Modelo de Árvore de Decisão para o dataset Kyphosis

Silva, Guilherme Aquino

09/11/2021

Dataset Kyphosis

Kyphosis é um dataset que representa os dados sobre crianças que passaram por cirurgia corretiva da coluna vertebral. Este data frame contém 81 linhas e 4 colunas, as seguintes colunas são:
1. **Kyphosis**: Indica com níveis de absent (ausente) e present (presente) um tipo de deformação após a operação
2. **Age**: Indica a idade em formato de meses.
3. **Number**: Indica o número de vértebras envolvidas.
4. **Start**: Indica o número da primeira vértebra operada.

O modelo de árvore de decisão utilizado será o rpart com o objetivo de prever se a pessoa pussui a tal deformação depois da operação.
Documentação disponível em Kyphosis (https://cran.r-project.org/web/packages/rpart/rpart.pdf) (pág. 6).

Pacotes utilizados

```
library(rpart)
library(ggplot2)
library(caTools)
library(rpart.plot)
```

Análisando o dataset

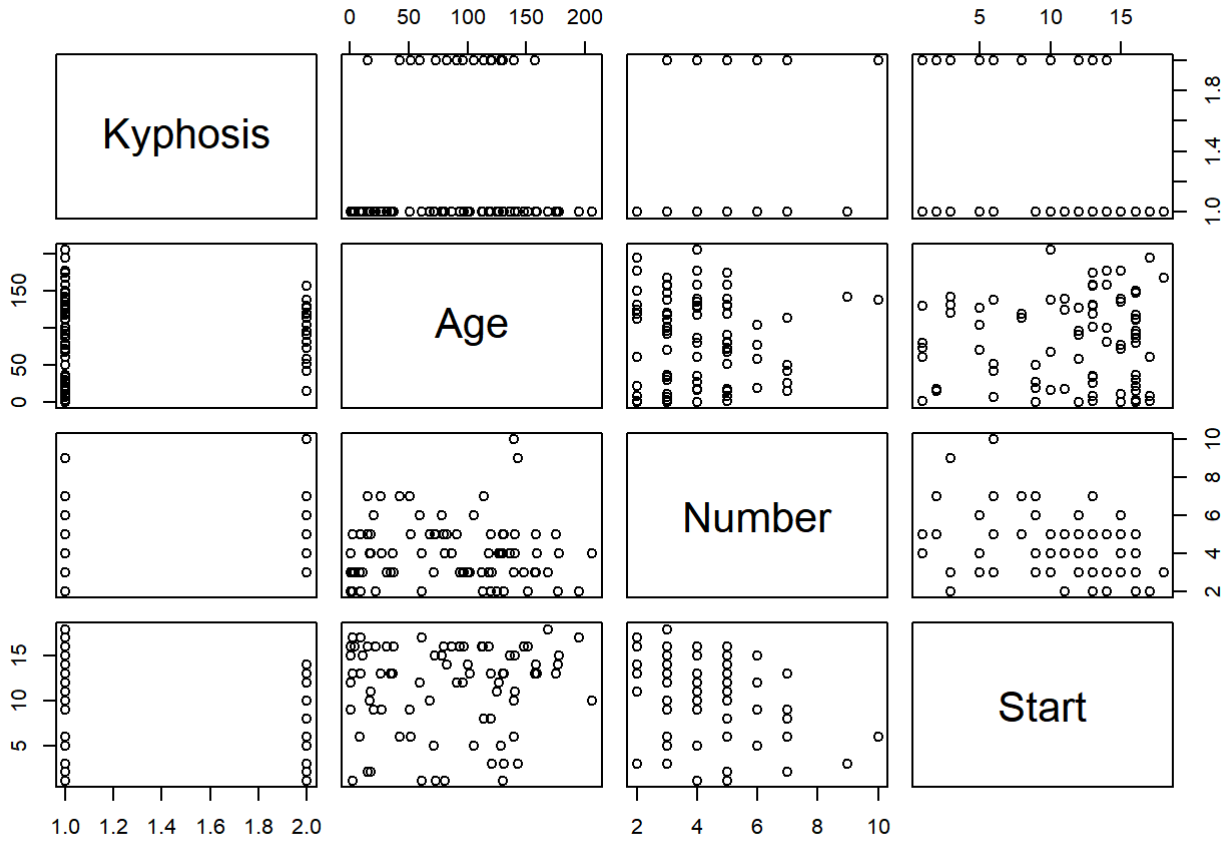
```
str(kyphosis)
```

```
## 'data.frame':    81 obs. of  4 variables:
##  $ Kyphosis: Factor w/ 2 levels "absent","present": 1 1 2 1 1 1 1 1 1 2 ...
##  $ Age     : int  71 158 128 2 1 1 61 37 113 59 ...
##  $ Number  : int  3 3 4 5 4 2 2 3 2 6 ...
##  $ Start   : int  5 14 5 1 15 16 17 16 16 12 ...
```

```
head(kyphosis)
```

```
##   Kyphosis Age Number Start
## 1   absent  71      3     5
## 2   absent 158      3    14
## 3  present 128      4     5
## 4   absent   2      5     1
## 5   absent   1      4    15
## 6   absent   1      2    16
```

```
plot(kyphosis)
```



