Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu

Practica 1 – Gender Discrimination

Carga de datos

Cargamos los datos. Problema de negocio: clasificar individuos como mujeres/hombres en funcion de la experiencia y el salario que perciben. Por lo tanto, nuestro modelo a estimar es Gender~ Experience + Salary.

```
setwd("C:/Users/Manuel/Desktop/CUNEF/Tecnicas de clasificacion/Practica1-arboles")
gender <- read.csv("http://www.biz.uiowa.edu/faculty/jledolter/DataMining/GenderDiscrimination.csv")
head(gender, 6)
```

```
##   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
```

```
# comprobamos el encabezado de gender
## # modelo a estimar Gender~ Experience + Salary

View(gender)
#para explorar los datos
```

Instalacion de los paquetes necesarios para el análisis:

```
#install.packages("rpart")
library(rpart)
#install.packages("rpart.plot")
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.4.2
```

```
#install.packages("partykit")
library(partykit)
```

```
## Warning: package 'partykit' was built under R version 3.4.2
```

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

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu

Creamos muestra aleatoria

Vamos a utilizar una muestra con set seed previo y definimos una muestra aleatoria de aprendizaje del arbol. Con esto evitamos usar toda la poblacion para no sobrecargar el arbol.

```
set.seed(1379)
train <- sample(nrow(gender), 0.7*nrow(gender))
#esto me selecciona al azar el 70% de la muestra

gender.train <- gender[train,]   #con los elementos de la muestra que
acabo de crear
gender.validate <- gender[-train,]  #con los elementos restantes
```

Queremos ver la frecuencia de la variable Gender en la muestra train y en la de validacion:

```
table(gender.train$Gender) #93 mujeres y 52 hombres

##
## Female   Male
##     93     52

table(gender.validate$Gender) #47 mujeres y 16 hombres

##
## Female   Male
##     47     16
```

Estimacion del arbol

Utilizamos la libreria rpart

```
arbol <- rpart(Gender ~ ., data=gender.train, method="class",
               parms=list(split="information"))

print(arbol) #esta info sera mas completa con la representacion grafica

## n= 145
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 145 52 Female (0.6413793 0.3586207)
##    2) Salary< 92300 117 31 Female (0.7350427 0.2649573)
##      4) Experience>=6.5 84 12 Female (0.8571429 0.1428571) *
##      5) Experience< 6.5 33 14 Male (0.4242424 0.5757576)
##       10) Salary< 70500 19  9 Female (0.5263158 0.4736842) *
##       11) Salary>=70500 14  4 Male (0.2857143 0.7142857) *
##    3) Salary>=92300 28  7 Male (0.2500000 0.7500000) *
```

