Análisis Discriminante

Objetivo

- Clasificar a una nueva observación en su población correspondiente, habiendo observado previamente muestras de las poblaciones involucradas con su identificador de población.

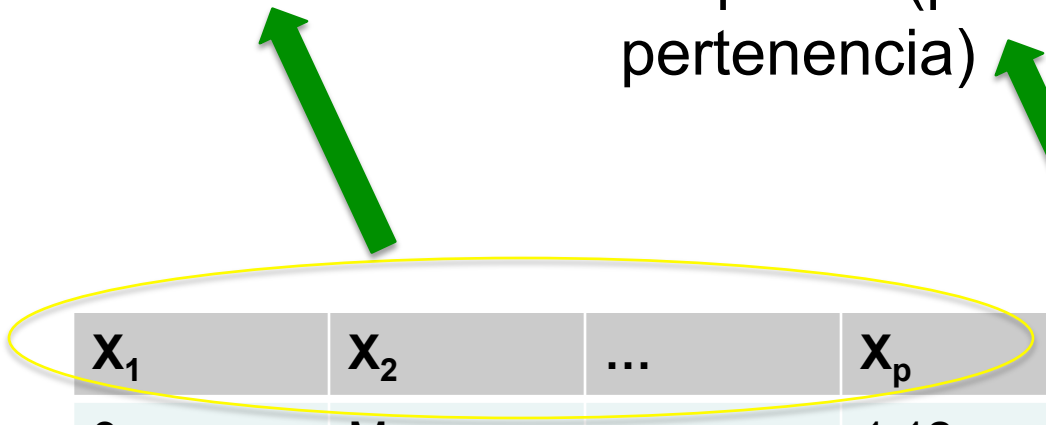
Algunos métodos

- K vecinos mas cercanos (KNN)
- Método lineal de Fisher (LDA)
- Método cuadrático de Fisher (QDA)
- Regresión logística
- Árboles de clasificación (CART)

Los datos

Atributos o covariables

Etiquetas (poblacion de pertenencia)



X_1	X_2	…	X_p	G
3	M	…	1.12	1
4	M	…	1.02	2
2	F	…	0.08	2
7	M	…	2.12	1
		…	3.22	…
3	M	…	1.56	2

El problema

Nuevos individuos

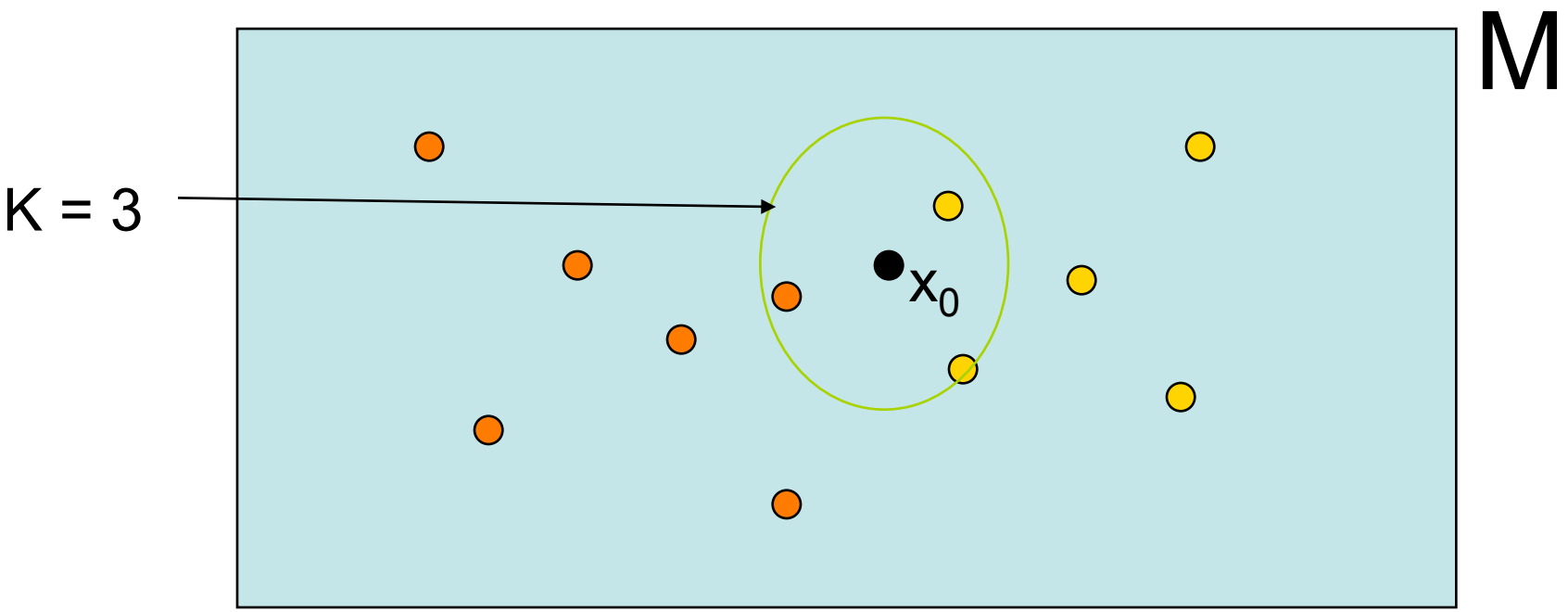
Etiquetas ???

