Practica 1 Clasificación

Federico Ros

31 de octubre de 2017

PRACTICA 1: GENDER DISCRIMINATION DECISSION TREE

1) Cargar el fichero y paquetes necesarios par el desarrollo del ejercicio:

```
setwd("/Users/fede/Downloads")
datos<- read.csv("GenderDiscrimination.csv")
head(datos)
```

```
##   Gender Experience Salary
## 1 Female         15  78200
## 2 Female         12  66400
## 3 Female         15  61200
## 4 Female          3  61000
## 5 Female          4  60000
## 6 Female          4  68000
```

```
# una vez cargardos los datos de nuestro csv para el ejercicio cargarmos los
#paquetes que necesitaremos:
library(tree)
library(rpart)
library(rpart.plot)
library(partykit)
```

```
## Loading required package: grid
```

2) Definicion del modelo y la variable a estudiar:

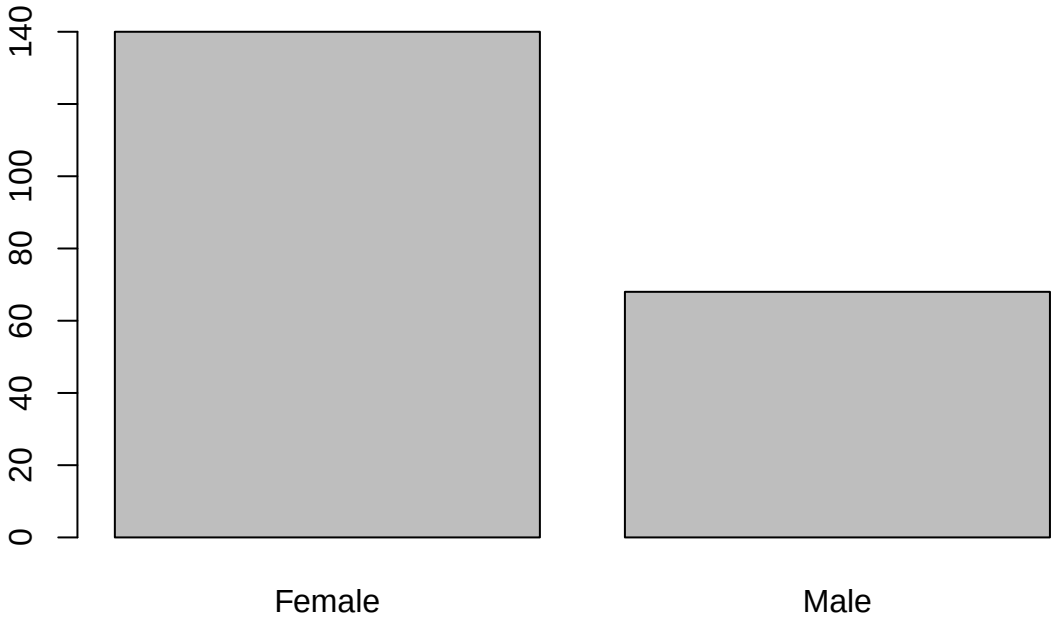
```
# vamos a estimar la variable Gender en funcion de las otras dos variables, expirience and salary:
# antes de nada comprobamos la estructura de nuestro data set en cuanto a
#proporcion de hombre y mujeres:
sum(datos$Gender=="Female")
```

```
## [1] 140
```

```
sum(datos$Gender=="Male")
```

```
## [1] 68
```

