Análisis de `rpart::kyphosis`

Pero Grullo

30 de marzo de 2023

1. Datos

Datos originales: `kyphosis`, distribuidos con el paquete `rpart` de R. Traducidos en `cifosis` como se indica en el apéndice A; contiene las variables:

`Cifosis` – Toma valores `presente` o `ausente` en función de si el individuo de la muestra presenta o no cifosis (cierta deformación de la columna vertebral) tras haber sido sometido a cirugía correctiva.

`Edad` – Edad en meses.

`Número` – Cuántas vértebras hay involucradas.

`Inicio` – Posición (desde lo alto) de la primera vértebra operada.

2. Objetivo

Obtener un modelo para predecir `Cifosis` a partir de las variables explicativas `Edad`, `Número` e `Inicio`. Describir la calidad del modelo.

3. Descriptiva

3.1. Univariante

`Cifosis` – Muchos más casos de cirugía exitosa (`ausente`). Frecuencias:

```
           ausente presente
absolutas       64       17
relat. (%)      79       21
```

`Edad` – Menores de 18 años, con acumulación de bebés (fig. 1, pág. 4).

`Número` – En general, de 2 a 5 vértebras afectadas (fig. 2, pág. 4).

`Inicio` – Casos desde la 1 (la más alta), pero mayor concentración entre la 13 y la 16 (fig. 3, pág. 5).

3.2. Bivariante

Correlación negativa entre `Número` e `Inicio`, como cabe esperar. Correlación nula en el resto de parejas de explicativas (fig. 4, pág. 5), sin que haya relación no lineal (fig. 5, pág. 6).

La variable respuesta está asociada a cada variable explicativa, sobre todo con `Inicio` (fig. 6, pág. 7). En gran medida, `Inicio` $> 12 \implies$ `ausente`, `Inicio` $< 11 \implies$ `presente`; $> 70\,\%$ de acierto (fig. 7, pág. 8).

4. Modelo predictivo

Usaremos la técnica de análisis discriminante por ser la única supervisada vista en esta asignatura.

```
##
##            FALSE TRUE
##   ausente     67    0
##   presente     0   14
```

El modelo construye el discriminante lineal 0,006 · `Edad` + 0,292 · `Número` + -0,17 · `Inicio`. Dicho modelo clasifica como `presente` para valores $> 1{,}2$ aproximadamente (fig. D, pág. 8). Los signos de los coeficientes indican que la predicción tiende a `presente`

- a mayor `Edad`,
- a mayor `Número`, y
- a menor `Inicio`.

Para cuantificar la importancia de las variables se puede aplicar el discriminante lineal a los datos tipificados (fig. D, pág. 9), lo que produce el modelo 0,343 · `Edad` + 0,472 · `Número` + -0,833 · `Inicio`, que refleja lo visto antes: `Inicio` es la variable más asociada a la respuesta.

Mediante validación cruzada dejando uno fuera (yacnaif) la matriz de confusión es

```
absolutas:
          pred
real       ausente presente
  ausente       58        6
  presente      10        7
```

```
porcentajes:
          pred
real       ausente presente
  ausente       91        9
  presente      59       41
```

Que para los `presente` haya más errores que aciertos puede llevar a poner en duda la conveniencia de usar probabilidades a priori estimadas a partir de la muestra. El análisis hecho con aprioris equiprobables da

```
absolutas:
          pred
real       ausente presente
  ausente       51       13
  presente       6       11
```

```
porcentajes:
          pred
real       ausente presente
  ausente       80       20
  presente      35       65
```

y podría ser preferido un resultado así, más equilibrado entre los dos grupos. Como desconocemos posibles cuantificaciones de pérdidas por error, mantendremos el modelo calculado en primer lugar.

Un discriminante cuadrático mejora algo los resultados (fig. D, pág. 9) cuando los aprioris se estiman de la muestra, pero no parece una mejora relevante.

5. Conclusiones

- La variable más relevante para predecir `Cifosis` es `Inicio`. Mediante ella sola se consigue más del 70 % de aciertos.
- Se ha llegado a la fórmula

$$F = 0{,}006 \cdot \texttt{Edad} + 0{,}292 \cdot \texttt{Número} + \text{-}0{,}17 \cdot \texttt{Inicio}$$

  que clasifica como `present` si $F > 1{,}2$ aproximadamente.
- Dicha fórmula consigue un 91 % de aciertos en `ausente` pero sólo un 41 % entre los `presente`; en total, un 80 %.