X_1	X_2	…	X_p	G
2	M	…	1.62	?
3	F	…	1.12	?
1	M	…	2.38	?
5	M	…	2.12	?
		…	3.12	…
6	M	…	1.16	?

K Vecinos Mas Cercanos (KNN)

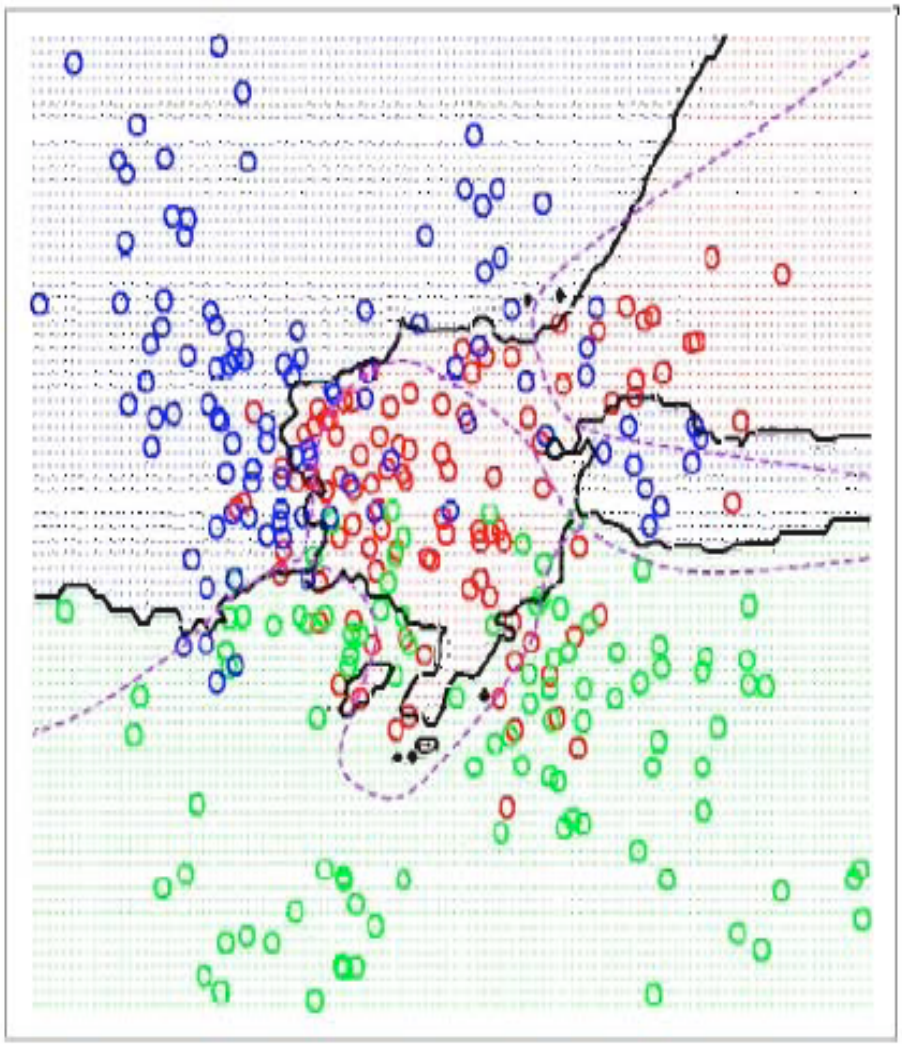
K vecinos mas cercanos (KNN)

- Dada una nueva observación X_0, la clasifico en aquella población que posee una representación mayoritaria entre los K vecinos mas cercanos a X_0.