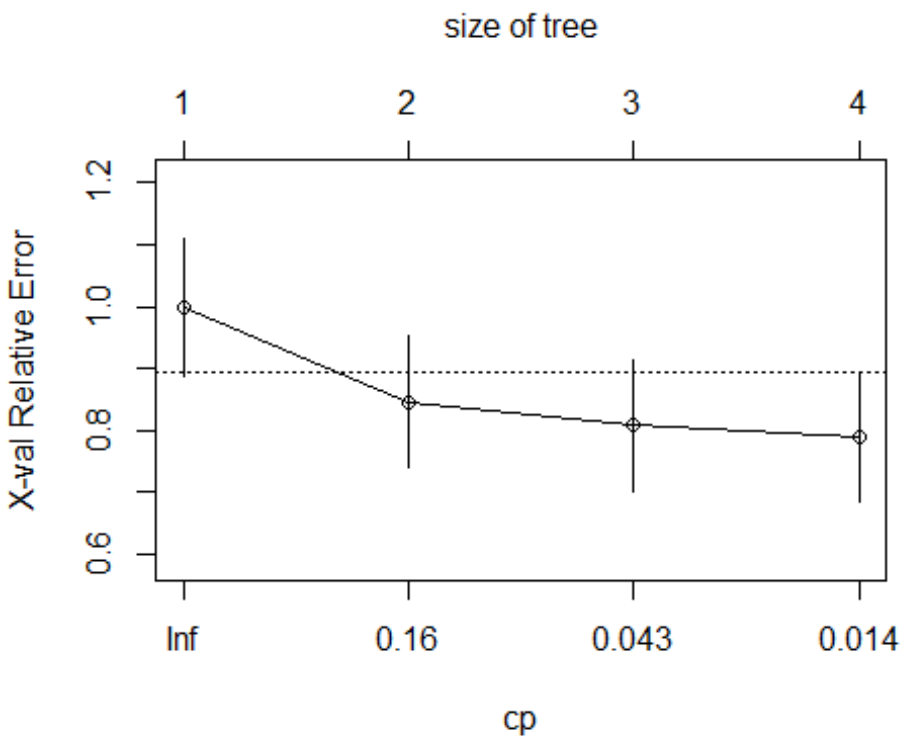
```
summary(arbol)
```

```
## Call:
## rpart(formula = Gender ~ ., data = gender.train, method = "class",
##     parms = list(split = "information"))
##   n= 145
##
##           CP nsplit rel error    xerror      xstd
## 1 0.26923077      0 1.0000000 1.0000000 0.1110595
## 2 0.09615385      1 0.7307692 0.8461538 0.1064632
## 3 0.01923077      2 0.6346154 0.8076923 0.1050403
## 4 0.01000000      3 0.6153846 0.7884615 0.1042849
##
## Variable importance
## Experience     Salary
##         52         48
##
## Node number 1: 145 observations,    complexity param=0.2692308
##   predicted class=Female  expected loss=0.3586207  P(node) =1
##     class counts:    93    52
##    probabilities: 0.641 0.359
##   left son=2 (117 obs) right son=3 (28 obs)
##   Primary splits:
##       Salary     < 92300 to the left,  improve=11.237700, (0 missing)
##       Experience < 6.5   to the right, improve= 4.235616, (0 missing)
##   Surrogate splits:
##       Experience < 20.5  to the left,  agree=0.841, adj=0.179, (0
split)
##
## Node number 2: 117 observations,    complexity param=0.09615385
##   predicted class=Female  expected loss=0.2649573  P(node) =0.8068966
##     class counts:    86    31
##    probabilities: 0.735 0.265
##   left son=4 (84 obs) right son=5 (33 obs)
##   Primary splits:
##       Experience < 6.5   to the right, improve=10.703500, (0 missing)
##       Salary     < 87900 to the left,  improve= 1.015875, (0 missing)
##
## Node number 3: 28 observations
##   predicted class=Male    expected loss=0.25  P(node) =0.1931034
##     class counts:     7    21
##    probabilities: 0.250 0.750
##
## Node number 4: 84 observations
##   predicted class=Female  expected loss=0.1428571  P(node) =0.5793103
##     class counts:    72    12
##    probabilities: 0.857 0.143
##
## Node number 5: 33 observations,    complexity param=0.01923077
##   predicted class=Male    expected loss=0.4242424  P(node) =0.2275862
```

```
##     class counts:    14    19
##    probabilities: 0.424 0.576
##   left son=10 (19 obs) right son=11 (14 obs)
##   Primary splits:
##       Salary     < 70500 to the left,  improve=0.9743636, (0 missing)
##       Experience < 4.5   to the right, improve=0.1545212, (0 missing)
##   Surrogate splits:
##       Experience < 4.5   to the left,  agree=0.727, adj=0.357, (0
split)
##
## Node number 10: 19 observations
##   predicted class=Female  expected loss=0.4736842  P(node) =0.1310345
##     class counts:    10     9
##    probabilities: 0.526 0.474
##
## Node number 11: 14 observations
##   predicted class=Male    expected loss=0.2857143  P(node) =0.09655172
##     class counts:     4    10
##    probabilities: 0.286 0.714
```

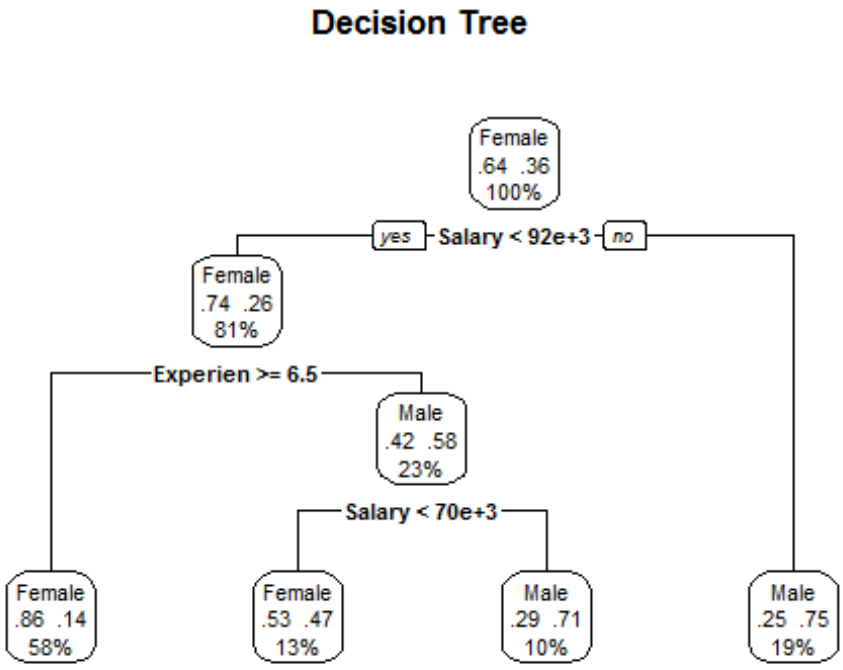
Representacion grafica

```
plotcp(arbol)
```




