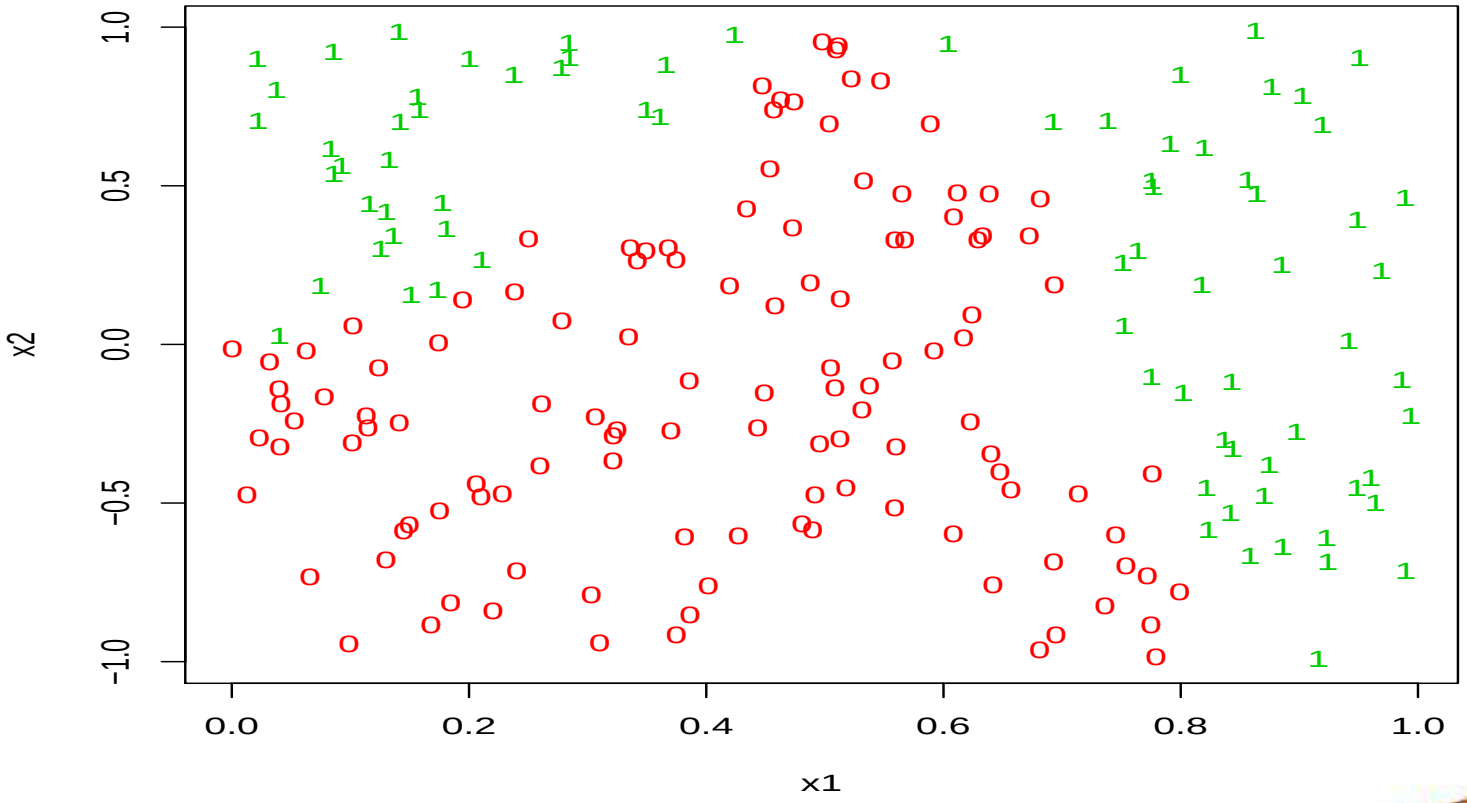
Prof Ernest Fokoué

Center for Quality and Applied Statistics
Rochester Institute of Technology

Spring 2013

Two Dimensional Motivating Example

Let $x_1 \in [0, 1]$ *and* $x_2 \in [-1, 1]$ *and consider following scatter*



How can a classification tree be built to minimize misclassification rate?