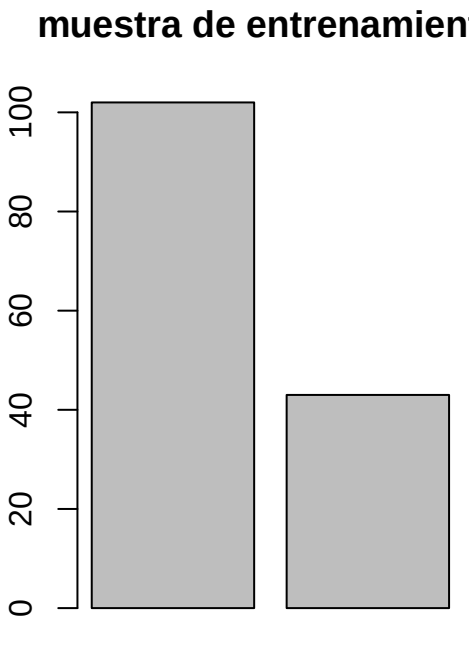
```
plot(datos$Gender)
```

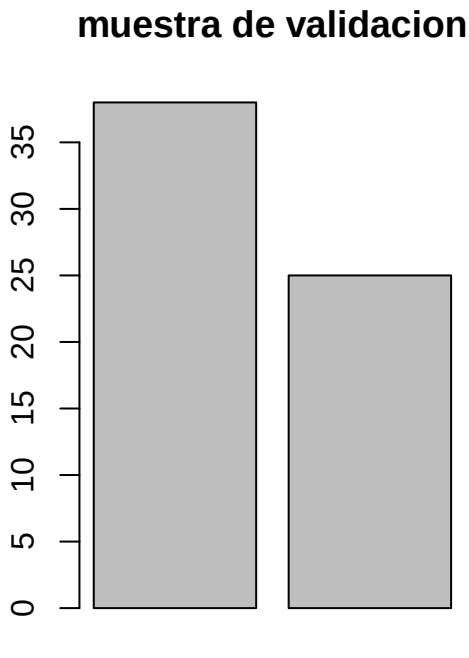



```
# por ultimo fijamos el data set "datos" para acceder a sus elementos sin necesidad de utilizar "$"
attach(datos)
```

3) Definicion de las muestras de validacion y entrenamiento:

```
# en primer lugar fijamos una semilla aleatoria
set.seed(1234)
# para continuar, definimos la muestra aleatoria de nuevo de aprendizaje,
#en la que utilizaremos el 70% de nuestra muestra tal como en el ejemplo realizado en clase:
train <- sample(nrow(datos), 0.7*nrow(datos))
par(mfrow=c(1,2))
# seguidamente definimos la muestra de entrenamiento y mostramos el grafico:
datos.train <- datos[train,]
plot(datos.train$Gender, main="muestra de entrenamiento")
# a continuacion definimos la muestra de validacion con el 30% de la muestra del estudio:
datos.validate <- datos[-train,]
plot(datos.validate$Gender, main="muestra de validacion")
```





4) El siguiente paso sera construir nuestro arbol de decision para, posteriormente ir ajustandolo para concluir nuestro ejercicio:

```
arboldecision <- rpart(Gender~ ., data=datos.train,method="class", parms=list(split="infor
mation"))
# pintamos el arbol
print(arboldecision)
```

```
## n= 145
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 145 43 Female (0.7034483 0.2965517)
##    2) Salary< 92300 123 26 Female (0.7886179 0.2113821)
##      4) Experience>=6.5 91  9 Female (0.9010989 0.0989011) *
##      5) Experience< 6.5 32 15 Male (0.4687500 0.5312500)
##       10) Experience>=4.5 19  8 Female (0.5789474 0.4210526) *
##       11) Experience< 4.5 13  4 Male (0.3076923 0.6923077) *
##    3) Salary>=92300 22  5 Male (0.2272727 0.7727273) *
```

1º. 123 personas no superan los 92.300 de salario de los cuales alrededor del 79% son mujeres y el 21% hombres.