A. Traducción

```
cifosis <- rpart::kyphosis
names(cifosis) <- c("Cifosis", "Edad", "Número", "Inicio")
levels(cifosis$Cifosis) <- c("ausente", "presente")
head(cifosis)
```

```
options (OutDec = ",")
histograma <- function (x, ...)
    hist(x, ..., main="", xlab="", ylab = "Frecuencias")
barras <- function (x, ...) barplot (table (x), ..., main="", xlab = "",
                                     ylab = "Frecuencias")
```

B. Descriptiva univariante

Figura 1: Distribución de la `Edad` (en meses).

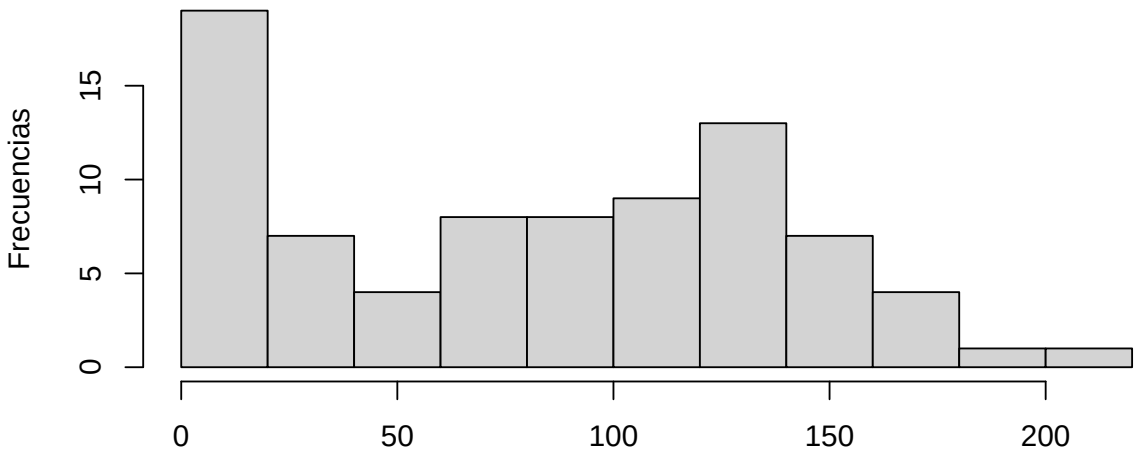


Figura 2: Distribución del `Número` de vértebras afectadas.

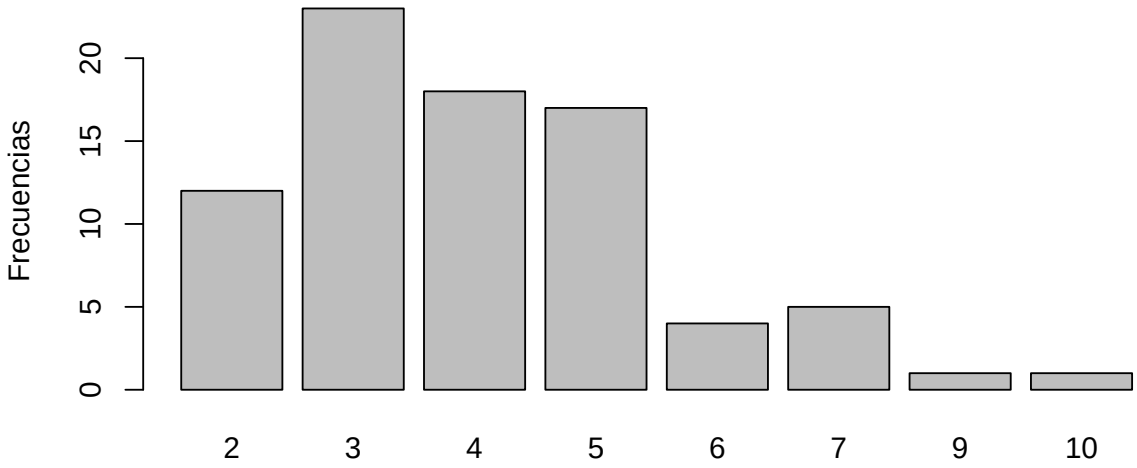
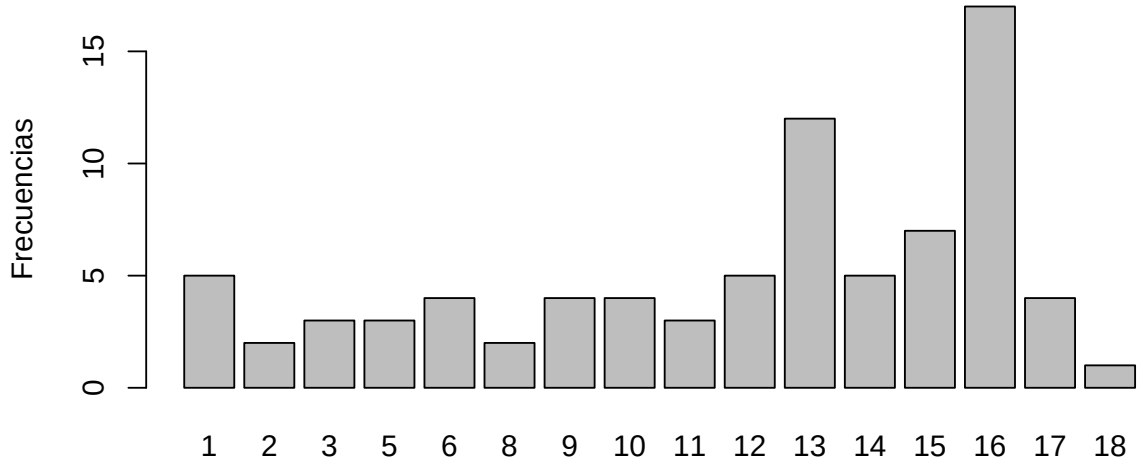