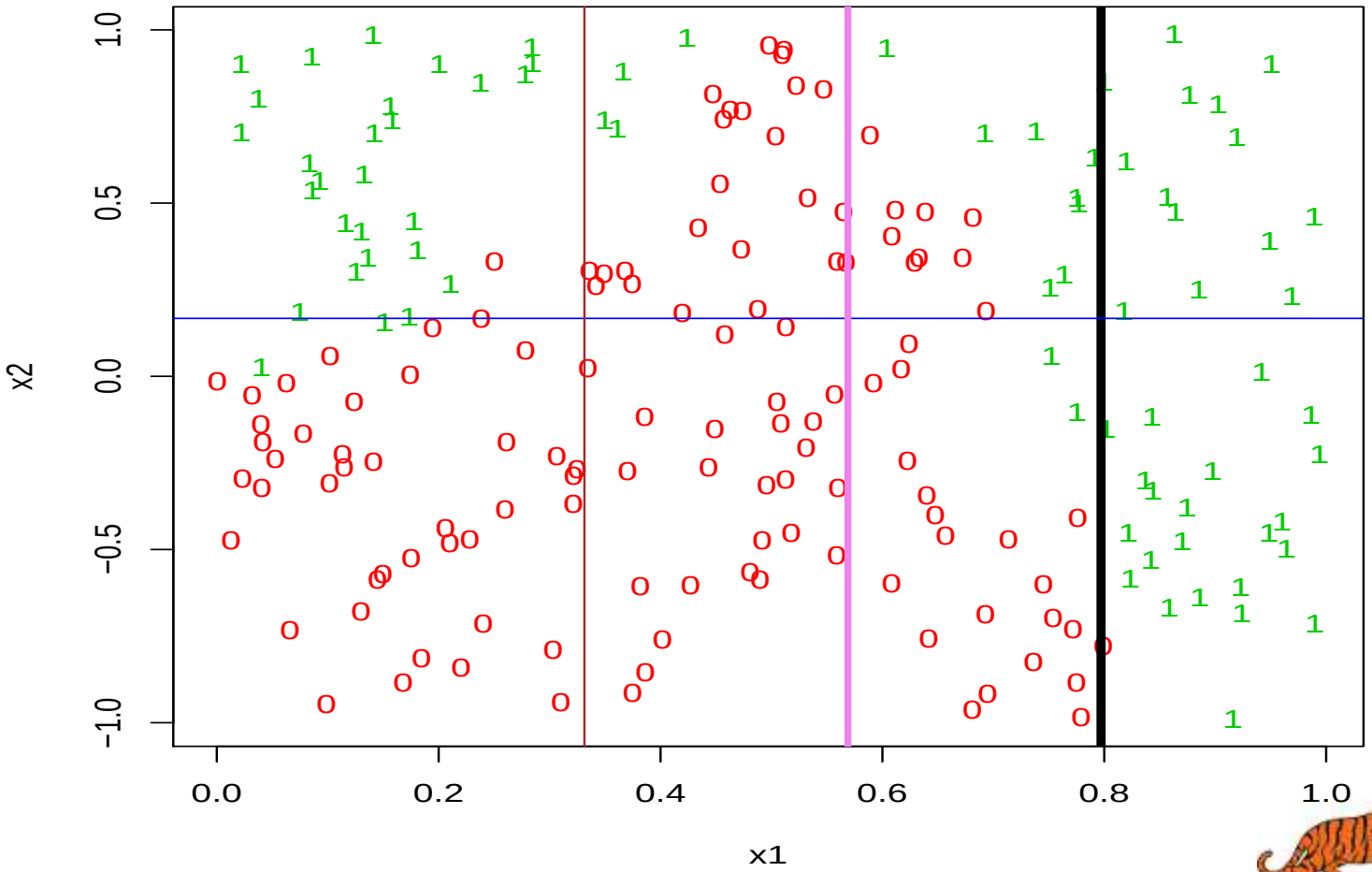
Intuition of Classification Trees

- *Tree Classification Principle:* *All the observations that fall in in the same node of the tree will be assigned the same class label, namely the majority class of that node.*

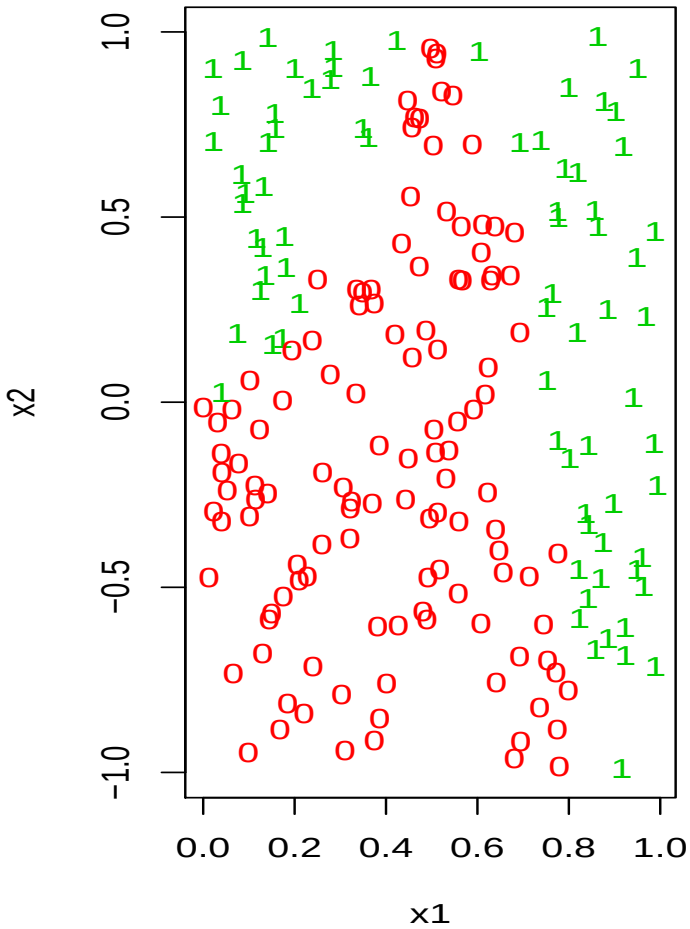
- *Classification Trees Steps:* *Let* T *denote the tree being built, and let* $\mathcal{X}$ *denote the input space.*
  - *Decide on (choose) some criterion like the node impurity* $\mathrm{Q}_\ell(\mathrm{T})$
  - *Partition* $\mathcal{X}$ *into* q *regions* $\mathrm{R}_1, \mathrm{R}_2, \cdots, \mathrm{R}_\mathrm{q}$ *according to* $\mathrm{Q}_\ell(\mathrm{T})$
  - *At each node, find the majority class*
  - *At each node, assign the same label (majority class) to all the observations that fall in that node. (This is why trees are referred to as piecewise constant estimators)*

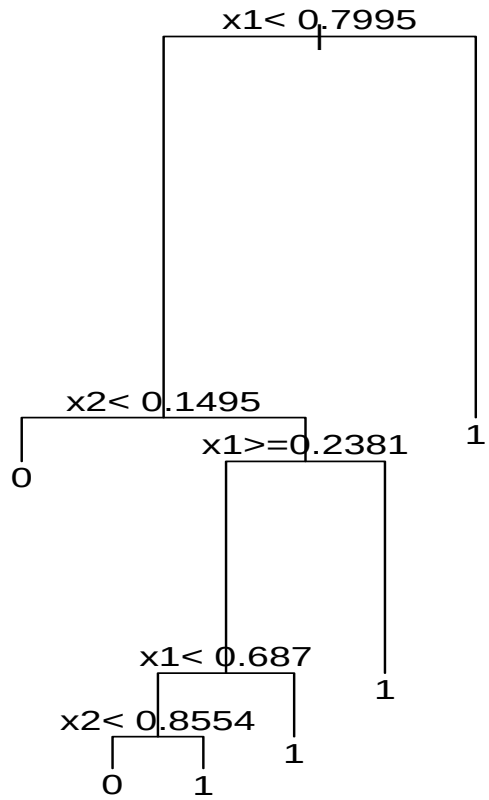
- *Tree Classification:* *The estimated class of a vector* $\mathbf{x}$ *is the majority class of the node in which* $\mathbf{x}$ *falls.*