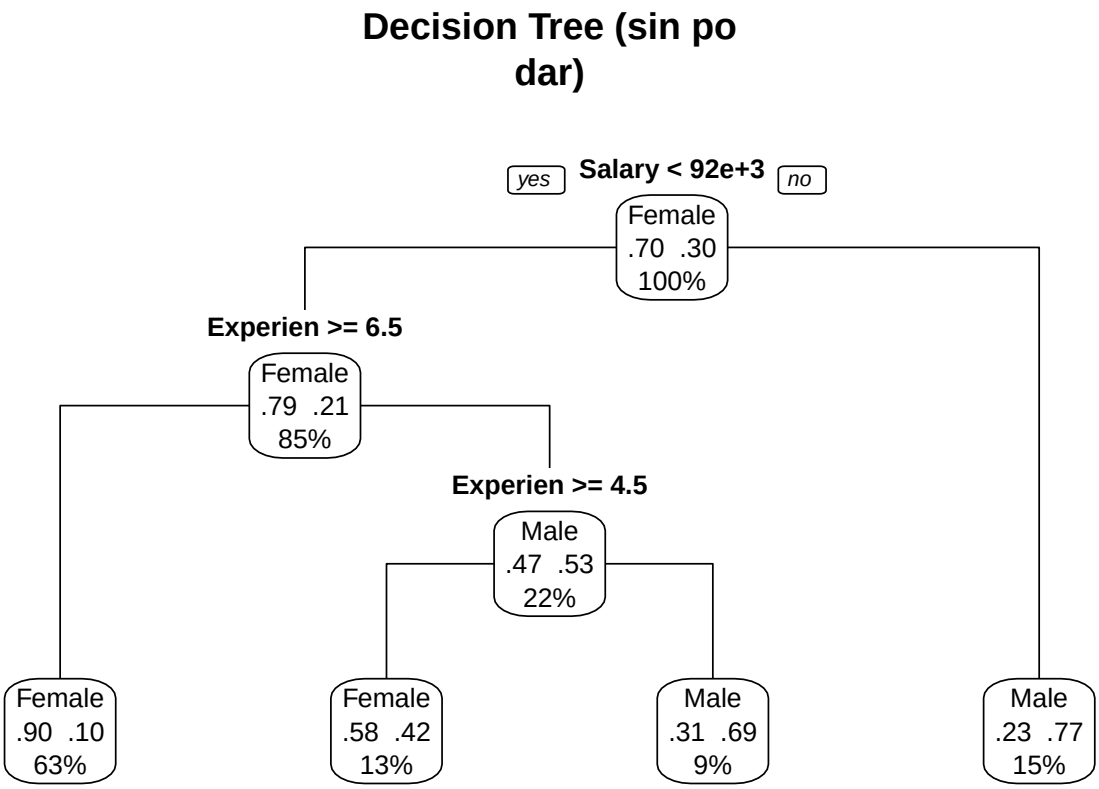
2º. 91 observaciones que cumplen la condicion de tener 6,5 años o mas de experiencia el 90% son mujeres y el 10% hombres.

3º. de las 32 observaciones que cumplen la condicion de menos de 6,5 años de experiencia alrededor del 53% son hombres y 47% mujeres.

4ºDe las observaciones que cumplen la condicion de tener menos de 6,5 años de experiencia volvemos a dividir donde los que tienen 4,5 años de experiencia o mas son 19 observacion en donde alrededor del 58% son mujeres y el 42% hombres. Y por ultimo, dentro de este mismo grupo las observaciones que tienen menos de 4,5 años de experiencia que son 13 alrededor del 70% son hombres y el 30% son mujeres.