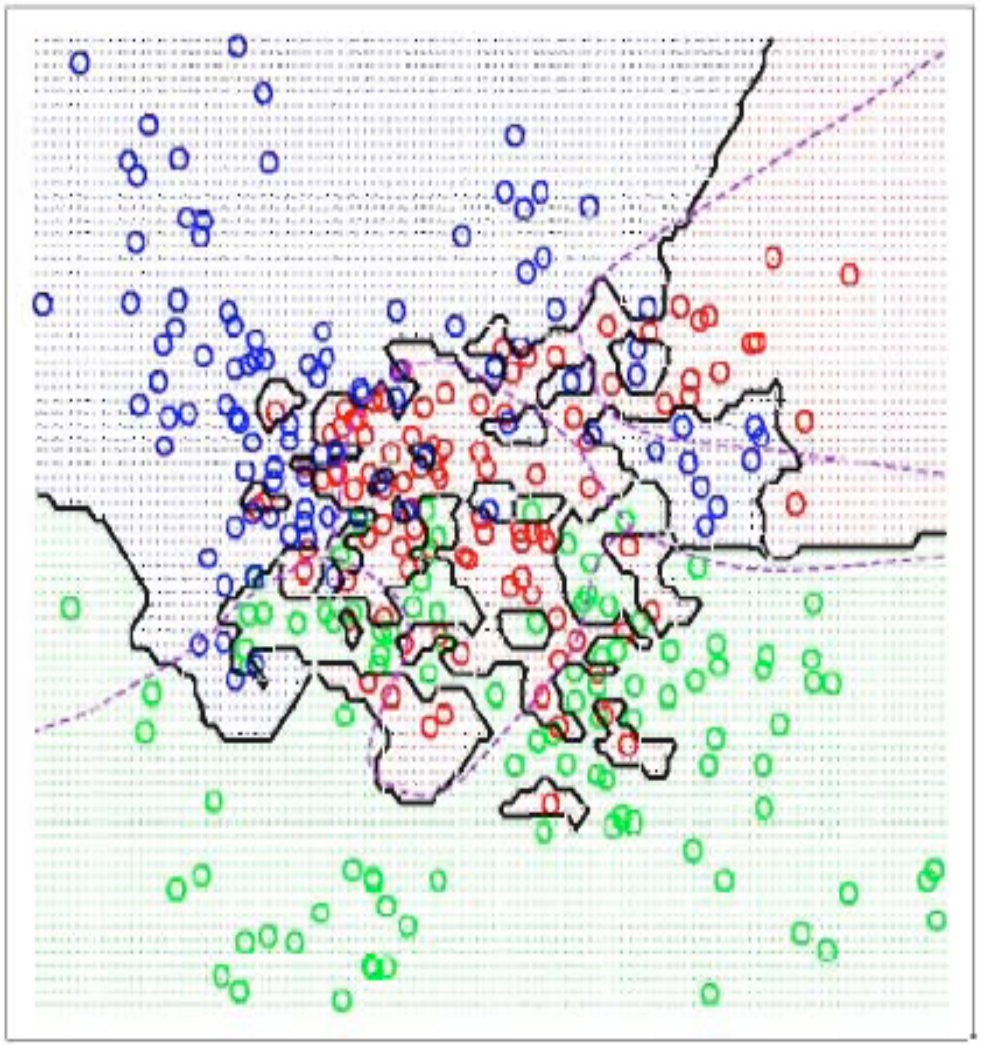


Ejemplo gráfico

15-Nearest Neighbors



1-Nearest Neighbor



Elección de K

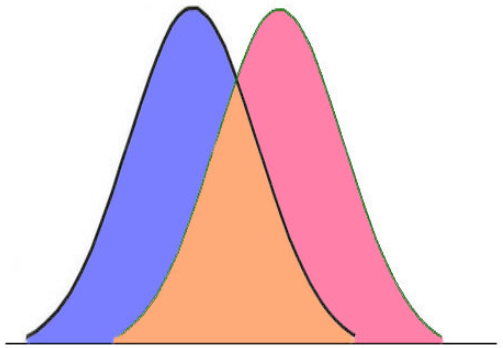
- Muy simple: El K óptimo es aquel que arroja la menor tasa (global) de mala clasificación (error).

Como calculo la tasa (global) de mala clasificación ?

