CART con R

Rodrigo Assar

Paquete de R

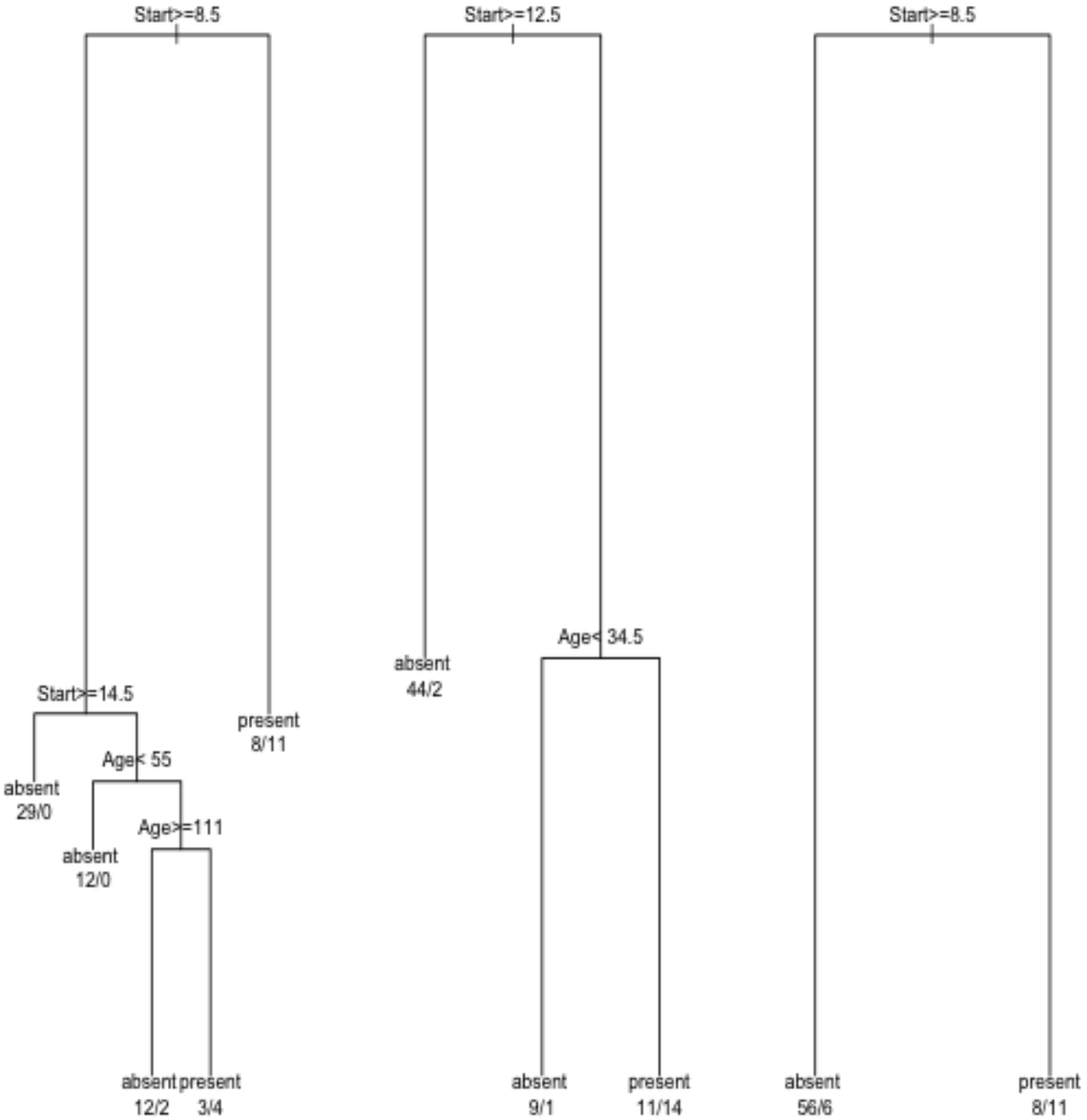
- rpart, en repositorio CRAN.

