THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Decision Tress

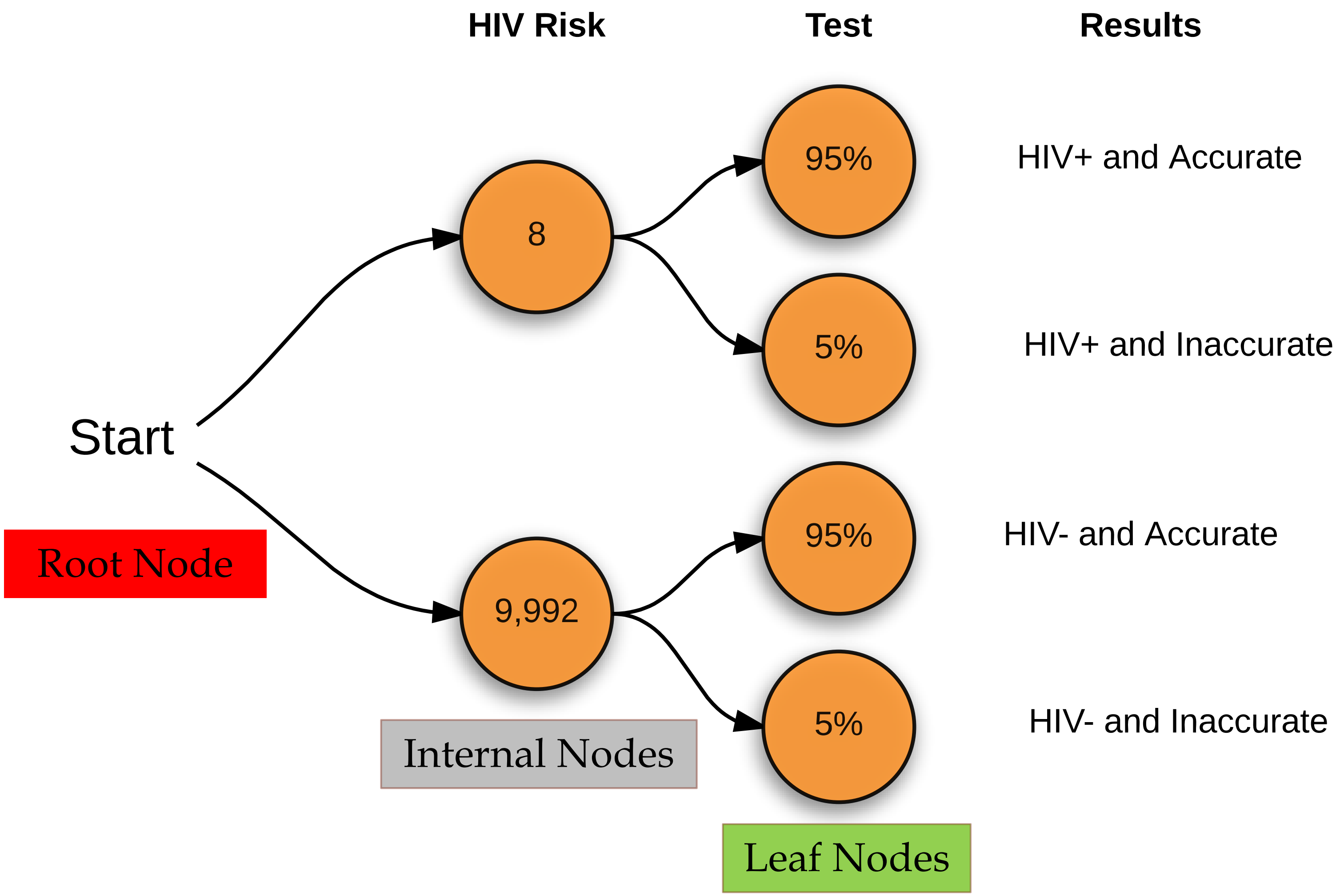
11/19/2019

Trees

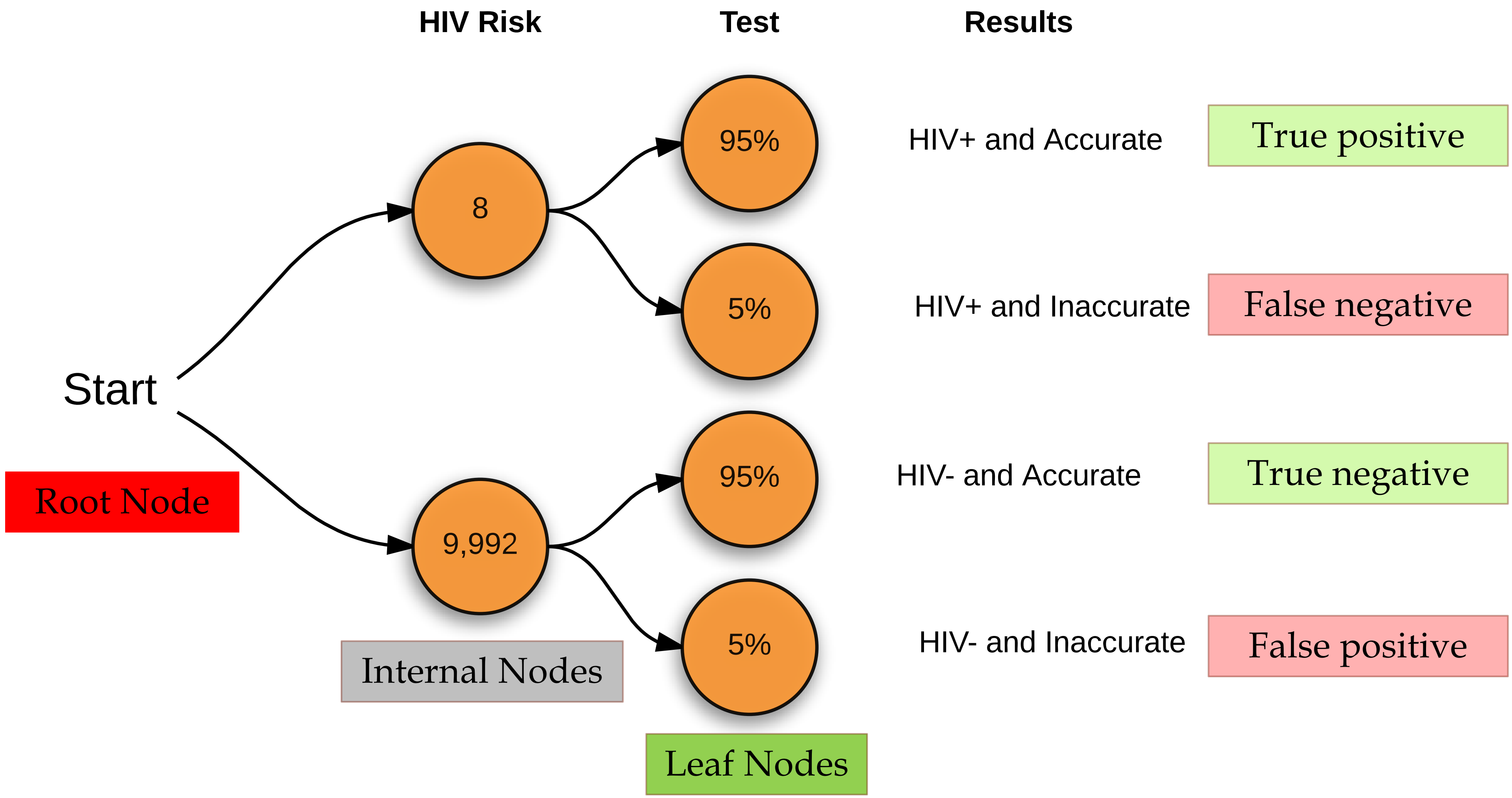
Machine Learning Algorithms *(sample)*

	Unsupervised	Supervised
Continuous	• Clustering & Dimensionality Reduction ○ SVD ○ PCA ○ K-means	• Regression ○ Linear ○ Polynomial • Regression Trees • Random Forests
Categorical	• Association Analysis ○ Apriori ○ FP-Growth • Hidden Markov Model	• Classification ○ KNN ○ Classification Trees ○ Logistic Regression ○ Naive-Bayes ○ SVM

Decision Trees in Statistics



Decision Trees in Statistics



Trees

- Decisions are often made by hierarchical rules.
  - ✓ Is your team going to playoff?
- Which variable is most important?
  - ✓ Decision trees can give good hints
- Categorical dependent variable: Classification tree
- Quantitative dependent variable: Regression tree
- Predictors: can mix categorical & quantitative
- Trees can be tuned and pruned.
- ***Random Forests*** – with many many trees, averaged, tuned, randomized (cross-validation etc)

Trees – pros and cons

- ✓ Easy to understand and interpret
- ✓ Insensitive to outliers
- ✓ Can handle both linear and non-linear relations
- ✓ Missing data handled elegantly
- ✓ Require little data pre-processing
- ✓ Large datasets work just fine.