Matriz de Confusión

Clasificación correta

Clasificación erronea



Clasificados

Reales

	P_1	P_2	Total
P_1	$N_{1,1}$	$N_{1,2}$	$N_{1,.}$
P_2	$N_{2,1}$	$N_{2,2}$	$N_{2,.}$
Total	$N_{.,1}$	$N_{.,2}$	$N_{.,.}$

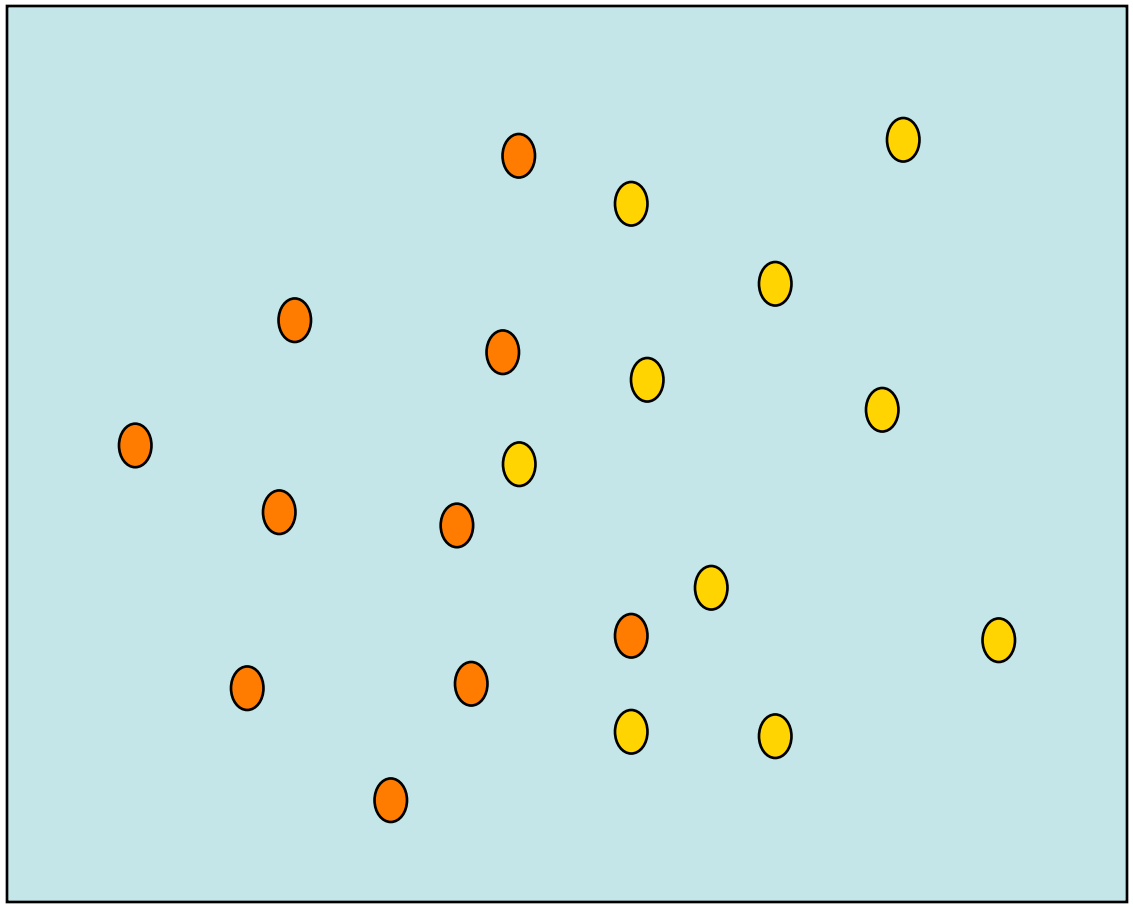
Tasa de error global = $(N_{1,2} + N_{2,1}) / N_{.,.}$

Construcción de la matriz de confusión

- Método ingenuo (naive)
- Partir la muestra (al azar) en dos partes: Muestra de entrenamiento y muestra de validación.
- Por validación cruzada.