Figura 3: Distribución del `Inicio` (vértebra inicial, desde arriba).



C. Descriptiva bivariante

Figura 4: Matriz de correlaciones entre variables explicativas.

```
              Edad     Número      Inicio
Edad    1,00000000 -0,0166875  0,05782789
Número -0,01668750  1,0000000 -0,42509875
Inicio  0,05782789 -0,4250988  1,00000000
```

Figura 5: Nubes de puntos de variables explicativas.

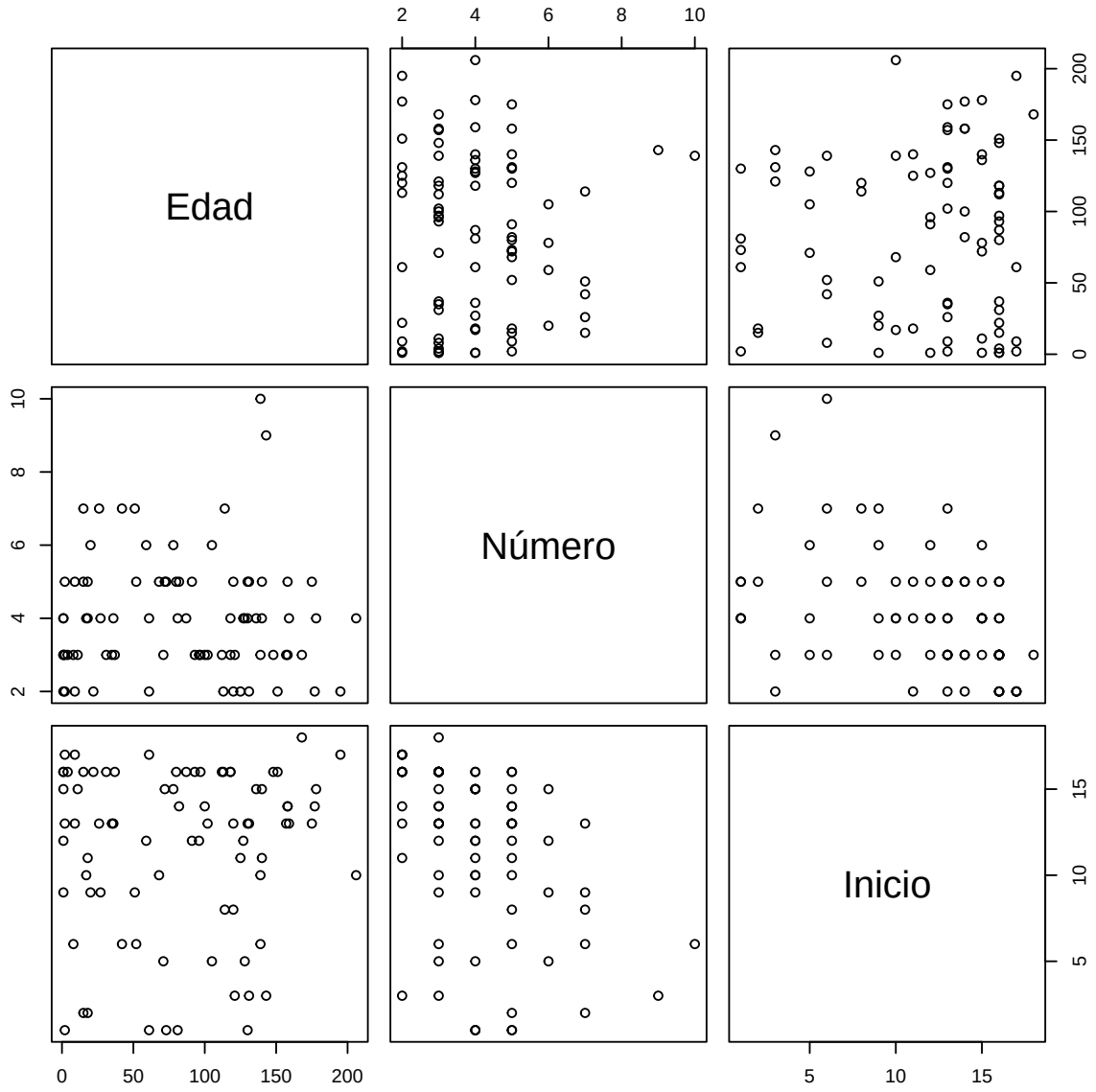


Figura 6: Diagrama de cajas según `Cifosis`.