Visual demonstration on Classification Trees



Visual demonstration on Classification Trees





1. $\mathcal{D} = \Big\{(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)\Big\}$, *with* $\mathbf{x}_i \in \mathcal{X}^p$, $Y_i \in \{1, \cdots, g\}$.
2. *Let* T *denote the tree represented by the partitioning of* $\mathcal{X}$ *into* q *regions* $R_1, R_2, \cdots, R_q$ *such that*

$$T = \cup_{\ell=1}^{q} R_\ell$$

3. *Given a new point* $\mathbf{x}^*$ *in node* ℓ*, its predicted response* $\hat{Y}^*$ *is*

$$\hat{Y}^*_{\texttt{Tree}} = \hat{f}_{\texttt{Tree}}(\mathbf{x}^*) = \arg \max_{j \in \{1, \cdots, g\}} \{p_{j\ell}\}$$

where

$$p_{j\ell} = \frac{1}{|R_\ell|} \sum_{\mathbf{x}_i \in R_\ell} I(Y_i = j)$$

estimates the proportion of node ℓ *observations that belongs to class* j*. Clearly,* $\mathbf{x}^*$ *is assigned the same label/class as all the points in its node.*

Measures of Node Impurity

Three of the most commonly used impurity measures are:

- *Misclassification rate:*

$$\frac{1}{|R_\ell|}\sum_{i\in R_\ell} I\left(y_i \neq \arg\max_{j\in\{1,\cdots,g\}}\{p_{j\ell}\}\right)$$

- *Gini Index:*

$$\sum_{j\neq j'} p_{j\ell}p_{j'\ell} = \sum_{j=1}^{g} p_{j\ell}(1-p_{j\ell})$$

- *Cross Entropy or Deviance:*

$$-\sum_{j=1}^{g} p_{j\ell}\log p_{j\ell}$$

Question: How does one decide which impurity measure to use?

Measures of Node Impurity in Binary Classification

For binary classification, we have

- *Misclassification rate:*

$$1 - \max(p, 1-p)$$

- *Gini Index:*

$$2p(1-p)$$

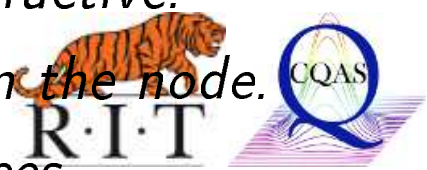
- *Cross Entropy or Deviance:*

$$-p \log p - (1-p) \log(1-p)$$

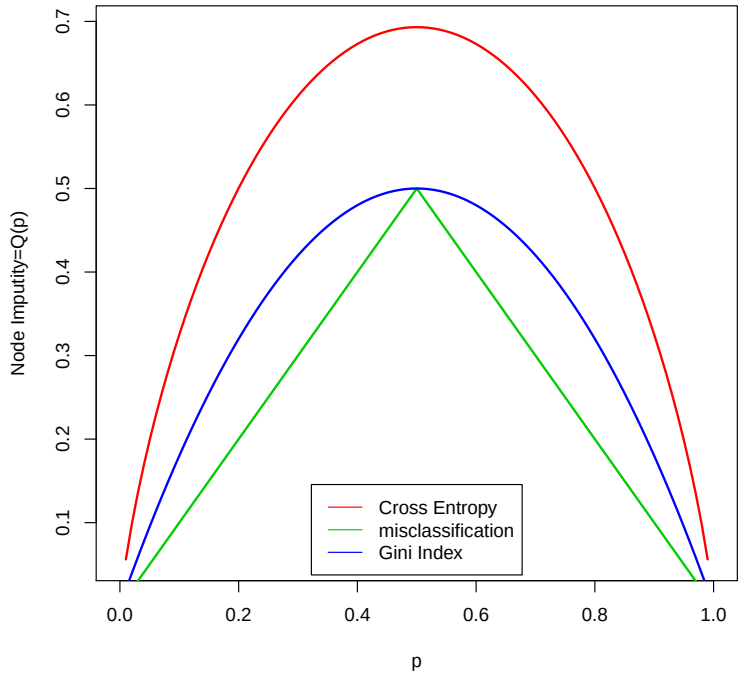
Question: Which impurity measure to use?

- *Gini and Cross Entropy are differentiable whereas the Misclassification rate is not. This makes them computationally more attractive.*
- *Gini and Cross Entropy are more sensitive to changes in the node.*

Recommendation: Use Gini or Cross Entropy for growing trees.



Measures of Node Impurity in Binary Classification



Remark: *From the above, it is easy to see that node impurity reaches its maximum in binary classification when the two classes in that given node are equally represented. The best split however, happens when one class has* 100%, *and the other has* 0%. *This confirms our intuition on purity.*

Building Classification and Regression Trees

Amongst the most crucial aspects of tree building are the following

- *Splitting criterion: How does one decide how to split an intermediary node?*
  - *The Gini index which measures the impurity is the most commonly used criterion. Essentially, the variable (attribute) that ends up driving the split along with its cut-off, are chosen so as to minimize the impurity*
  - *The Information Gain, which is derived from entropy is another criterion commonly used*
- *Terminal Node (leaf): How does one decide to declare a node a terminal node?*
- *Regional/Nodal Estimate of the response: Once the recursion gets down to a leaf (terminal node), how is the estimate of the response determined there?*

More on Splits in Tree Building

Since tree-based methods proceed by recursively partitioning the input space, it makes sense that the central idea in classification and regression trees is the concept of split.

- *For continuous and counts variables, splits of the type* $x_j < t$ *versus* $x_j \geqslant t$ *are considered, where the choice of* t *is guided by the minimization of the nodal impurity measure* $Q_\ell(T)$
- *For ordered factors, splits of the type* $x_j < t$ *versus* $x_j \geqslant t$ *are considered, where the choice of* t *is again guided by the minimization of the nodal impurity measure* $Q_\ell(T)$
- *For general factors (like nominal), the levels are divided into two classes, so that if the factor has* L *levels, one has to consider* $2^{L-1} - 1$ *possible splits if order and empty splits are disallowed.*

Appeal/Strengths of Classification Trees

- *The ever increasing popularity of classification trees is primarily due to their interpretability.*
- *Mix-typed input spaces: Decision trees are also appealing because they can naturally handle input spaces containing variables (attributes) of various different types*
- *Implicit variable selection: Trees naturally only build their rules around the most important variables, thereby providing (implictly/indirectly) a handle on variable importance*
- *Unified regression and classification: Seamless and identical representation of both classification and regression*
- *Straightforward prediction: Performing prediction with trees is as simple as traversing it from the root to a terminal node (leaf).*

Weaknesses/Limitations of Classification Trees

- *Instability:* Arguably their most significant weakness, the instability of trees is easily evidence by seeing how little changes in the data can lead to big changes in the tree. Fortunately, ensemble techniques like bagging, boosting and random forest help reduce the variance of trees substantially.
- *Overfitting:* Trees will be default tend to overfit the data in their quest to minimize the misclassification rate on the present data. Fortunate, trees can be regularized by various pruning techniques that help achieve bias-variance trade-off.
- *Lack of Smoothness:* This is clealy due to the fact trees are piecewise constant estimator of the true underlying function. As a result of the piecewise constantness, trees are definitely not smooth. This is better appreciated when one sees regression trees.
- *Categorical variables and missing values:* These two issues are not as crucial as the previous ones.