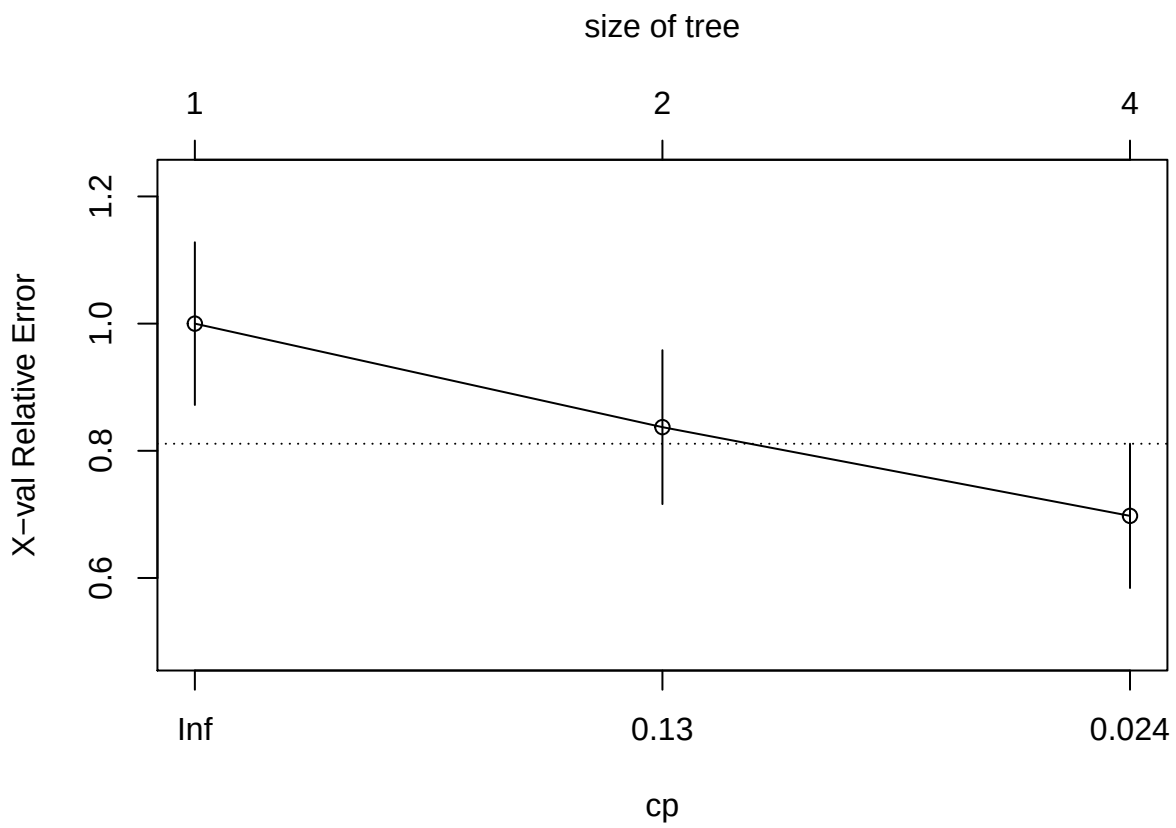
5) Arbol de decision y complejidad.

```
prp(arboldecision, type = 1, extra = 104, fallen.leaves = TRUE, main="Decision Tree (sin po
dar)")
```



Dibujamos ahora el grafico y tabla de complejidad de nuestro arbol:

```
plotcp(arboldecision)
```




```
arboldecision$cptable
```

```
##           CP nsplit rel error    xerror      xstd
## 1 0.27906977      0 1.0000000 1.0000000 0.1279033
## 2 0.05813953      1 0.7209302 0.8372093 0.1209796
## 3 0.01000000      3 0.6046512 0.6976744 0.1134376
```

Cogeriamos el nsplit con menor xerror pero en este caso, tal y como hemos comentado en clase, si el menor xerror sumado a su xstd es mayor o igual que el superior se coge el superior, por tanto cogemos la opcion del nodo 2.

6) Poda de nuestro arbol de decision.