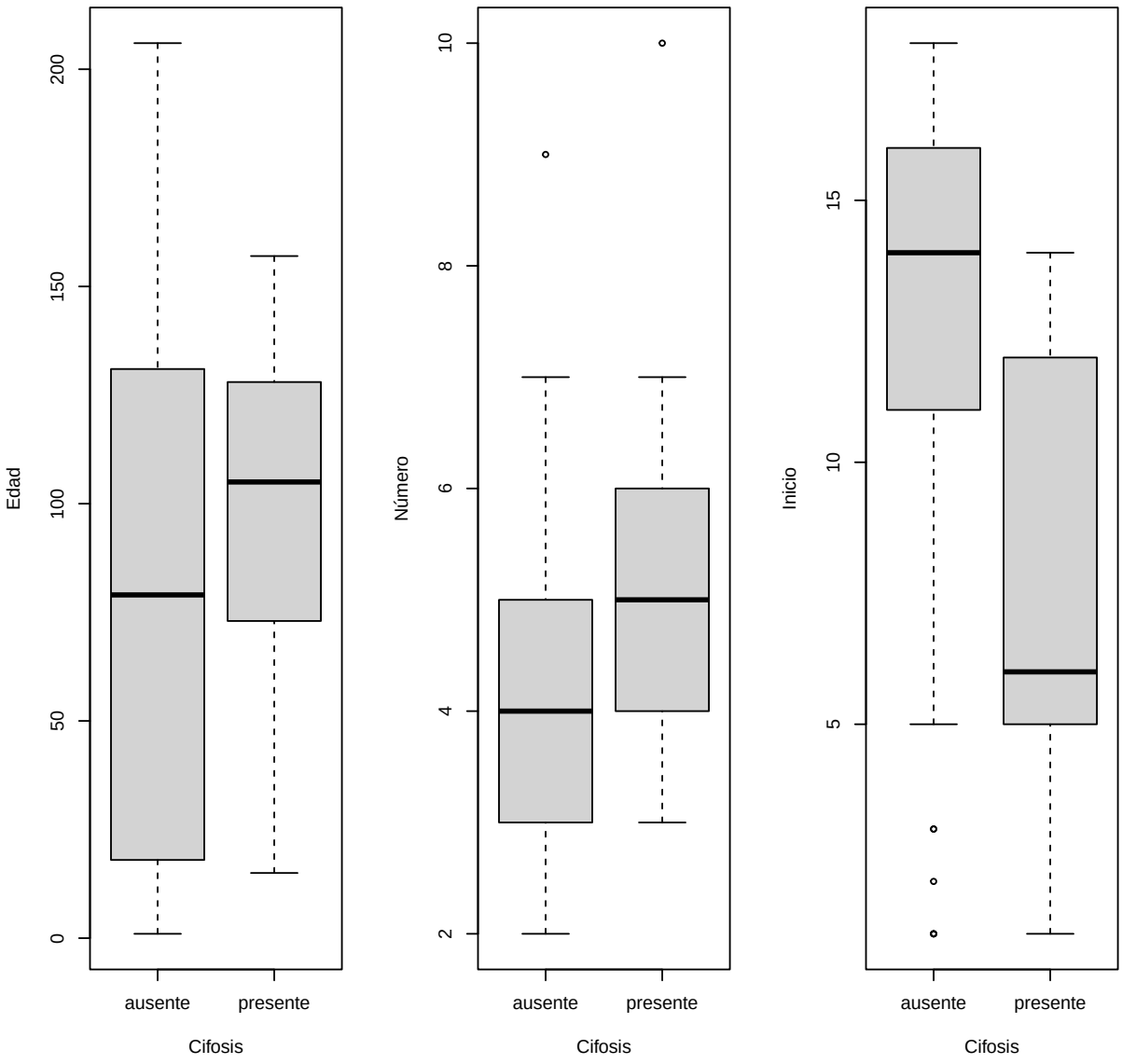


Figura 7: Mejor matriz de confusión usando sólo `Inicio`.

```
prop.table (table (real=cifosis$Cifosis, pred=cifosis$Inicio<11), "real")

##           pred
## real           FALSE      TRUE
##   ausente  0,7656250 0,2343750
##   presente 0,2941176 0,7058824
```

D. Discriminante

Figura 8: Discriminante lineal con variables originales.

```
library (MASS)
modelo <- lda (Cifosis ~ ., cifosis)
pmodelo <- predict (modelo)
table (pmodelo$class, pmodelo$x > 1.2)

##
##            FALSE TRUE
##   ausente     67    0
##   presente     0   14
```

Figura 9: Discriminante lineal con variables tipificadas.

```
library (MASS)
modelo <- lda (Cifosis ~ ., cifosis)
pmodelo <- predict (modelo)
table (pmodelo$class, pmodelo$x > 1.2)

##
##             FALSE TRUE
##   ausente      67    0
##   presente      0   14
```

Figura 10: Discriminante cuadrático.

```
Aprioris MUESTRALES
absolutas:
          pred
real       ausente presente
  ausente       59        5
  presente      10        7
```

```
Aprioris EQUIPROB.
absolutas:
          pred
real       ausente presente
  ausente       51       13
  presente       6       11
```

```
porcentajes:
          pred
real       ausente presente
  ausente       92        8
  presente      59       41
```

```
porcentajes:
          pred
real       ausente presente