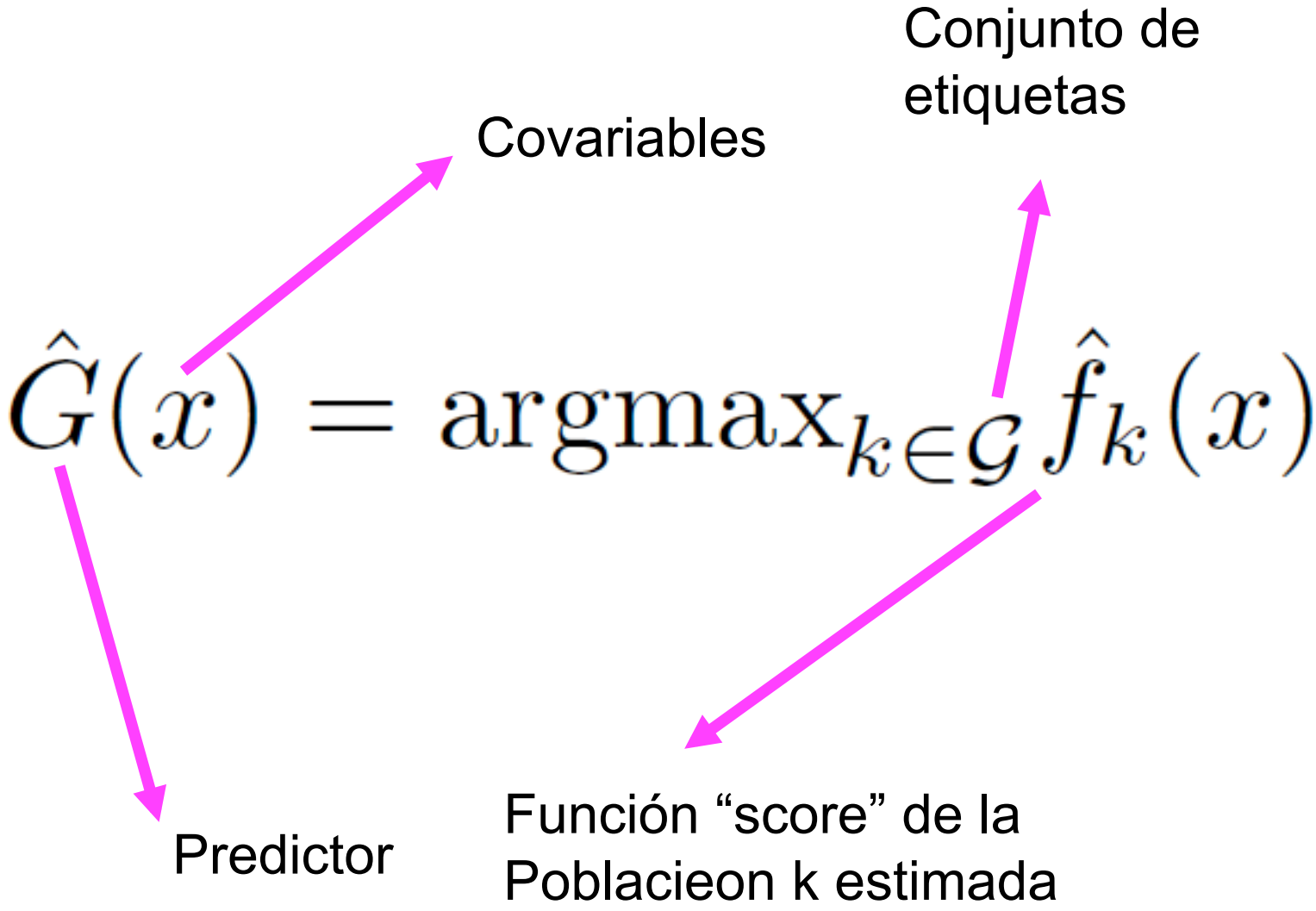
Ejercicio de construcción de matriz de confusión para k = 3