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

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu



Como podemos observar, el arbol tiene 4 nodos terminales y tres niveles distintos que hacen split en función de la variable salario y la experiencia. El primer nivel distingue en función de si el salario es superior o inferior a 92300.

El segundo nivel, que parte del grupo con salario inferior a 92300 (81%), distingue en función de si se tiene una experiencia laboral superior o inferior a 6.5 años. El terce nivel distingue entre aquellos individuos que, con un salario menor a 92300 y una experiencia laboral inferior a 6.5 años, tienen un salario menor o mayor a 70000.

Llama la atención que en este primer arbol de decisión los individuos con mas de 6.5 años de experiencia y con salario inferior a 92300 sean, en un 86%, mujeres.

Poda del arbol

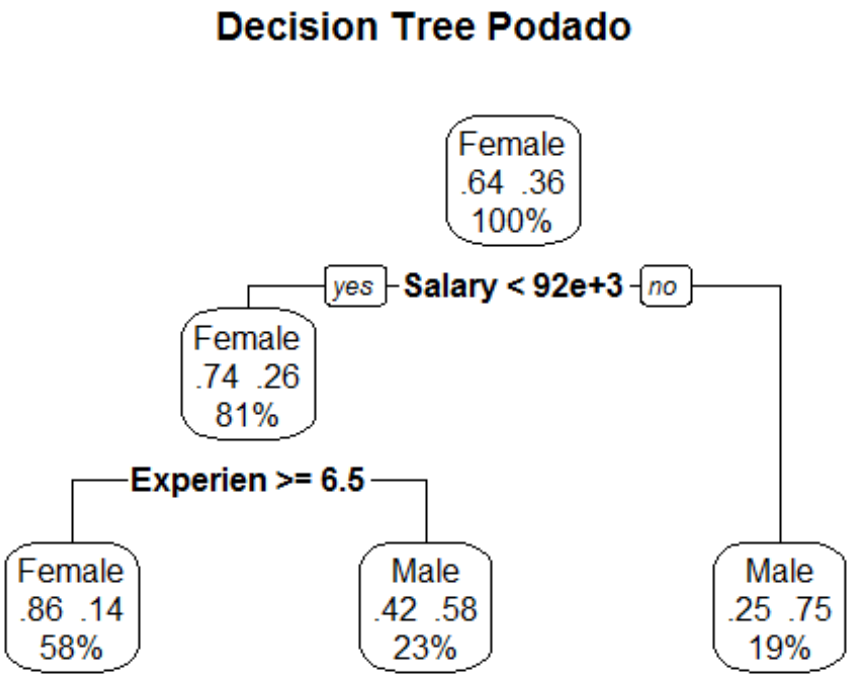
Ahora tenemos que podar el arbol. Para ellos nos guiaremos con la tabla de complejidad parametrica.Cogemos el xerror mas pequeño y le hacemos +/- la desv tipica. Si hay algun xerror mas pequeño que lo que nos salga, subo un nivel. Utilizamos la libreria rpart.plot para representar graficamente el arbol

```
arbol$cptable
```

```
##           CP nsplit rel error    xerror      xstd
## 1 0.26923077      0 1.0000000 1.0000000 0.1110595
## 2 0.09615385      1 0.7307692 0.8461538 0.1064632
## 3 0.01923077      2 0.6346154 0.8076923 0.1050403
## 4 0.01000000      3 0.6153846 0.7884615 0.1042849
```