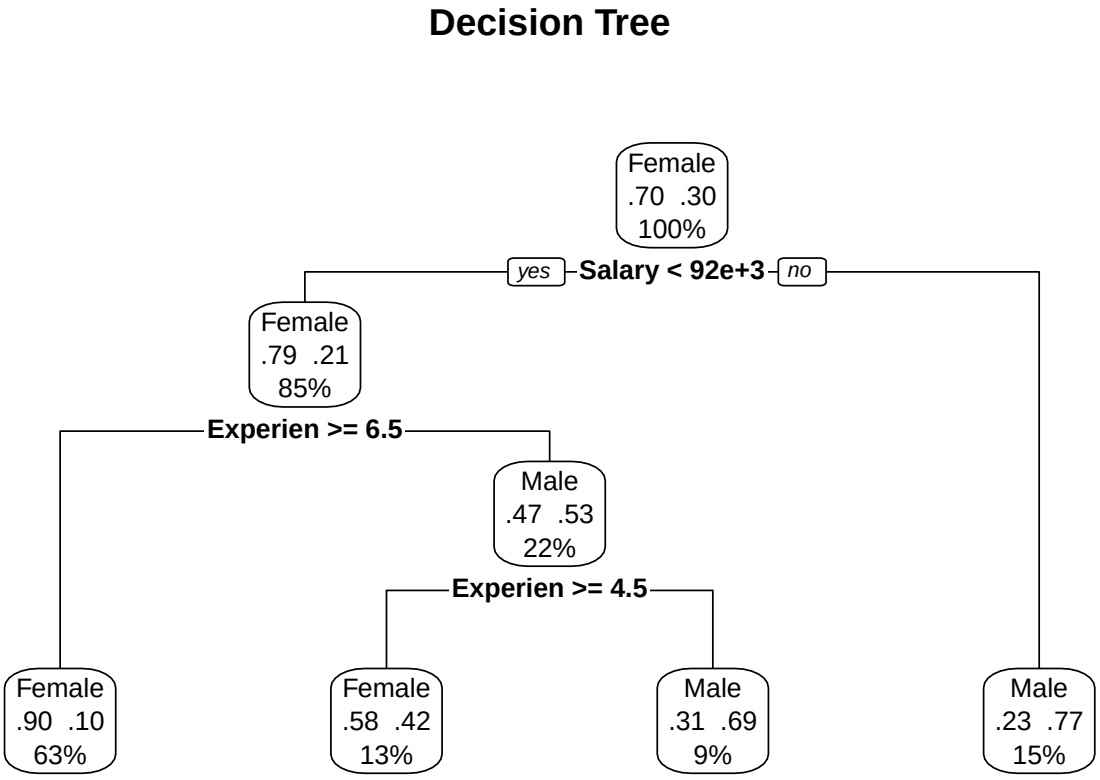
```
arbolPodado <- prune (arboldecision, cp=0.05813953)
arbolPodado
```

```
## n= 145
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 145 43 Female (0.7034483 0.2965517)
##    2) Salary< 92300 123 26 Female (0.7886179 0.2113821)
```