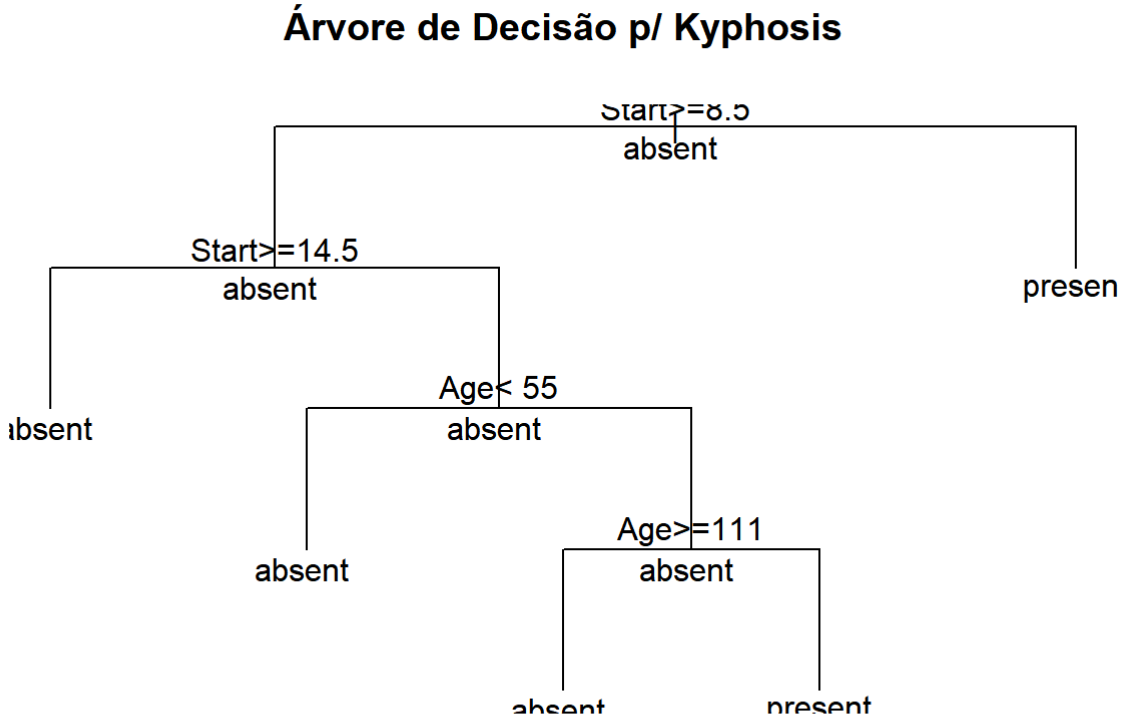
Modelo de Árvore de Decisão

```
arvore <- rpart(Kyphosis ~ ., method = "class", data = kyphosis)
printcp(arvore)
```

```
##
## Classification tree:
## rpart(formula = Kyphosis ~ ., data = kyphosis, method = "class")
##
## Variables actually used in tree construction:
## [1] Age   Start
##
## Root node error: 17/81 = 0.20988
##
## n= 81
##
##         CP nsplit rel error xerror    xstd
## 1 0.176471      0   1.00000      1 0.21559
## 2 0.019608      1   0.82353      1 0.21559
## 3 0.010000      4   0.76471      1 0.21559
```