M

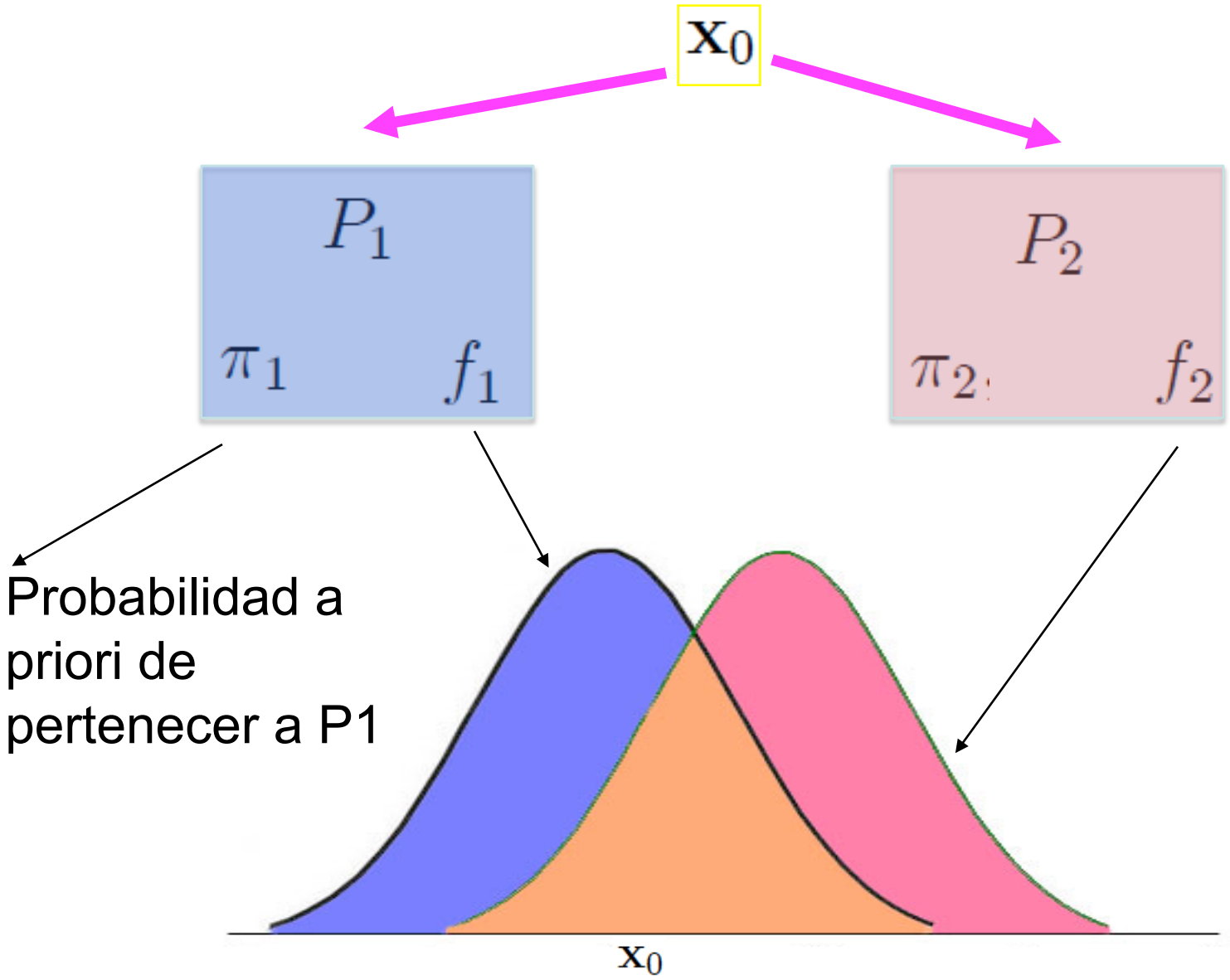


LDA y QDA de Fisher

Idea Fundamental



Definiciones (2 poblaciones)



Principio discriminante (2 poblaciones)

$$P(1/\mathbf{x}_0) = \frac{f_1(\mathbf{x}_0)\pi_1}{f_1(\mathbf{x}_0)\pi_1 + f_2(\mathbf{x}_0)\pi_2}$$

$$P(2|\mathbf{x}_0) = \frac{f_2(\mathbf{x}_0)\pi_2}{f_1(\mathbf{x}_0)\pi_1 + f_2(\mathbf{x}_0)\pi_2}$$

$$\mathbf{x}_0 \in P_2$$

$$P(2|\mathbf{x}_0) > P(1/\mathbf{x}_0) \Longleftrightarrow \pi_2 f_2(\mathbf{x}_0) > \pi_1 f_1(\mathbf{x}_0)$$

