Exploring Decision Trees with R

J. Alfredo Freites

OCRUG, February 2018

Sources

rpart *vignette* http://127.0.0.1:30960/help/library/rpart/doc/longintro.pdf

rpart.plot *vignette* http://127.0.0.1:30960/help/library/rpart.plot/doc/prp.pdf

Max Kuhn and Kjell Johnson *Applied Predictive Modeling,* Springer, 2013. ISBN 978-1-4614-6848-6

Tom Mitchell Machine Learning, McGraw-Hill, 1997. ISBN 0070428077

Decision Tree modeling

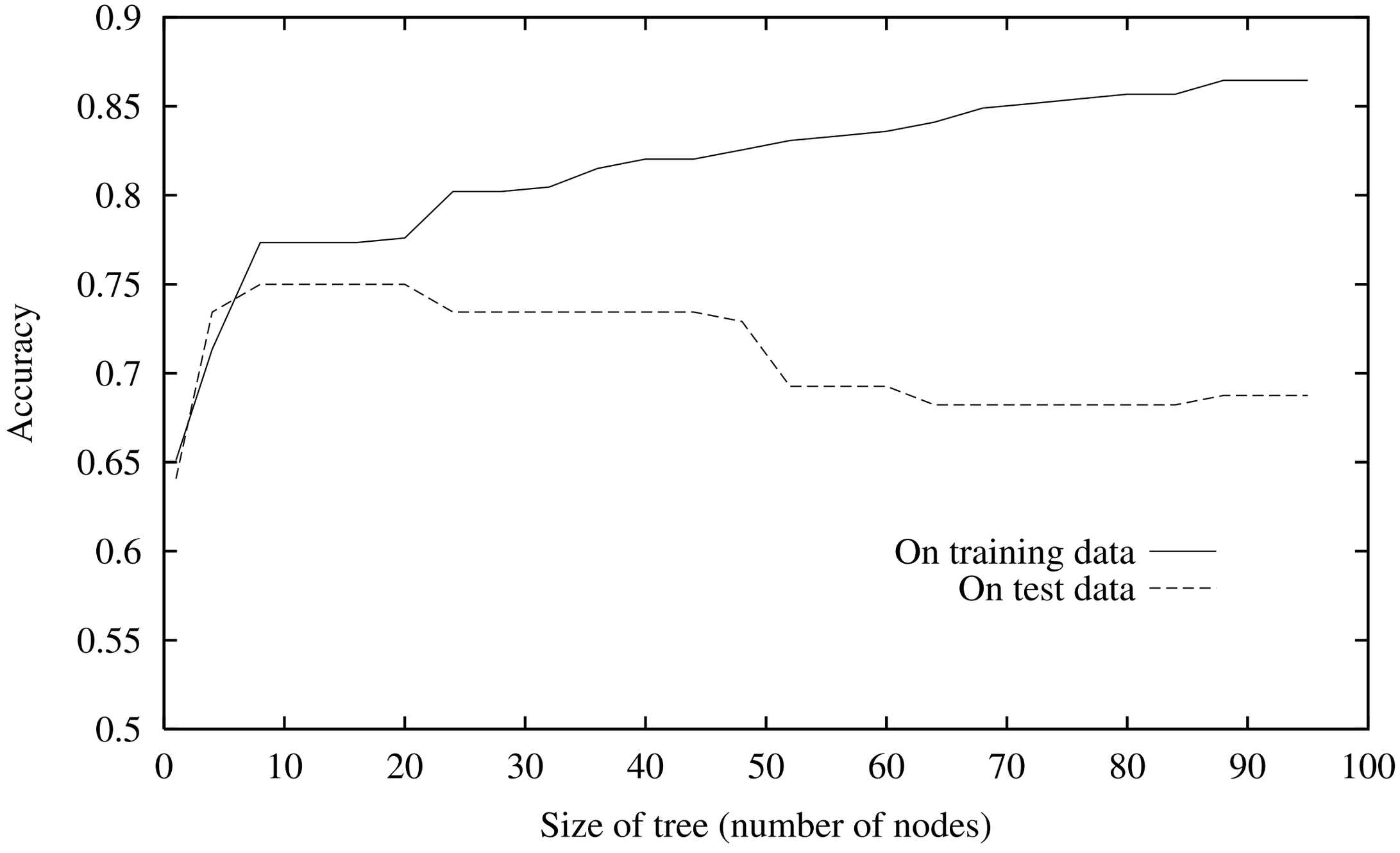
Partition the data into smaller groups that are more homogeneous with respect to the response

- real-valued response: Regression Trees
- categorical (discrete): Classification Trees

Known knowns (the bad)

- Overfitting
- Model instability
- If-then statements produce rectangular partitions, which may negatively affect predictive performance

Overfitting



Machine Learning, ©Tom M. Mitchell, McGraw Hill, 1997

Decision Tree modeling

Known knowns (the good)

White box. Highly interpretable. I know how it was done.