- Funcion rpart():
  - Componentes:
    - formula, method, cost
  - En parms:
    - prior=c(0.65,0.35) #ejemplo
    - Split: ‘gini’ o ‘information’

Ejemplo

```
fit <- rpart(Kyphosis ~ Age + Number + Start, data=kyphosis)
fit2 <- rpart(Kyphosis ~ Age + Number + Start, data=kyphosis,
              parms=list(prior=c(.65,.35), split='information'))
fit3 <- rpart(Kyphosis ~ Age + Number + Start, data=kyphosis,
              control=rpart.control(cp=.05))
par(mfrow=c(1,3), xpd=NA) # para juntar graficos
plot(fit)
text(fit, use.n=TRUE)
plot(fit2)
text(fit2, use.n=TRUE)
plot(fit3)
text(fit3, use.n=TRUE)
```

Resultado (visualizacion)



Viendo descripcion: print(fit)

```
n= 81

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 81 17 absent (0.79012346 0.20987654)
   2) Start>=8.5 62  6 absent (0.90322581 0.09677419)
     4) Start>=14.5 29  0 absent (1.00000000 0.00000000) *
     5) Start< 14.5 33  6 absent (0.81818182 0.18181818)
      10) Age< 55 12  0 absent (1.00000000 0.00000000) *
      11) Age>=55 21  6 absent (0.71428571 0.28571429)
        22) Age>=111 14  2 absent (0.85714286 0.14285714) *
        23) Age< 111 7  3 present (0.42857143 0.57142857) *
   3) Start< 8.5 19  8 present (0.42105263 0.57894737) *
```

Viendo descripcion: print(fit2)

n= 81

node), split, n, loss, yval, (yprob)

* denotes terminal node

1) root 81 28.350000 absent (0.65000000 0.35000000)

2) Start>=12.5 46 3.335294 absent (0.91563089 0.08436911) *

3) Start< 12.5 35 16.453120 present (0.39676840 0.60323160)

6) Age< 34.5 10 1.667647 absent (0.81616742 0.18383258) *

7) Age>=34.5 25 9.049219 present (0.27932897 0.72067103) *

Viendo descripcion: print(fit3)

n= 81

node), split, n, loss, yval, (yprob)

* denotes terminal node

1) root 81 17 absent (0.79012346 0.20987654)

2) Start>=8.5 62 6 absent (0.90322581 0.09677419) *

3) Start< 8.5 19 8 present (0.42105263 0.57894737) *

Con funcion summary()

```
summary(fit2)
Call:
rpart(formula = Kyphosis ~ Age + Number + Start, data = kyphosis,
    parms = list(prior = c(0.65, 0.35), split = "information"))
  n= 81

          CP nsplit rel error    xerror      xstd
1 0.3019958      0 1.0000000 1.0000000 0.2155872
2 0.2023372      1 0.6980042 0.9960609 0.1941501
3 0.0100000      2 0.4956670 0.7293855 0.1528517
```

Con funcion summary() (cont.)

Node number 1: 81 observations, complexity param=0.3019958

predicted class=absent expected loss=0.35

class counts: 64 17

probabilities: 0.650 0.350

left son=2 (46 obs) right son=3 (35 obs)

Primary splits:

Start < 12.5 to the right, improve=13.152920, (0 missing)

Age < 39.5 to the left, improve= 6.035255, (0 missing)

Number < 4.5 to the left, improve= 5.602041, (0 missing)

Surrogate splits:

Number < 3.5 to the left, agree=0.667, adj=0.229, (0 split)

Con funcion summary() (cont.)

Node number 2: 46 observations

predicted class=absent expected loss=0.07250639

class counts: 44 2

probabilities: 0.916 0.084

Node number 3: 35 observations, complexity param=0.2023372

predicted class=present expected loss=0.4700893

class counts: 20 15

probabilities: 0.397 0.603

left son=6 (10 obs) right son=7 (25 obs)

Primary splits:

Age < 34.5 to the left, improve=4.335641, (0 missing)

Start < 8.5 to the right, improve=2.310569, (0 missing)

Number < 4.5 to the left, improve=2.028844, (0 missing)

Con funcion summary() (cont.)