Study of Diabetes among Pima Indian Women

Motivating Example: *A study originally published by the National Institute of Diabetes and Digestive and Kidney Diseases sought to determine the relationship between the incidence of diabetes in Pima Indian Women and some specific medical and personal characteristics. A population of women at least 21 years old and of Pima Indian heritage living near Phoenix were chosen and were tested for diabetes.*

`npreg`	*Number of pregnancies*
`glu`	*Plasma glucose concentration*
`bp`	*Diastolic blood pressure (mm Hg)*
`skin`	*Triceps skin fold thickness (mm)*
`bmi`	*Body mass index kg/m2*
`ped`	*Diabetes pedigree function*
`age`	*Age (years)*

The response is type : Yes = diabetic; No = Non diabetic.

Study of Diabetes among Pima Indian Women

Example: *The dataset pima-tr.csv from the R package MASS has* $\mathtt{n} = 200$ *observations. A portion of pima-tr.csv is:*

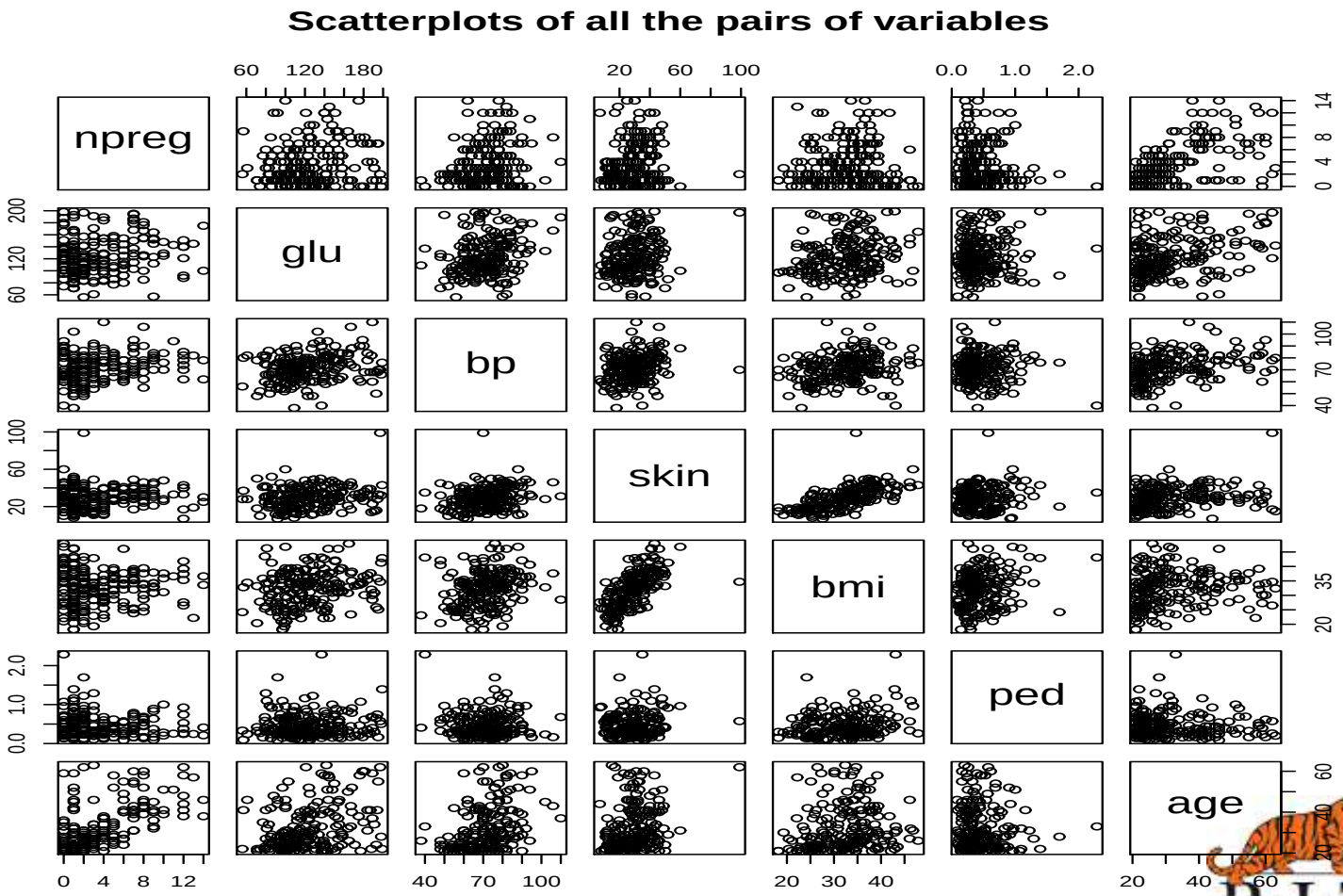
npreg	glu	bp	skin	bmi	ped	age	type
5	86	68	28	30.2	0.364	24	No
7	195	70	33	25.1	0.163	55	Yes
5	77	82	41	35.8	0.156	35	No
5	97	76	27	35.6	0.378	52	Yes

- *Are all the above variables significantly related to the incidence of diabetes in Pima Indian women?*
- *Can the attributes be summarized into uncorrelated meaningful concepts in decreasing order of importance?*

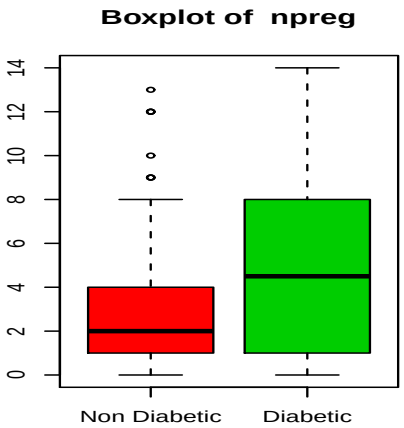
Modeling: *How can a classification tree help build an interpretable and predictively optimal representation of the data?*