Can handle all kinds of variables.

Can handle missing data.

Base for successful ensemble methods.

rpart (recursive partitioning)

CART (Classification and Regression Trees)

Breiman, Friedman, Olshen and Stone book.

can do

- ***Categorical***
- ***Linear regression***
- ***Poisson***
- ***Exponential***

Two stages

First:

Find the single variable which BEST splits the data into two groups

Apply separately two each sub-group RECURSIVELY until minimum size or no improvement

Second:

Use cross-validation to prune the resulting tree

Key points

Splitting criteria: Gini, entropy.

Pruning method: Cross validation

Missing data: if there's at least one independent variable, it will be used!

Maximize this thing

I: Impurity measure

p: probability

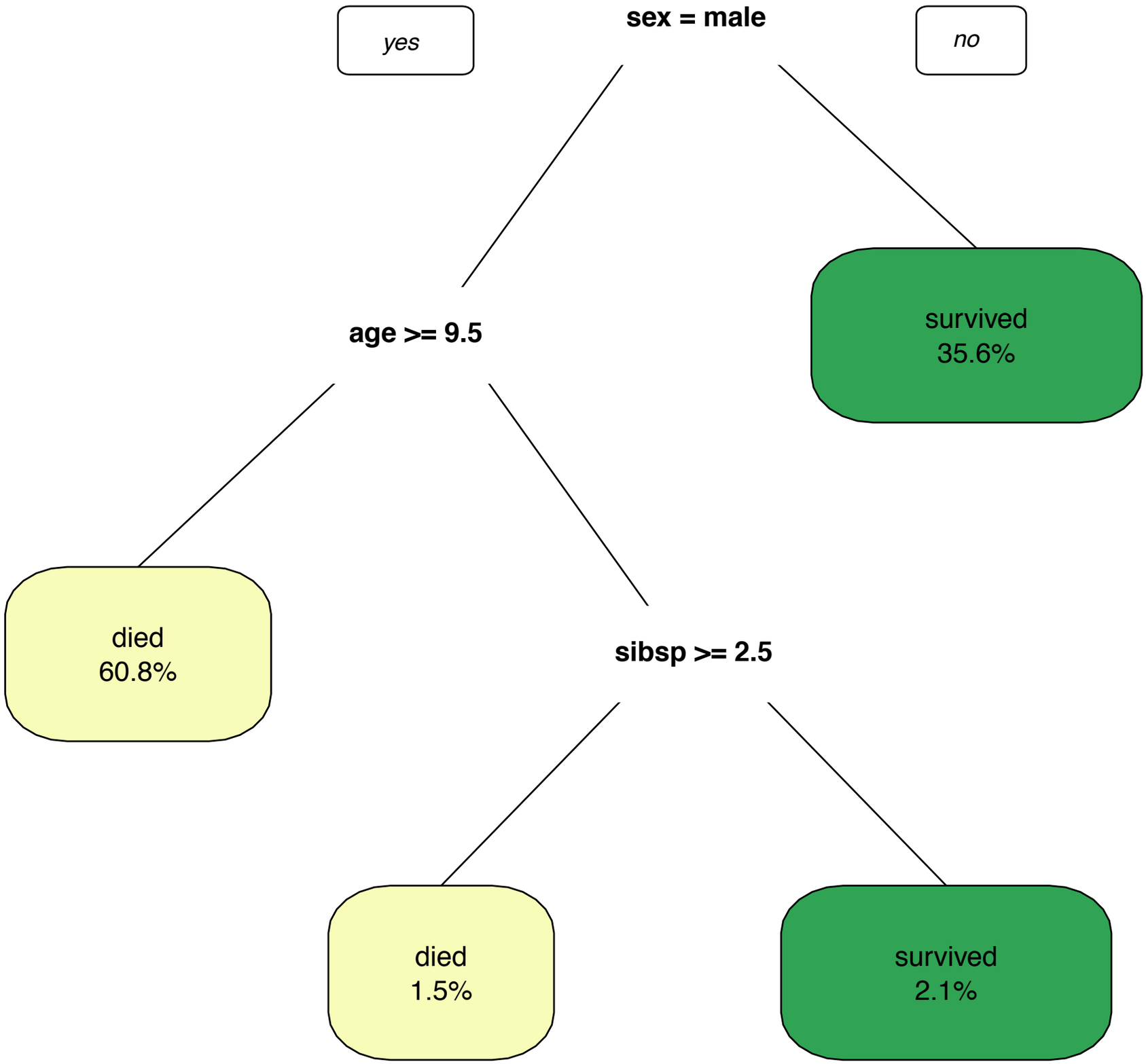
A: the parent node (a rule and a set of data)