  ausente       80       20
  presente      35       65
```

E. Código R de los análisis

Descriptiva univariante:

```
rbind (absolutas = (t <- table(cifosis$Cifosis)),
       "relat. (%)" = round(100*prop.table(t)))
histograma(cifosis$Edad)
barras(cifosis$Número)
barras(cifosis$Inicio)
```

Descriptiva bivariante:

```
cor (subset (cifosis, , -Cifosis))
pairs (subset (cifosis, , -Cifosis))
par(mfrow=c(1,3))
boxplot (Edad   ~ Cifosis, cifosis)
boxplot (Número ~ Cifosis, cifosis)
boxplot (Inicio ~ Cifosis, cifosis)
par(mfrow=c(1,1))
```

Discriminante lineal:

```
library (MASS)
modelo <- lda (Cifosis ~ ., cifosis)
pmodelo <- predict (modelo)
table (pmodelo$class, pmodelo$x > 1.2)
tabla <- table (real=cifosis$Cifosis, pred=lda(Cifosis~.,cifosis,CV=TRUE)$class)
cat ("absolutas:")
tabla
cat ("porcentajes:")
round (100 * prop.table (tabla, "real"))
```

Discriminante cuadrático:

```
cat ("Aprioris MUESTRALES")
tabla <- table (real=cifosis$Cifosis,
                pred=qda(Cifosis~.,cifosis,CV=TRUE)$class)
cat ("absolutas:")
tabla
cat ("porcentajes:")
round (100 * prop.table (tabla, "real"))
cat ("Aprioris EQUIPROB.")
tabla <- table (real=cifosis$Cifosis,
                pred=qda(Cifosis~.,cifosis,CV=TRUE,prior=c(.5,.5))$class)
cat ("absolutas:")
tabla
```

```
cat ("porcentajes:")
round (100 * prop.table (tabla, "real"))
```