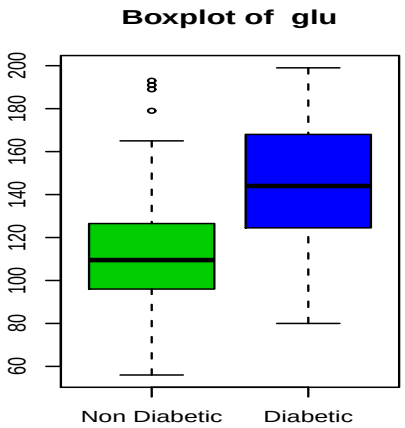
Exploratory Data Analysis Prior to Classification

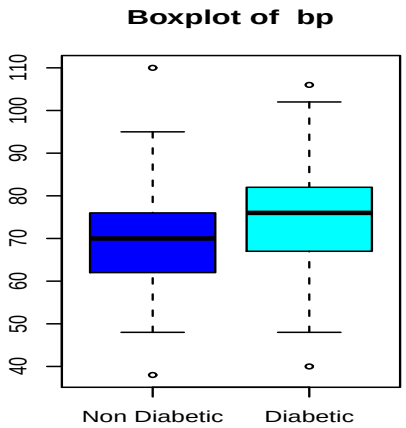
Classifying Pima Indian Diabetes: Let's have a look at the Pima Indian Diabetes data set prior to building a classifier for it.

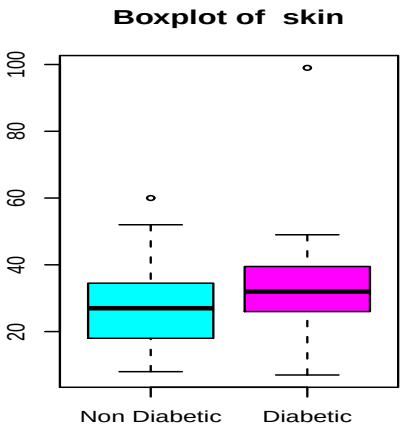


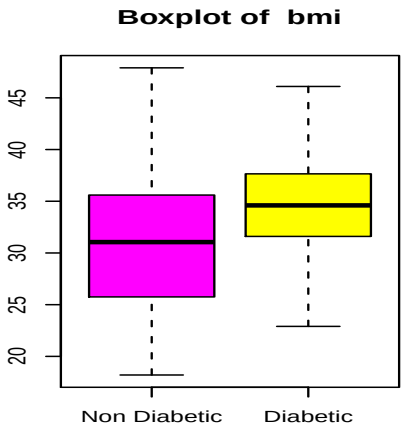
Exploratory Data Analysis Prior to Classification

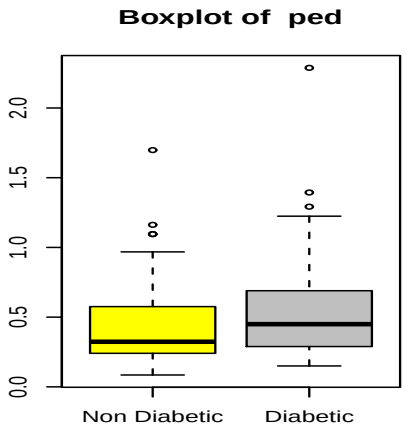












- *Is there any one single variable that alone help discriminate between the diabetic and the non diabetic Pima Indian Women?*

Classification Trees in R

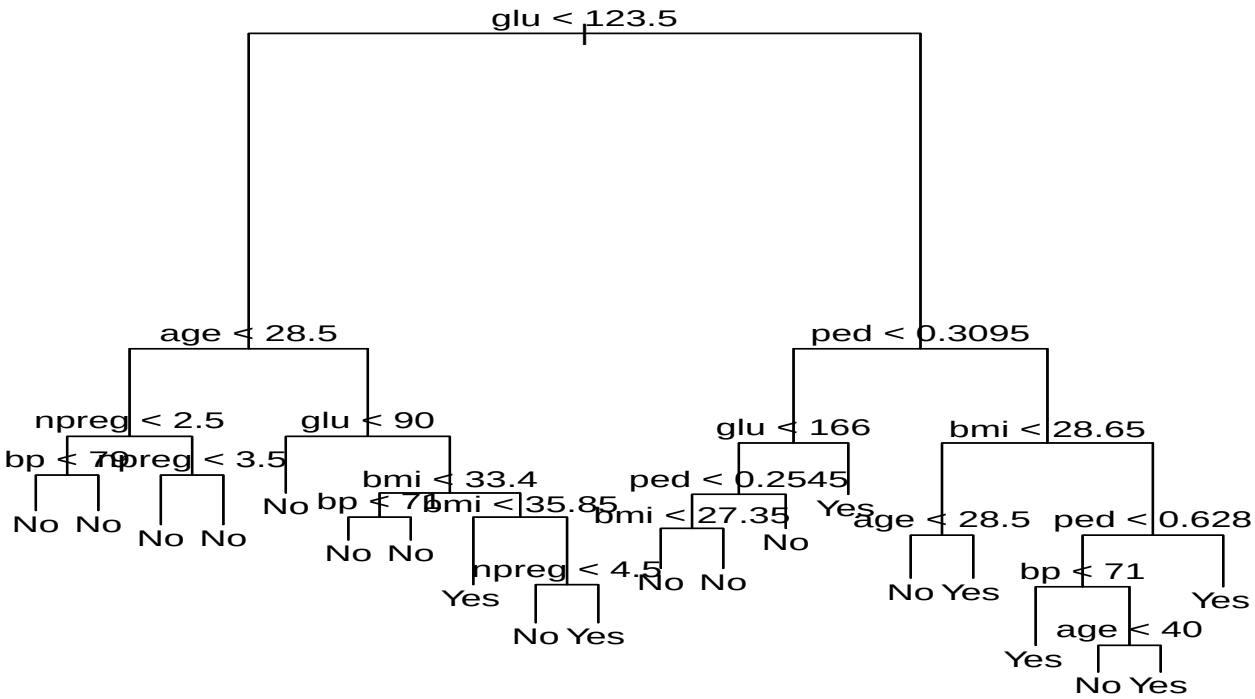
The R command for trees is pretty straightforward and is found in the R package tree.

```
library(tree)
tree(formula, data, ...)
```

Classification trees on the famous Pima Indian Diabetes

```
library(tree)
library(MASS)
tree.pima <- tree(type~., data=Pima.tr)
summary(tree.pima)
plot(tree.pima)
text(tree.pima)
```