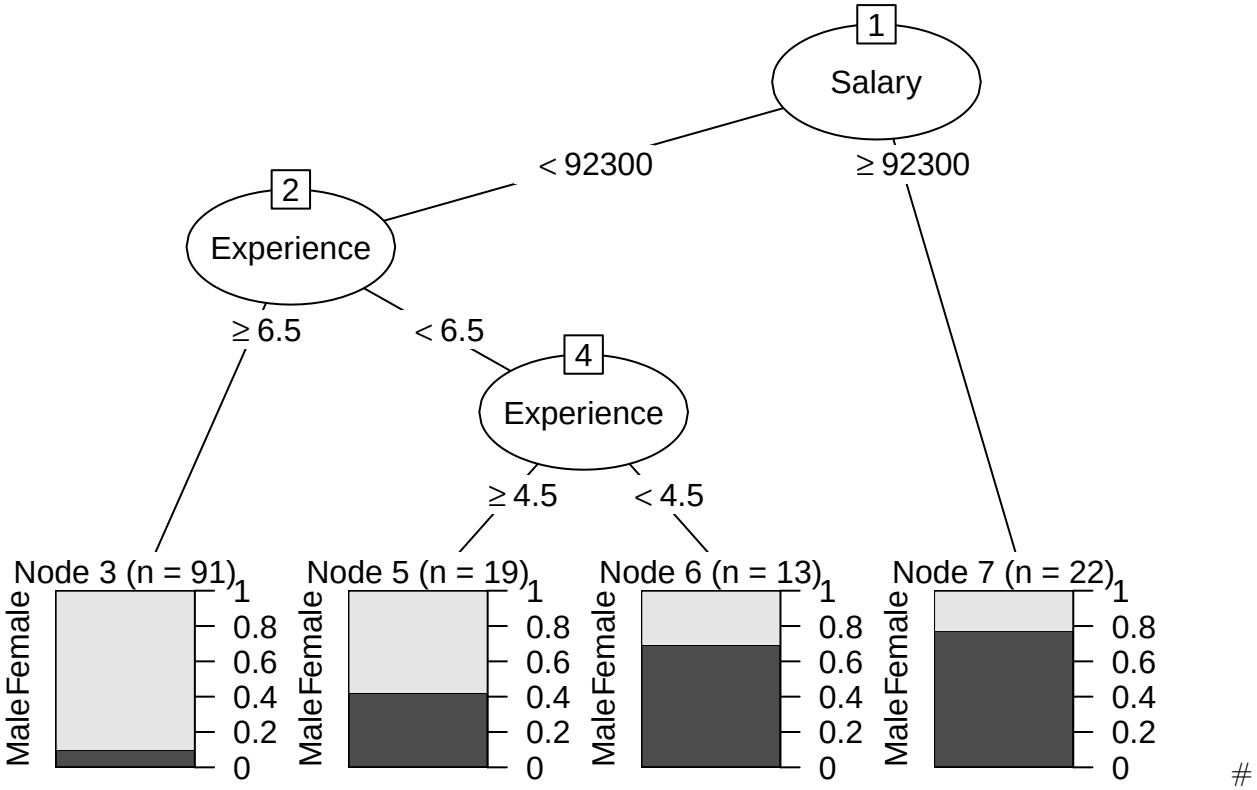
```
##      4) Experience>=6.5 91  9 Female (0.9010989 0.0989011) *
##      5) Experience< 6.5 32 15 Male (0.4687500 0.5312500)
##       10) Experience>=4.5 19  8 Female (0.5789474 0.4210526) *
##       11) Experience< 4.5 13  4 Male (0.3076923 0.6923077) *
##    3) Salary>=92300 22  5 Male (0.2272727 0.7727273) *
```

Dibujamos el nuevo arbol con las dos formas que hemos visto en clase.

```
prp(arbolPodado, type = 2, extra = 104, fallen.leaves = TRUE, main="Decision Tree
")
```




```
plot(as.party(arbolPodado))
```



#

Finalmente realizamos la comprobacion de la mejora del arbolPodado con respecto al arboldecision:

```
# Arbol sin podar
arbol.pred <- predict(arboldecision, datos.validate, type="class")
arbol.perf <- table(datos.validate$Gender, arbol.pred, dnn=c("Actual", "Predicte
d"))
arbol.perf
```

```
##          Predicte
## d
## Actual   Female Male
##   Female     27   11
##   Male       12   13
```

Arbol podado:

```
arbol.pred <- predict(arbolPodado, datos.validate, type="class")
arbol.perf<- table(datos.validate$Gender, arbol.pred, dnn=c("Actual", "Predicte
d"))
arbol.perf
```

```
##          Predicte
## d
## Actual   Female Male
##   Female     27   11
##   Male       12   13
```

7) Conclusiones:

Como interpretacion final, concluimos que, dada esta muestra el porcentaje de mujeres que se registran (alrededor del 70%) explica que el porcentaje de observaciones que reciben un salario superior a 92.300 sean mujeres y tan solo alrededor de un 12% sean hombres. Si analizamos mas a fondo estos datos con los arboles de decision observamos que el porcetanje de hombres que perciben un salario superior a 92300 es proporcionalmente mayor al numero de mujeres, si bien la variable experiencia podemos concluir que no tiene mucha importancia en esta muestra. De los 67 hombres existentes en los datos,alrededor de un 40% tienen un sueldo superior a los 92300 euros y tan solo 11 de las mujeres (alrededor de un 7% de la muestra) tienen un sueldo superior a los 92300 euros.