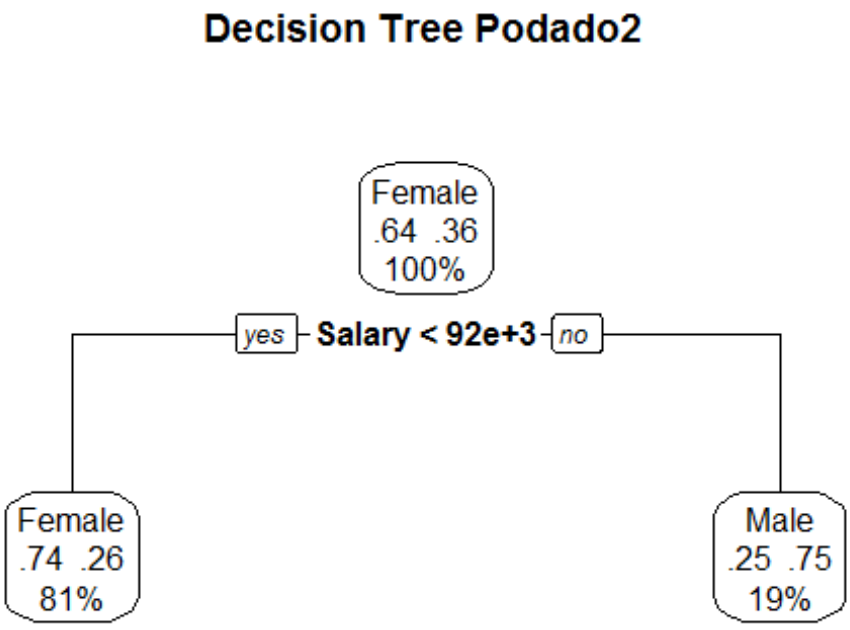
Podaremos en el nivel 3 primeramente y, viendo que se vuelve a cumplir la condicion de poda, volveremos a podar otro nivel. Obtenemos el arbol de decision arbol.podado2 con el que seguiremos el proceso.

```
arbol.podado <- prune(arbol, cp=0.01923077)
prp(arbol.podado, type = 2, extra = 104,
    fallen.leaves = TRUE, main="Decision Tree Podado")
```




```
arbol.podado$cptable

##           CP nsplit rel error    xerror      xstd
## 1 0.26923077      0 1.0000000 1.0000000 0.1110595
## 2 0.09615385      1 0.7307692 0.8461538 0.1064632
## 3 0.01923077      2 0.6346154 0.8076923 0.1050403

arbol.podado2 <- prune(arbol.podado, cp=0.09615385)

prp(arbol.podado2, type = 2, extra = 104,
    fallen.leaves = TRUE, main="Decision Tree Podado2")
```

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu



El arbol que usaremos para clasificar los individuos como hombres y mujeres utiliza solamente el primer criterio. Observamos que el 19% de la muestra cobra mas de 92300, de los cuales el 25% son mujeres y el 75% son hombres. La gran mayoria, el 81%, cobra menos de 92300 siendo el 74% mujeres y el 26% hombres.

Prediccion con la muestra de validacion con arbol.podado

Para verificar que la segunda poda es fructífera comparemos la capacidad de predicción de arbol.podado y arbol.podado2.

```
arbol.pred <- predict(arbol.podado, gender.validate, type="class")

arbol.perf <- table(gender.validate$Gender, arbol.pred,
                    dnn=c("Actual", "Predicted"))

arbol.perf
```

```
##          Predicted
## Actual    Female Male
##   Female     33   14
##   Male        5   11
```

Como podemos observar el nº total de errores que comete arbol.podado es 19.

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu

Prediccion con la muestra de validacion con arbol.podado2

```
arbol.pred <- predict(arbol.podado2, gender.validate, type="class")

arbol.perf <- table(gender.validate$Gender, arbol.pred,
                    dnn=c("Actual", "Predicted"))

arbol.perf
```