A_L, A_R: the left and right children

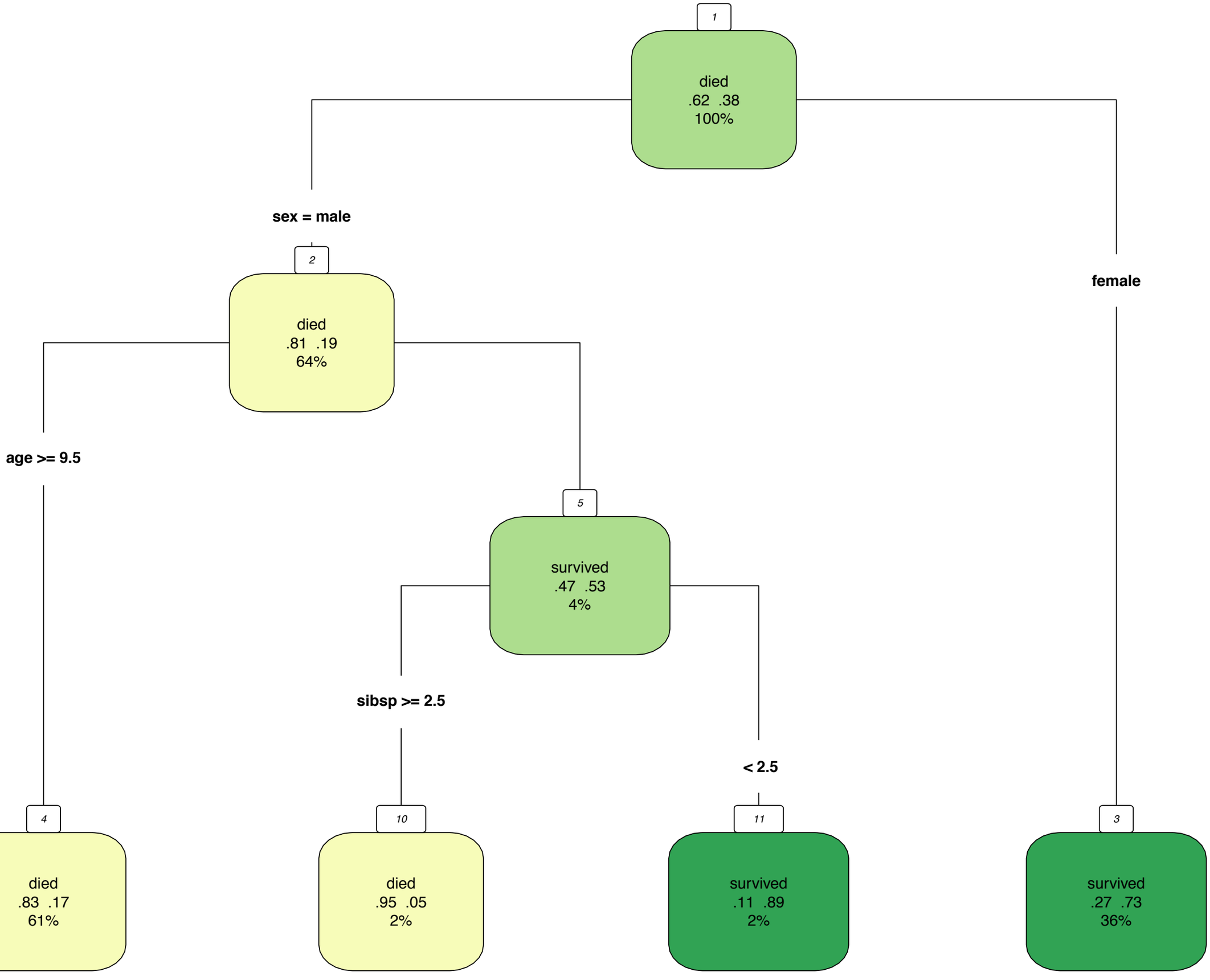
$$\Delta I = p(A)I(A) - p(A_L)I(A_L) - p(A_R)I(A_R)$$