Classification tree on Pima Indian Data



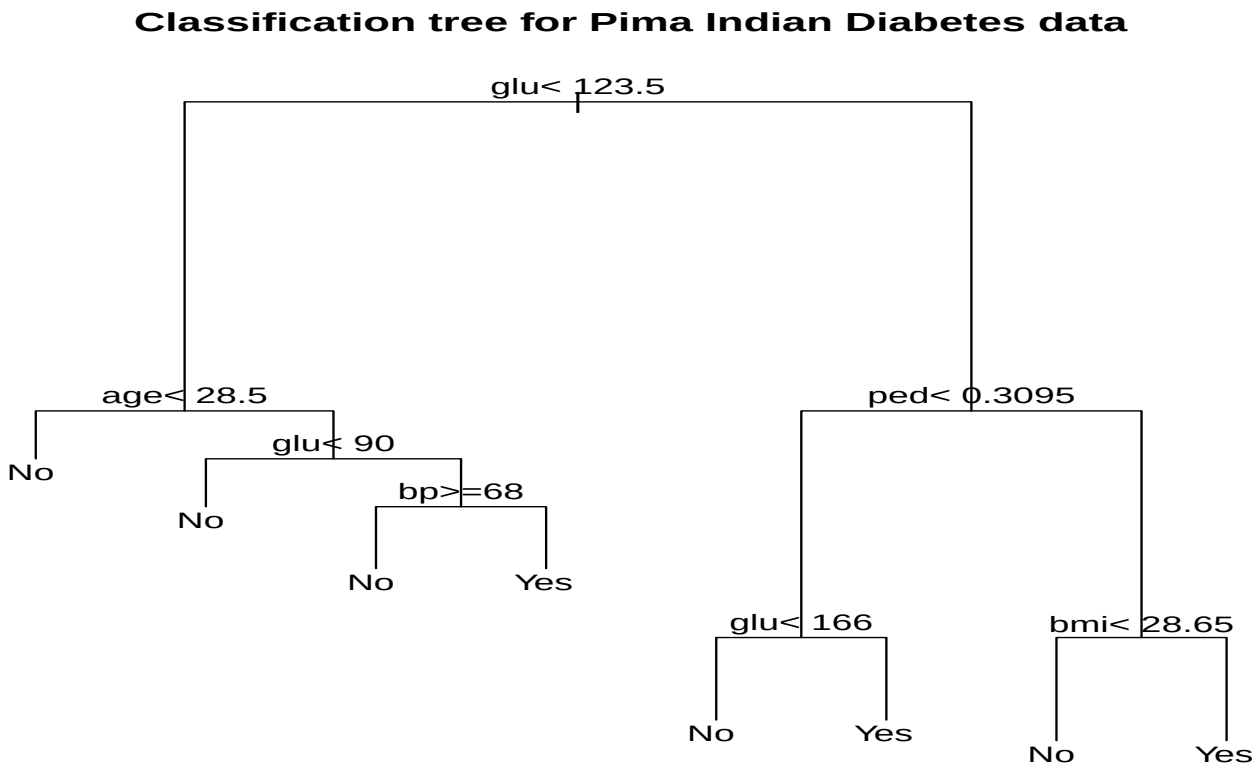
- *The variable glu does indeed come first at the root of the tree*
- *This tree is too complex as it stands, will need to be pruned*

Another R command for trees is pretty straightforward and is found in the R package rpart.

```
library(rpart)
rpart(formula, data, ...)
```

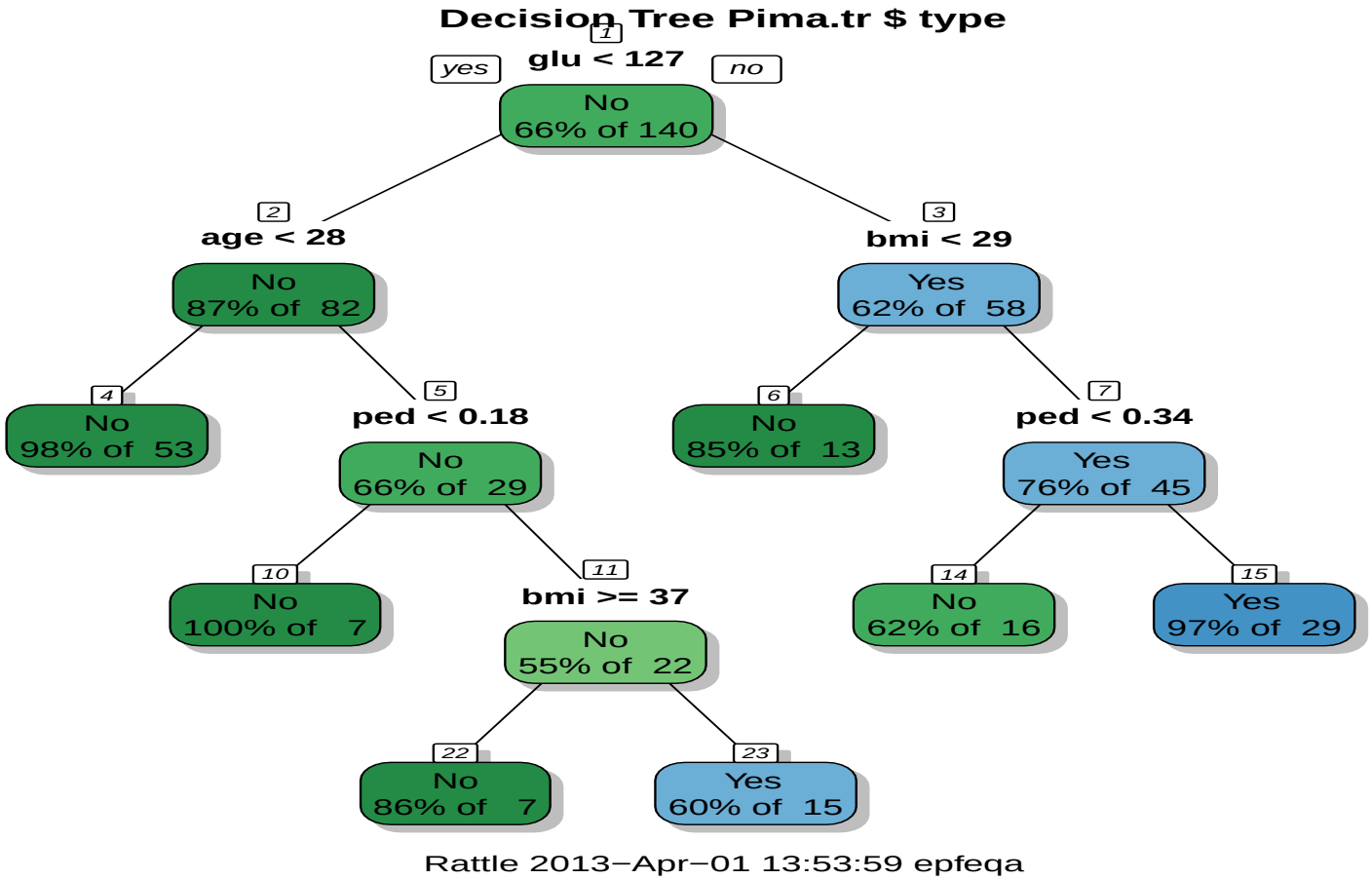
Classification trees on the famous Pima Indian Diabetes

```
library(rpart)
library(MASS)
rpart.pima <- rpart(type~., data=Pima.tr)
summary(rpart.pima)
plot(rpart.pima)
text(rpart.pima)
```