$$\mathbf{x}_0 \in P_1$$

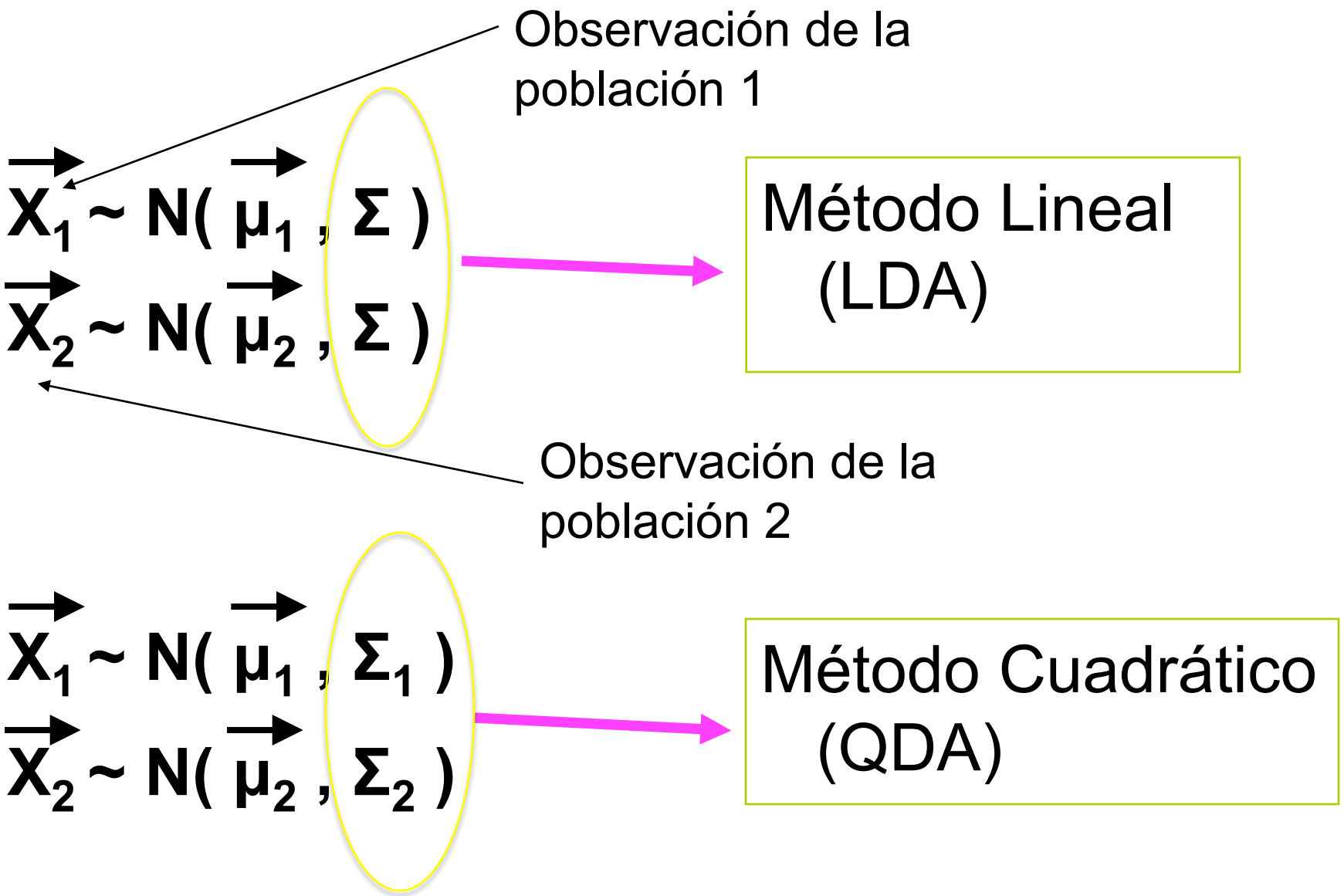
$$P(2|\mathbf{x}_0) < P(1/\mathbf{x}_0) \Longleftrightarrow \pi_2 f_2(\mathbf{x}_0) < \pi_1 f_1(\mathbf{x}_0)$$

Costos de mala clasificación

Clasifico en P_1 si $\dfrac{f_2(\mathbf{x}_0)\pi_2}{c(2|1)} < \dfrac{f_1(\mathbf{x}_0)\pi_1}{c(1|2)}$.

Clasifico en P_2 si $\dfrac{f_2(\mathbf{x}_0)\pi_2}{c(2|1)} > \dfrac{f_1(\mathbf{x}_0)\pi_1}{c(1|2)}$.

Análisis Discriminante de Fisher



Función disciminante lineal

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|V|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}$$

$$S(i) = -\boldsymbol{\mu}_i'\mathbf{V}^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}_i'\mathbf{V}^{-1}\boldsymbol{\mu}_i - \log\frac{\pi_i}{c(i|j)}$$

Clasifico en P1 si $\frac{f_2(\mathbf{x}_0)\pi_2}{c(2|1)} < \frac{f_1(\mathbf{x}_0)\pi_1}{c(1|2)} \Longleftrightarrow$ S(1) < S(2)

Clasifico en P2 si $\frac{f_2(\mathbf{x}_0)\pi_2}{c(2|1)} > \frac{f_1(\mathbf{x}_0)\pi_1}{c(1|2)} \Longleftrightarrow$ S(2) < S(1)