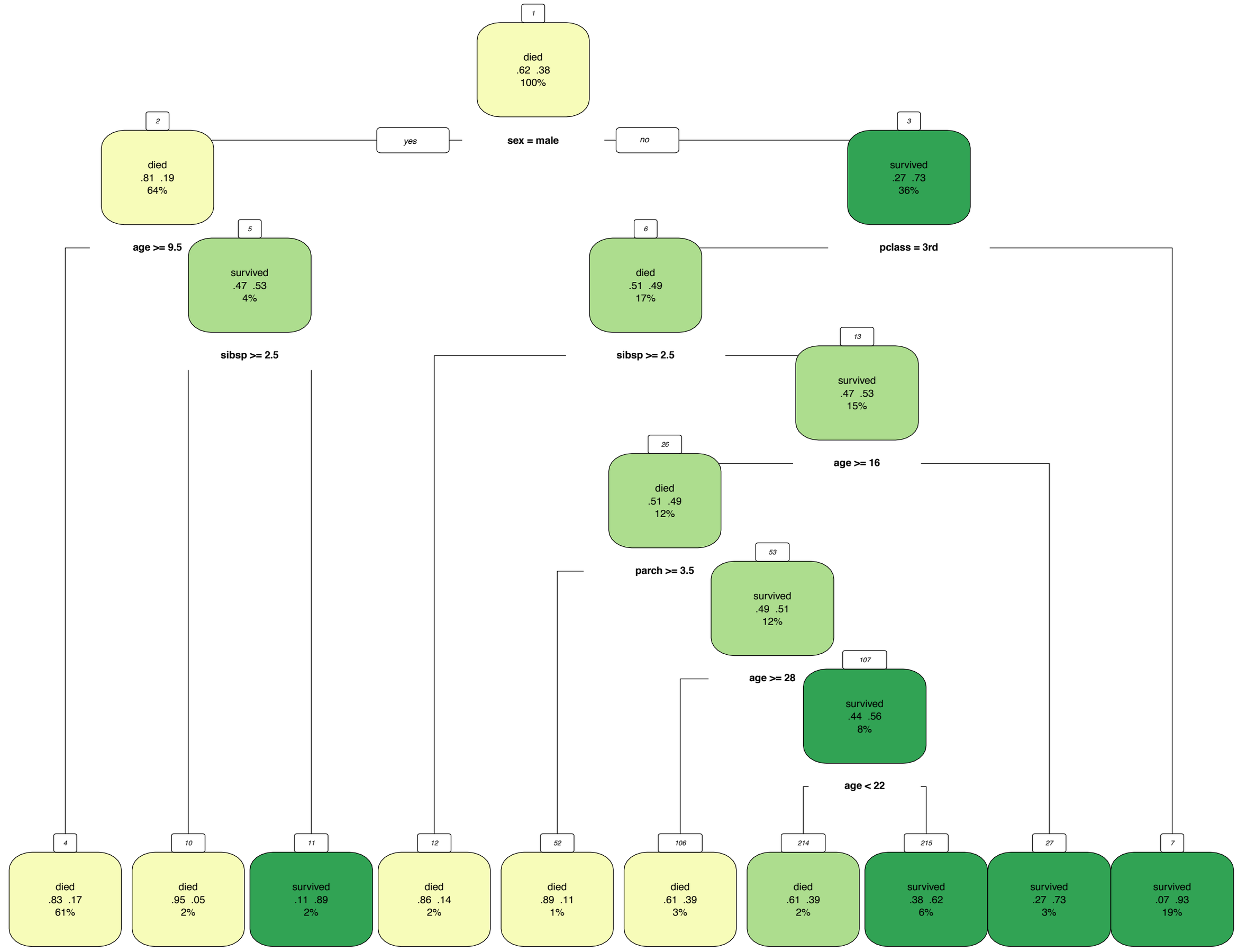

LEONARDO
DICAPRIO
KATE
WINSLET
WRITTEN AND DIRECTED BY
JAMES CAMERON
TITANIC
DIGITALLY REMASTERED IN HIGH DEFINITION

Who survived?



Who Survived?




1
died
.62 .38
100%
yes
sex = male
no
2
died
.81 .19
64%
3
survived
.27 .73
36%
age >= 9.5
5
survived
.47 .53
4%
pclass = 3rd
6
died
.51 .49
17%
sibsp >= 2.5
sibsp >= 2.5
13
survived
.47 .53
15%
26
died
.51 .49
12%
age >= 16
parch >= 3.5
53
survived
.49 .51
12%
age >= 28
107
survived
.44 .56
8%
age < 22
4
died
.83 .17
61%
10
died
.95 .05
2%
11
survived
.11 .89
2%
12
died
.86 .14
2%
52
died
.89 .11
1%
106
died
.61 .39
3%
214
died
.61 .39
2%
215
survived
.38 .62
6%
27
survived
.27 .73
3%
7
survived
.07 .93
19%

Cost of splitting

cp: complexity parameter

R is the risk

T is the number of terminal nodes

T_1 is the one-node tree

$$R_{cp}(T) \equiv R(T) + cp * |T| * R(T_1)$$