- ✗ Large trees – difficult to interpret
- ✗ High variance, low performance
- ✗ Overfitting easily

Trees – Classification or Regression

- Both types:
  - use divide & conquer strategy
  - make usually binary splits
  - goal is to maximize increase in homogeneity after each split
  - can couple with other techniques such as CV, bagging, boosting, etc

Trees – two types

- Classification Trees – categorical outcome
  - The outcome (leaf node) predicts the label
  - Each split is decided by maximizing the information gained on resulting configuration
  - Info gain = Entropy(parent) – Entropy(children)
  - Common model evaluation by accuracy, precision, recall, F1 score, AUC/ROC

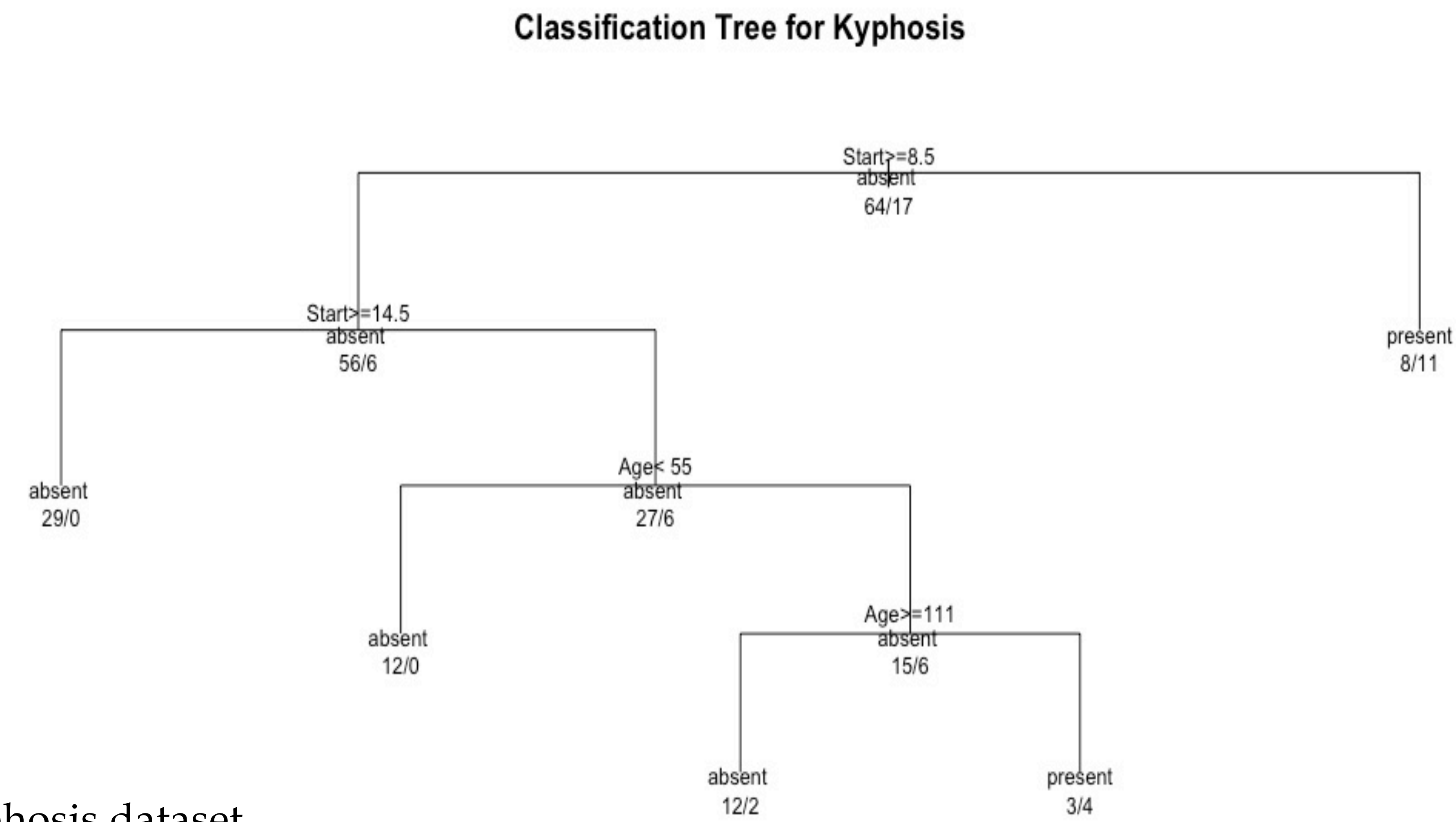
Trees – two types

- Regression Trees – numerical outcome
  - The outcome predicts the average value of the target variable, somewhat like K-means
  - Instead of entropy, each split is decided by variance, std deviation, absolute deviation and the likes.
  - Common model evaluation/metric by root-mean-squared-error (RMSE) or mean-absolute-error (MAE)