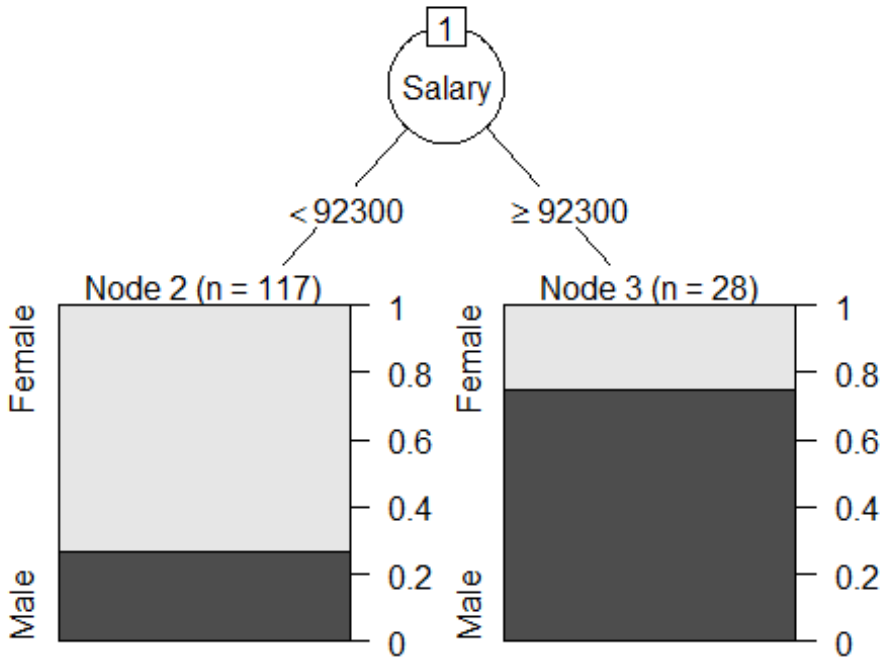
```
##          Predicted
## Actual   Female Male
##   Female     45    2
##   Male       11    5
```

Como podemos observar el numero de errores en arbol.podado2 es 13, menor que en el caso anterior. La segunda poda ha sido correcta.

Representacion grafica arbol.podado2

Utilizamos la libreria partykit para graficar

```
plot(as.party(arbol.podado2))
```



Hemos descartado la variable "experiencia" en nuestro arbol de decisión. Distinguiendo 2 nodos fundamentales en función del salario, obtenemos dos nodos

con una pureza aceptable. El nodo 1 está compuesto por un 23% aprox. de hombres y un 77% de mujeres. El nodo 3 está compuesto por un 77% aprox de hombres y un 23% de mujeres.

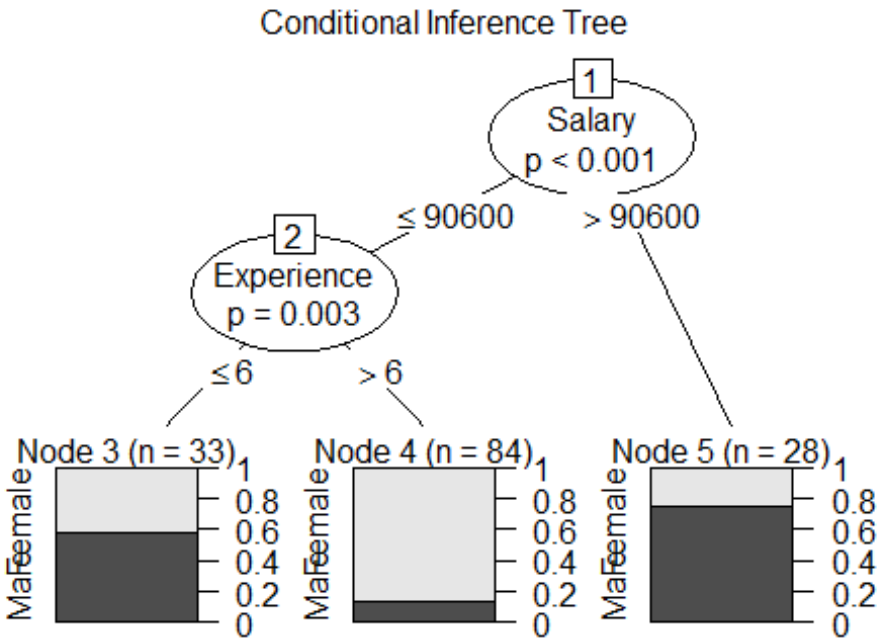
En este tipo de gráficos tenemos que observar la pureza de los nodos. Deben ser mayoritariamente blancos, o negros. La mezcla en proporciones similares no es buena señal en terminos de pureza y homogeneidad. Como podemos observar, nuestros nodos tienen una composición bastante homogénea.