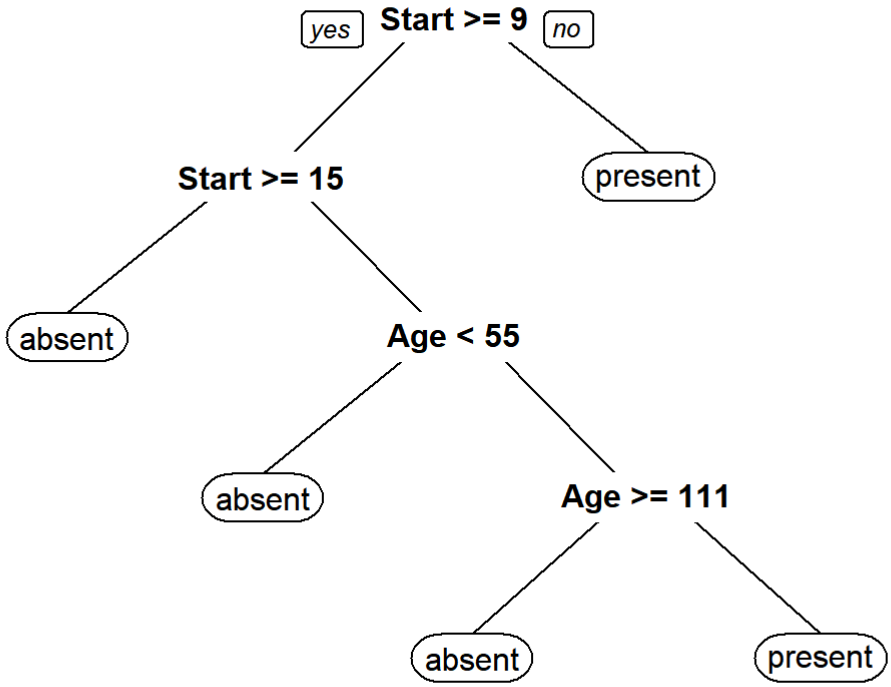
Plotando a Árvore de Decisão e inserindo as descrições

```
plot(arvore, uniform = T, main = "Árvore de Decisão p/ Kyphosis")
text(arvore, splits = T, all = T)
```



Plotagem de forma mais simples e menos poluída

```
prp(arvore)
```



Separando em dados de treino e teste

```
split = sample.split(kyphosis$Kyphosis, SplitRatio = 0.70)

treino <- subset(kyphosis, split == T)
teste <- subset(kyphosis, split == F)
```

Criando um novo modelo a partir dos dados de treino

```
arvore_2 <- rpart(Kyphosis ~ ., method = "class", data = treino)
```

Previsão

```
previsao <- predict(arvore_2, teste[-1], type = "class")
```

Comparando os valores reais com os previstos

```
realxprev <- data.frame(teste$Kyphosis, previsao)
print(realxprev)
```

```
##    teste.Kyphosis previsao
## 2          absent   absent
## 3         present  present
## 4          absent  present
## 8          absent   absent
## 10        present   absent
## 12         absent   absent
## 14         absent   absent
## 19         absent   absent
## 36         absent   absent
## 37         absent   absent
## 39         absent   absent
## 40        present   absent
## 42         absent   absent
## 43         absent  present
## 44         absent  present
## 45         absent   absent
## 47         absent   absent
## 48         absent   absent
## 51         absent   absent
## 62        present  present
## 71         absent   absent
## 73         absent   absent
## 76         absent   absent
## 80        present  present
```