Classification Tree Example

- Predict kyphosis occurrences in patients
  - Example from www.statmethods.net/advstats/cart.html
- Use **binary recursive splitting**
- The predictor solution space (with both x variables being numerical) is divided into regions
- Branches are created by **cutpoints** that divide the space into regions
- Cutpoints are determined by maximizing information gain, or minimal entropy.
- Cuts continue recursively

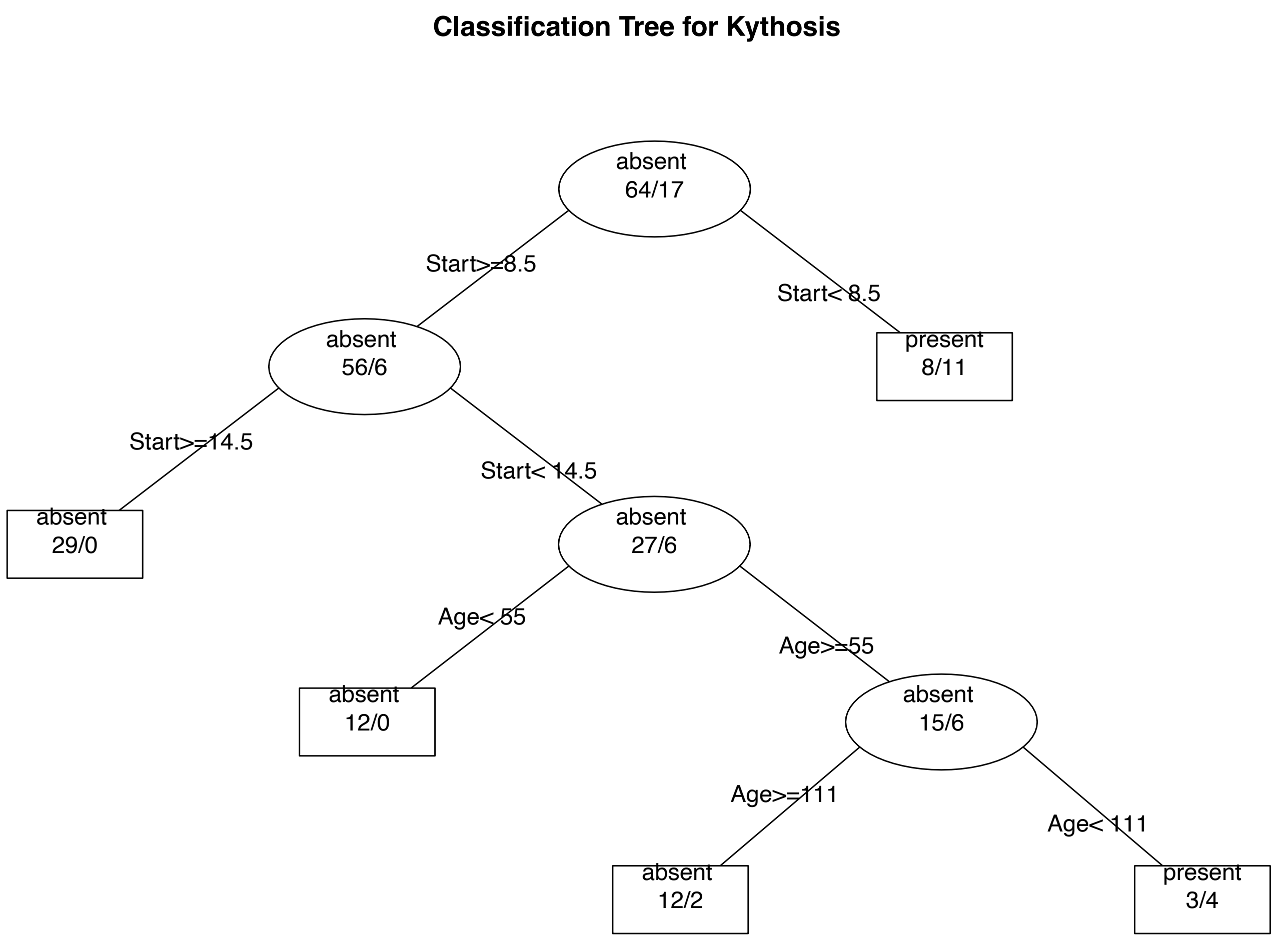
Classification Tree Example



Kyphosis dataset
Source: https://www.statmethods.net/advstats/cart.html

Classification Tree Example

Or a nicer postscript plot using post( ) function

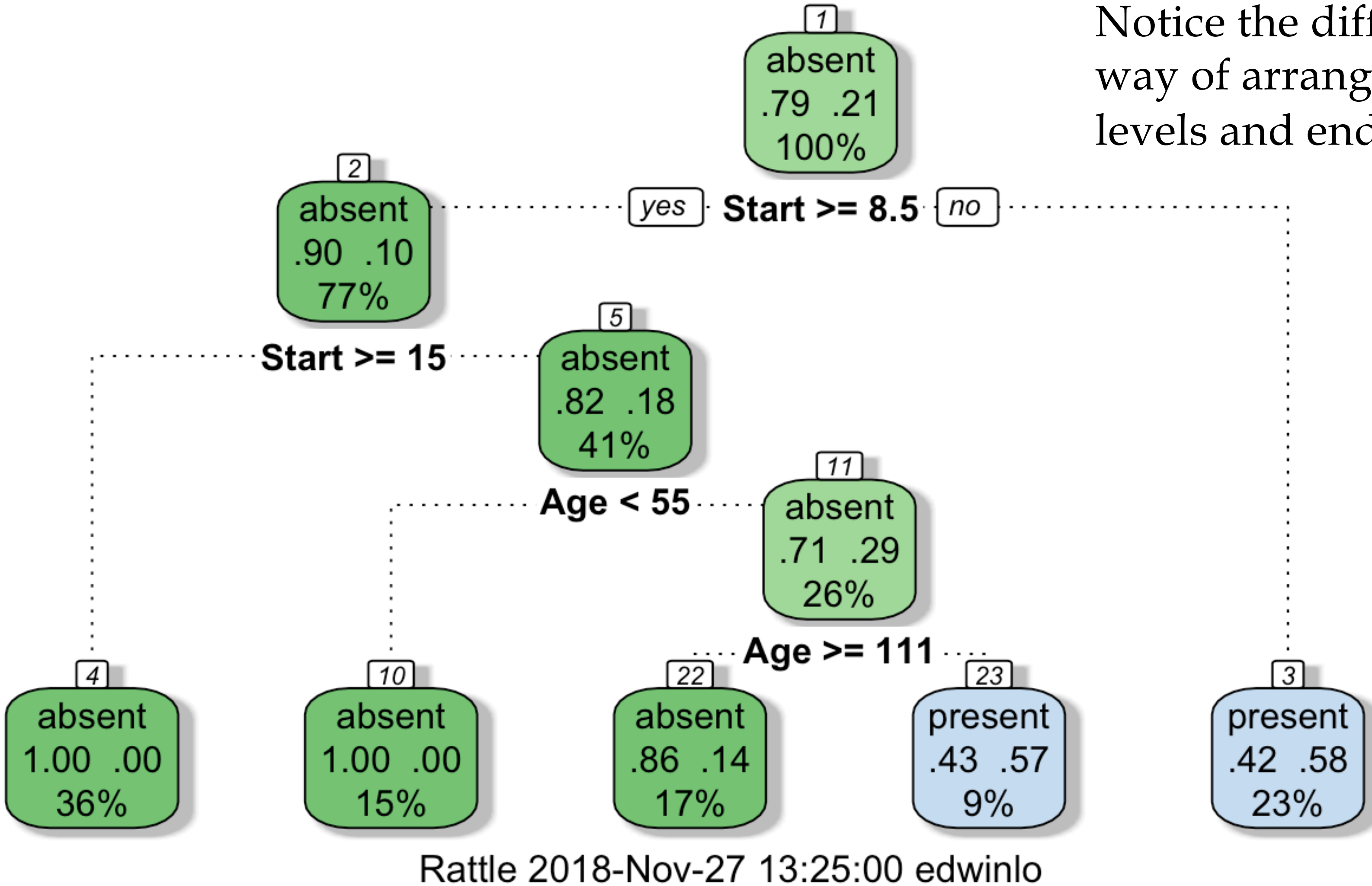


THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Classification Tree Example

Kyphosis dataset with fancyRpartPlot( ) (library: rattle)

Notice the different way of arrange the levels and end notes.