Podemos interpretar que existe una brecha salarial importante entre hombres y mujeres puesto que la mayoría de los individuos con salario superior a 92300 son hombres y, al contrario, la mayoría de los individuos con salarior inferior a dicha cifra son mujeres.

Esto podría poner de relevancia uno de los principales conflictos de la sociedad del siglo XXI en los paises dearrollados pero, para poder afirmar si existe discriminación por género, tendríamos que ver si los hombres y mujeres de la muestra trabajan en el mismo sector y en cargos similares.

Metodo alternativo - Arboles basados en la inferencia. Conditional Inference Trees

```
fit.ctree <- ctree(Gender~ Experience + Salary, data=gender.train)

plot(fit.ctree, main="Conditional Inference Tree")
```



Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu

Los árboles basados en la inferencia (conditional inference tree) constituyen una variante importante de los árboles de decisión tradicional. Los árboles basados en la inferencia son similares a los tradicionales pero las variables y divisiones se basan en la significatividad de algunos contrastes más que en las medidas de puridad u homogeneidad. Así, nos fijaremos en el contraste del pvalor.

Este arbol utiliza la variable salario (menor o igual y mayor que 90600) como primer criterio de clasificacion. A diferencia del arbol anterior, utiliza la experiencia (menor o igual y mayor que 6) como segundo criterio de clasificación. Da lugar por tanto a 3 nodos terminales, 3, 4 y 5, de 33, 84 y 28 observaciones, respectivamente.

Aunque el nodo 4 y el nodo 5 son bastante homogéneos, el nodo 3 se sitúa en torno a una proporción 60-40, lo que incrementa la probabilidad de error.

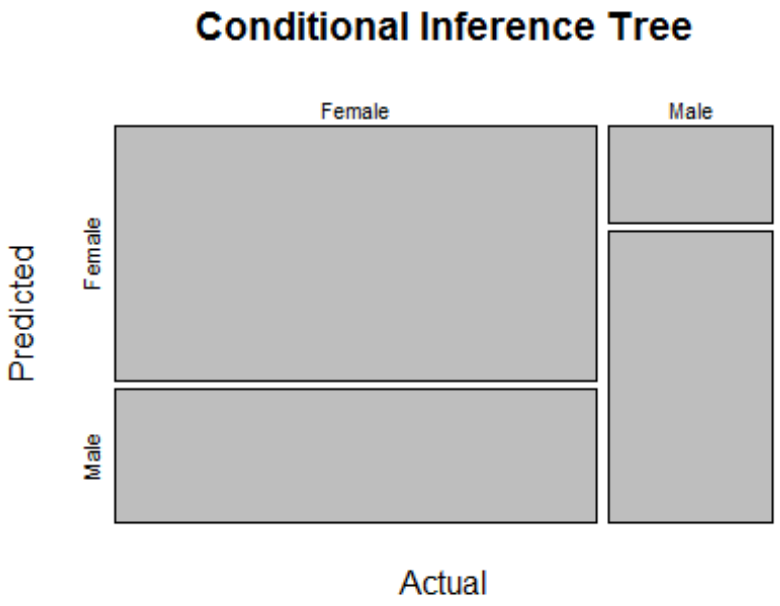
Hemos de comprobar si el arbol basado en la inferencia tiene proporciona una prediccion mas acertada a la hora de clasificar individuos en comparacion a arbol.podado2.

```
ctree.pred <- predict(fit.ctree, gender.validate, type="response")

ctree.perf <- table(gender.validate$Gender, ctree.pred,
                    dnn=c("Actual", "Predicted"))

ctree.perf
```

```
##          Predicted
## Actual   Female Male
##   Female     31   16
##   Male        4   12
```

```
plot(ctree.perf, main="Conditional Inference Tree")
```

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu



Como podemos observar el arbol basado en la inferencia da lugar a 20 errores, uno mas que nuestro arbol tradicional tras la primera poda y 7 mas que arbol.podado2. Aun asi, tal y como se observa en el grafico, este arbol permite corroborar la impresion general que generaban los arboles tradicionales:

- Existe una brecha salarial, en general las mujeres cobran menos.
- Para un mismo nivel de experiencia, los hombres cobran mas.

Sin embargo, como hemos señalado anteriormente, antes de hacer una afirmación rotunda habría que comprobar la distribución por sectores y cargos para comprobar que los individuos son comparables.

Podríamos decir que el modelo es relativamente robusto y parsimonioso puesto que llega a resultados parecidos por diferentes caminos y se explica con muy pocas variables.