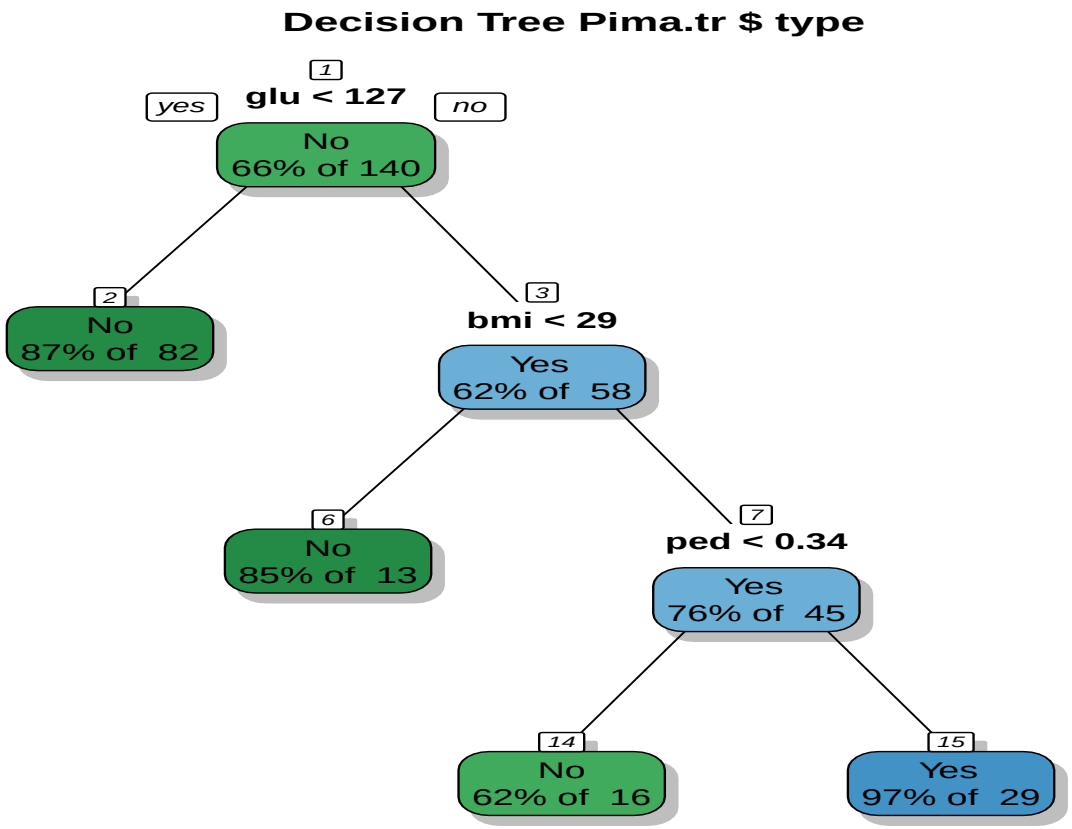
- *The variable glu does indeed come first at the root of the tree*
- *This tree is now simpler and more readable (interpretable).*

Classification Trees using Rattle



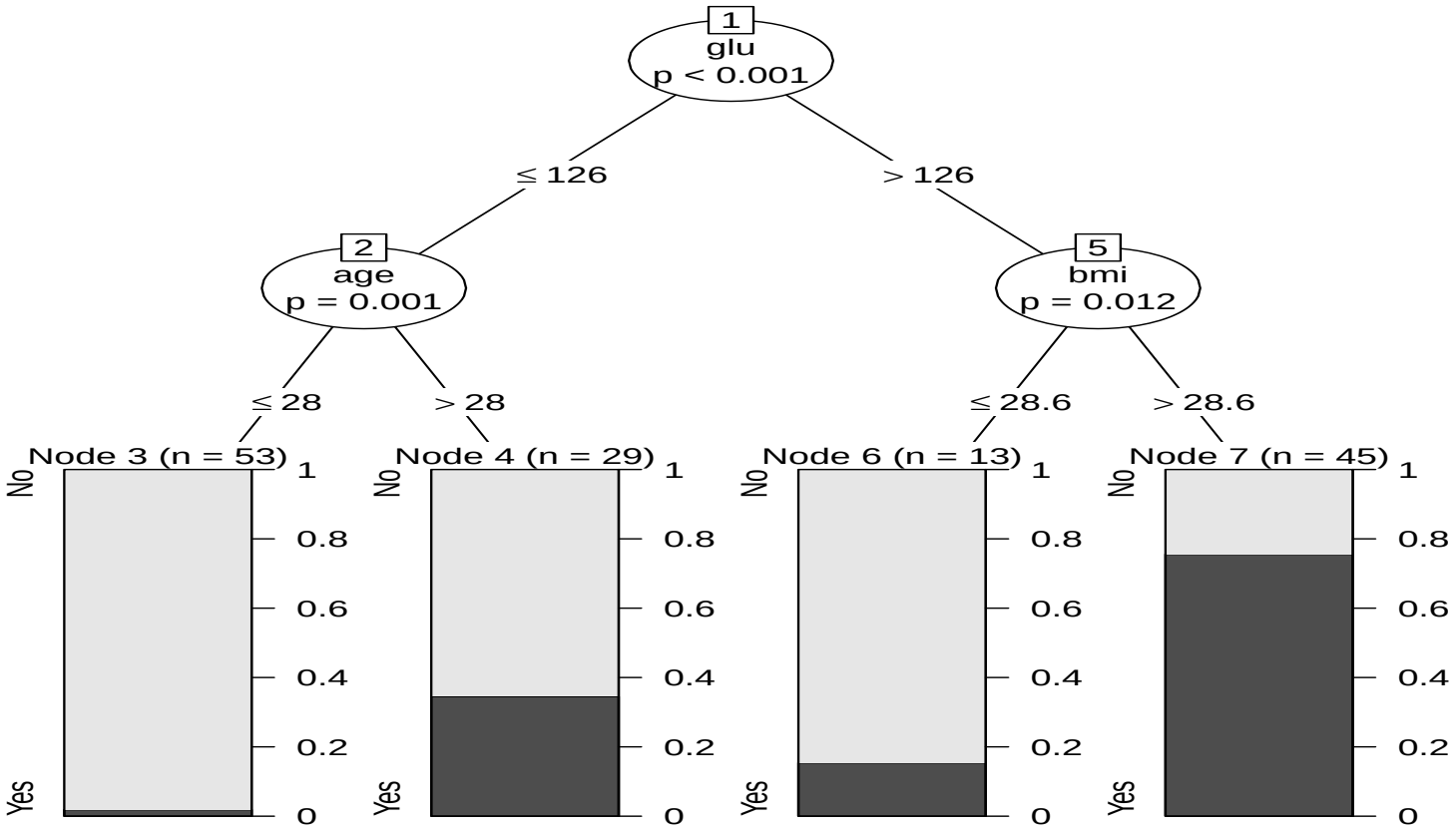
- *A colorful tree indeed, but a little too large (too many terminal nodes). Some pruning would help.*