Node number 6: 10 observations
  predicted class=absent   expected loss=0.1667647
    class counts: 9   1
   probabilities: 0.816 0.184

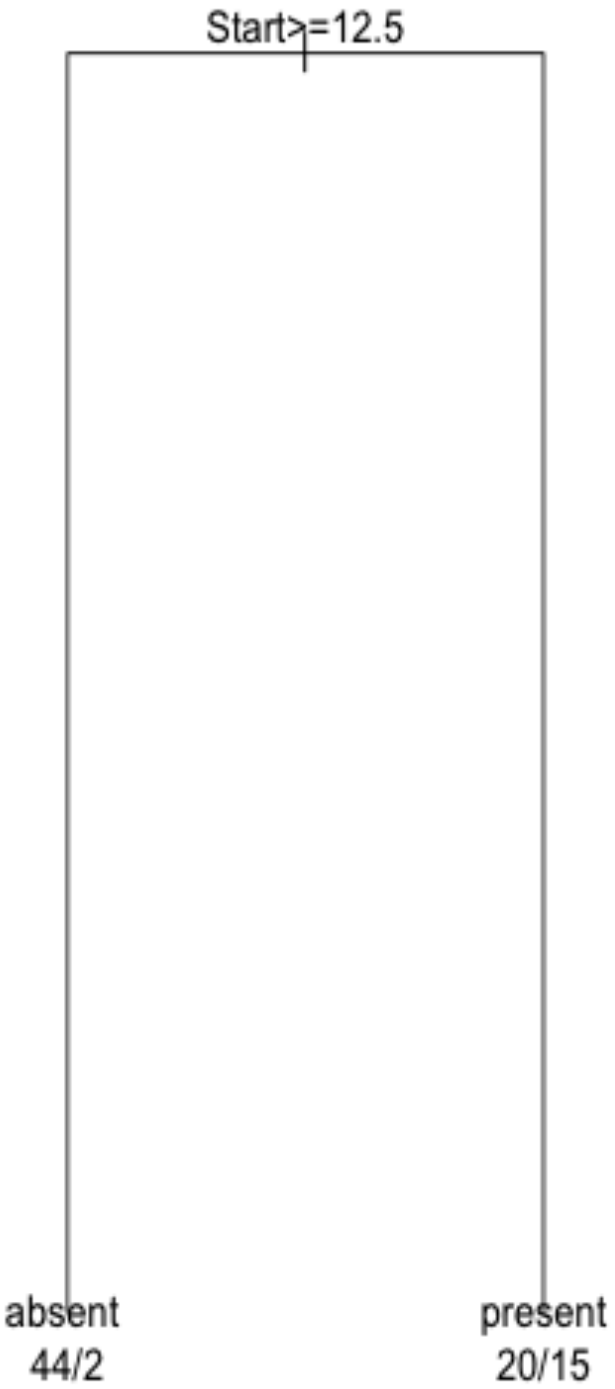
Node number 7: 25 observations
  predicted class=present expected loss=0.3619688
    class counts: 11   14
   probabilities: 0.279 0.721

Podando

- Con funcion prune()
  - A traves del campo cp
  - Ejemplo:

```
fit2 <- rpart(Kyphosis ~ Age + Number + Start, data=kyphosis, parms=list(prior=c(.65,.35), split='information'))
fit2_2 <- prune(fit2, cp=0.25)
par(mfrow=c(1,2), xpd=NA)
plot(fit2)
text(fit2, use.n=TRUE)
plot(fit2_2)
text(fit2_2, use.n=TRUE)
```


Start>=12.5
absent
44/2
Age< 34.5
absent
9/1
present
11/14


Start>=12.5
absent
44/2
present
20/15

Prediciendo

- predict()
  - Ejemplo:

```
sub=sample(1:81, 60)
fit2 <- rpart(Kyphosis ~ Age + Number + Start, data=kyphosis,
      parms=list(prior=c(.65,.35), split='information'),subset=sub)
plot(fit2)
valid=setdiff(1:81,sub)
prediccion=predict(fit2,kyphosis[valid,],type='class)
table(prediccion,kyphosis[valid,1])
```

	absent	present
4	0.2550943	0.7449057
5	1.0000000	0.0000000
7	0.3910227	0.6089773
8	1.0000000	0.0000000
9	1.0000000	0.0000000
12	1.0000000	0.0000000
13	0.2550943	0.7449057
14	1.0000000	0.0000000
19	0.3910227	0.6089773

…

prediccion	absent	present
absent	13	2
present	4	2