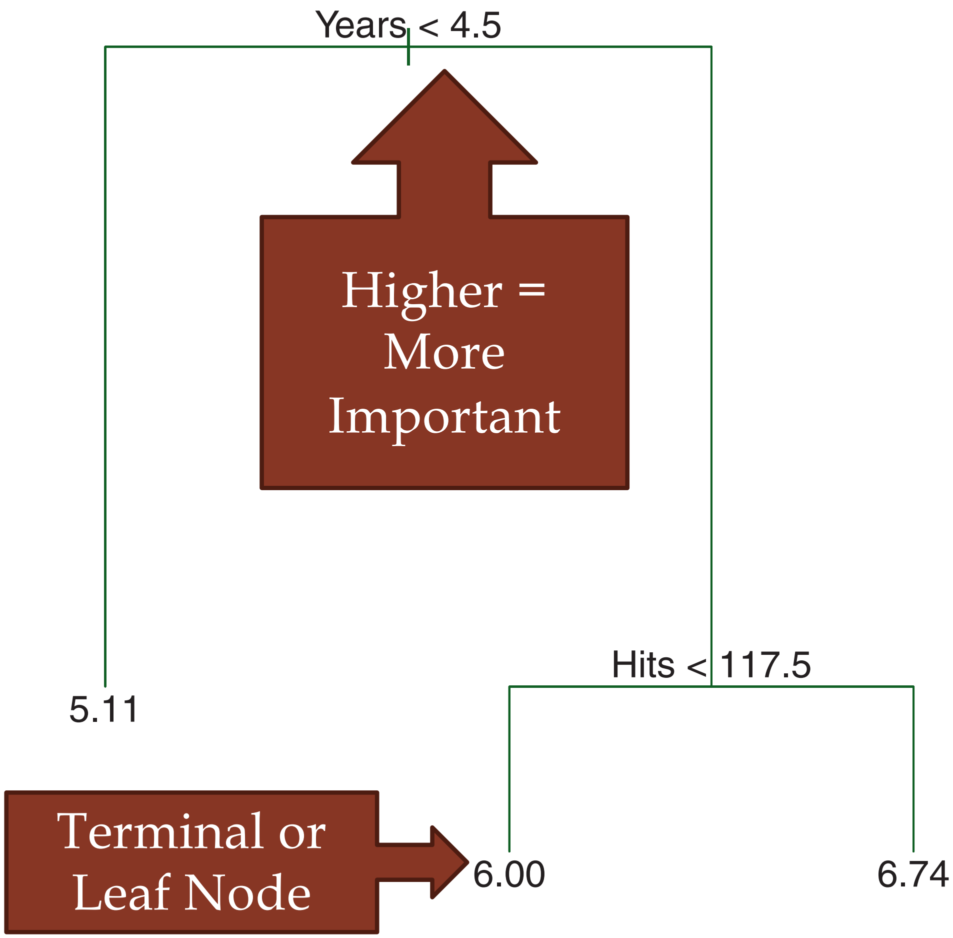


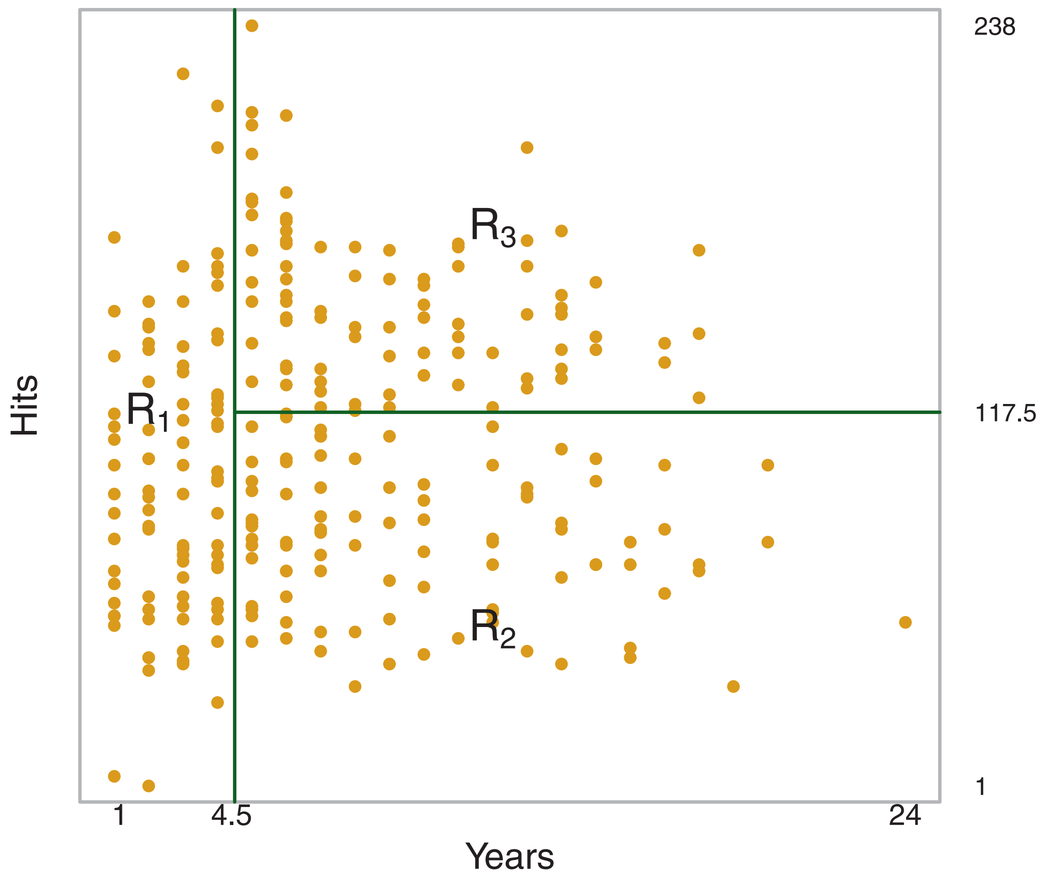
Regression Tree Example

- Predict baseball player salary
  - Example by Rebecca C. Steorts, Duke University
- Use **binary recursive splitting**
- The predictor solution space is divided into regions
- Branches are created by **cutpoints** that divide the space into regions
- Each region is defined by its average values of the variables (indept and/or dept), c.f. K-means
- Cutpoints are determined by minimizing MSE of both new regions.
- Cuts continue recursively