Classification Trees using Rattle



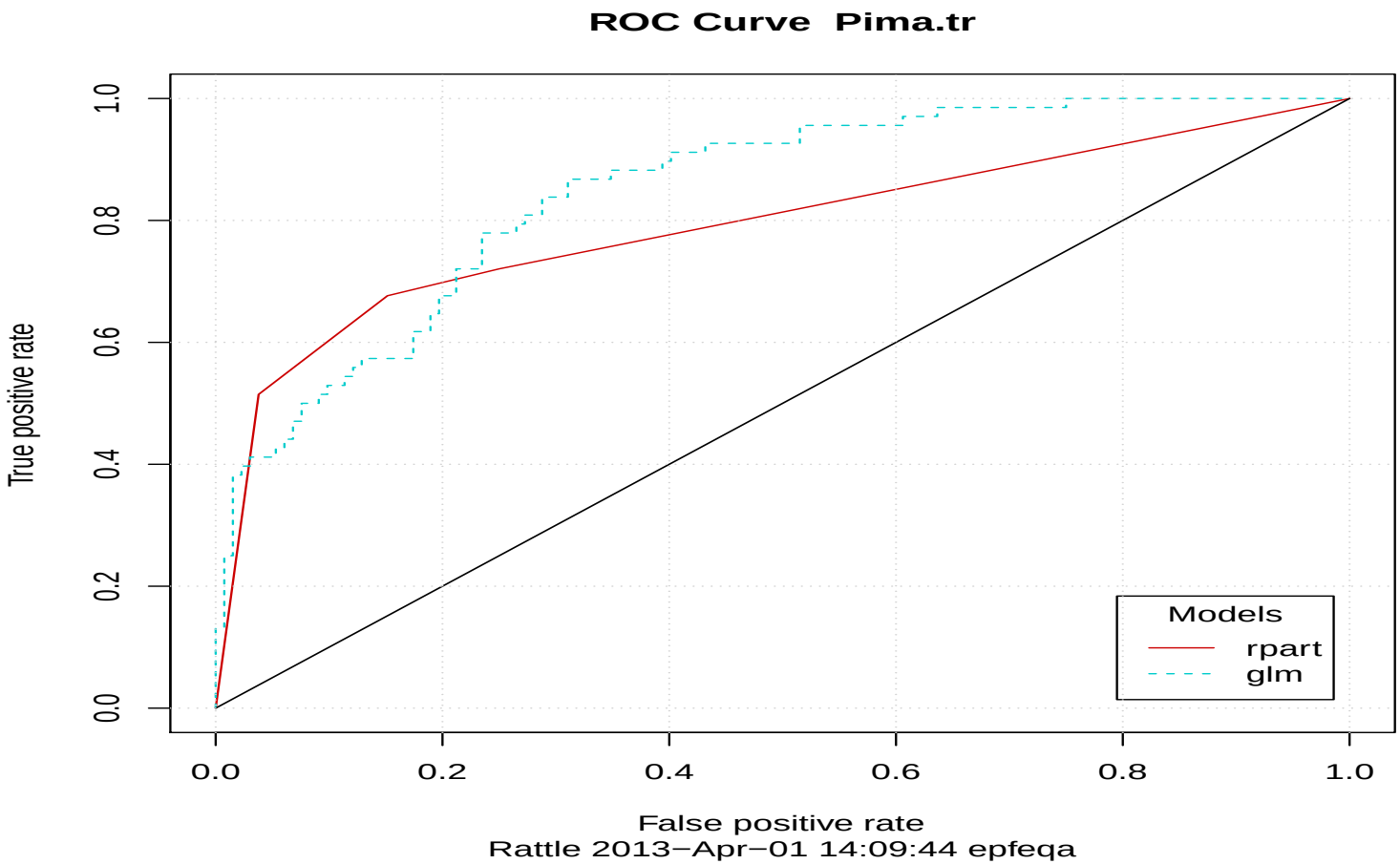
- *A colorful tree indeed, now less bushy!*

Classification Trees using Rattle



- *Visual compelling representation of terminal nodes.*

ROC of Trees vs Logistic in Rattle



- *No obvious winner.*

Exercise session

1. *Consider the Crabs Leptograpsus dataset and build a classification tree for it*
   - *Comment on the complexity of the tree you obtain*
   - *Remove 10 observations at random from the dataset, and rebuild the tree. What do you see?*
2. *Consider the German credit dataset and build a tree for it, then compare its predictive performance to that of logistic regression and LDA.*
3. *Build a classification tree for the Wisconsin breast cancer dataset and compare its performance to that of QDA*
4. *Provide your own dataset and perform a tree building on it*
5. *Explain in your own words what you understand pruning a tree to be all about. What statistical properties does pruning affect?*