Regression Tree Example

Recursive Binary Splitting





Regression Tree Example

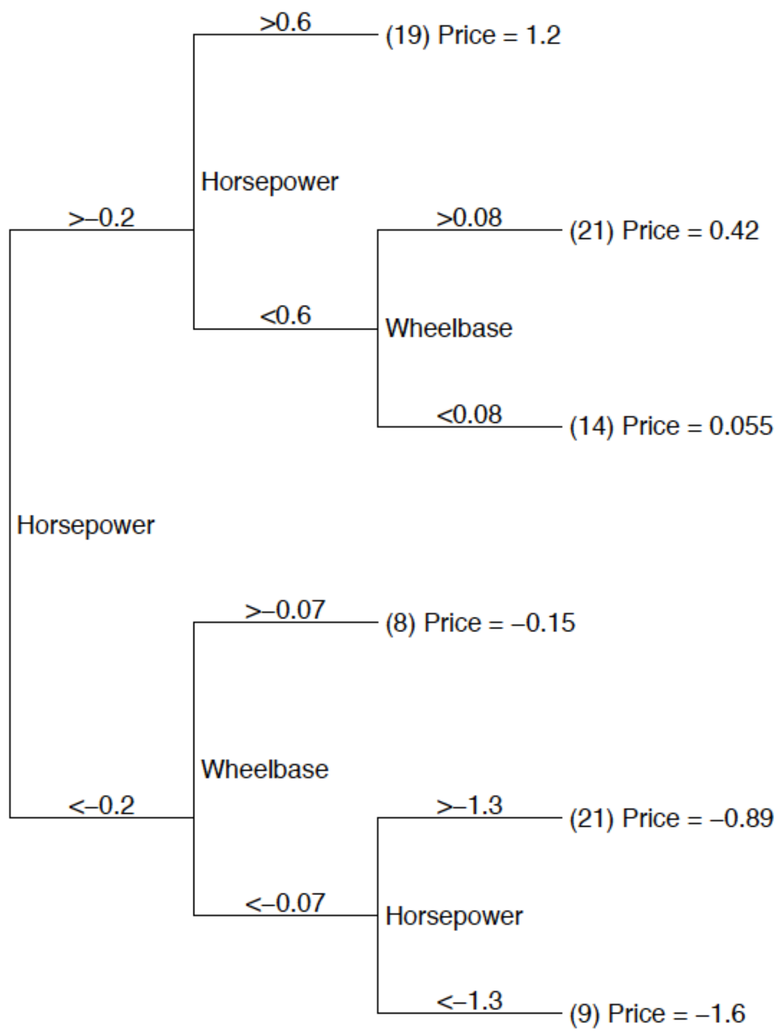


Figure 1: Regression tree for predicting price of 1993-model cars. All features have been standardized to have zero mean and unit variance. Note that the order in which variables are examined depends on the answers to previous questions. The numbers in parentheses at the leaves indicate how many cases (data points) belong to each leaf.

Predicting price of 1993-model cars.

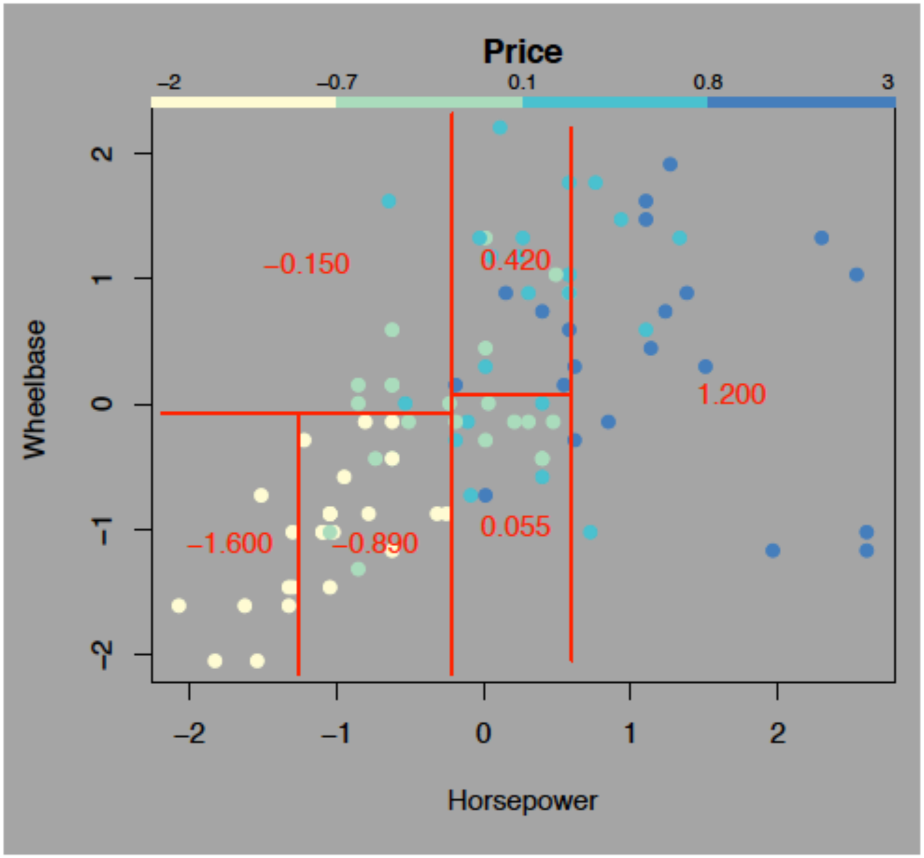
All variables standardized.

Note the order of variables examined depends previous questions.

The numbers in parentheses at the leaves indicate how many cases (data points) belong to each leaf.

Source: CMU Data Mining lectures http://www.stat.cmu.edu/~cshalizi/350-2006/lecture-10.pdf

Regression Tree Example



Predicting price of 1993-model cars.

Partitions of solution space

Quantitative response variable represented by color.

Source: CMU Data Mining lectures http://www.stat.cmu.edu/~cshalizi/350-2006/lecture-10.pdf

Figure 2: The partition of the data implied by the regression tree from Figure 1. Notice that all the dividing lines are parallel to the axes, because each internal node checks whether a single variable is above or below a given value.

Random Forest

- Forest contains many many trees
- Each tree from different sample
- Each split chosen from random sample of possible predictors
- Usually $m = \sqrt{p}$ possible predictors are chosen at each split
- Useful when many independent variables or one strong and many weak predictors
- If predicting within the space of the sample, decision tree might show better accuracy
- When predicting unexpected (untrained) data, random forest usually much better