Función disciminante cuadrática

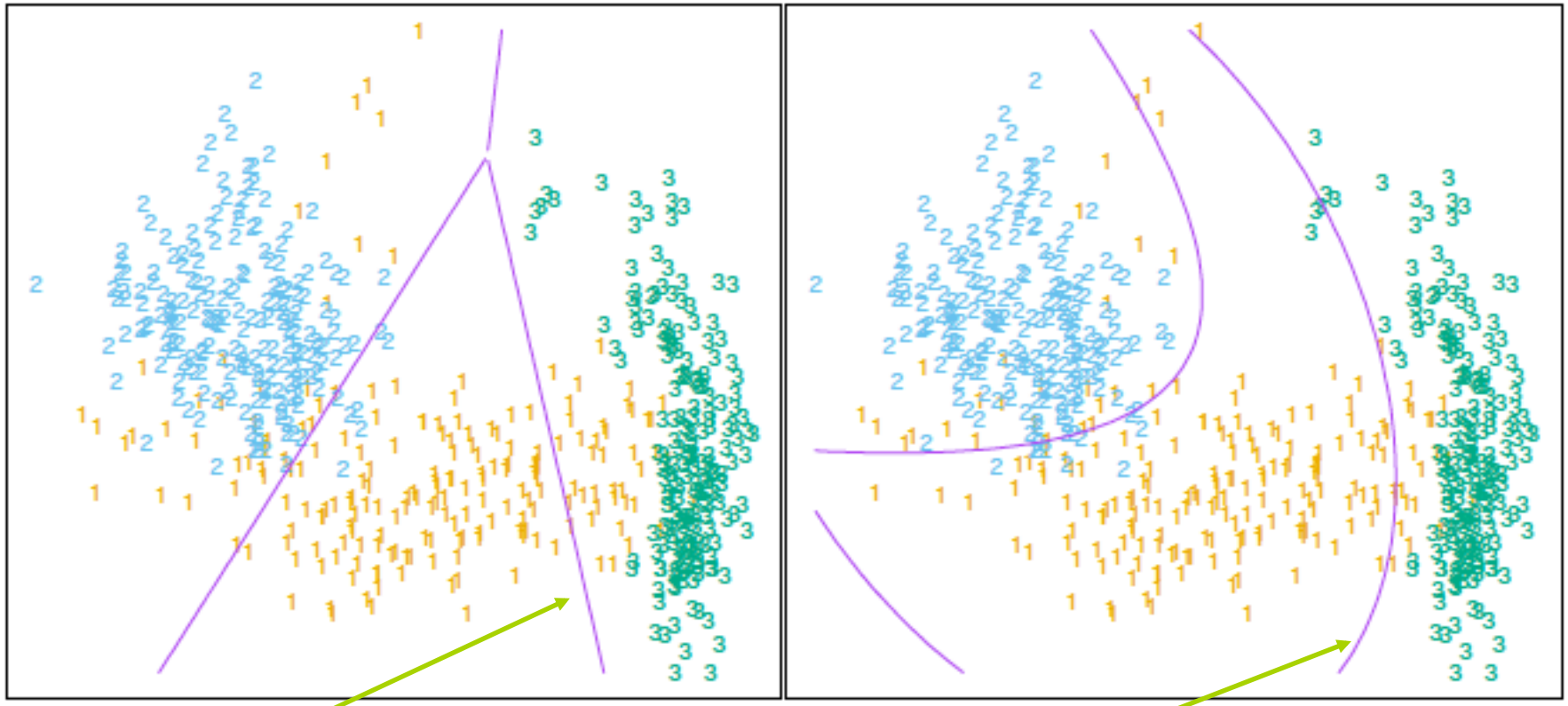
Clasifico en aquella población que satisfaga

$$\min_{j\in(1,\ldots,G)} \left[\frac{1}{2}\log|\mathbf{V}_j| + \frac{1}{2}(\mathbf{x_0}-\boldsymbol{\mu}_j)'\mathbf{V}_j^{-1}(\mathbf{x_0}-\boldsymbol{\mu}_j) - \ln(C_j\pi_j)\right]$$

Varianza generalizada de la población j

Costo de clasidicar mal una observación de j

Lineal Vs. Cuadrático



Fronteras lineales

Fronteras cuadráticas

Recordando la Deviance

- Es una medida de Bondad de Ajuste de un modelo, fijado un conjunto de datos.

$$D(y) = -2[\log\{p(y|\hat{\theta}_0)\} - \log\{p(y|\hat{\theta}_s)\}]$$

Conjunto de observaiones

Verosimilitud

Parámetros estimados

Modelo saturado

Regresión Logística

Regresión Logística

$$y \sim binomial(1, p) \Longrightarrow y \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$$

Link

$$g(E(y)) = Ln\left(\frac{E(y)}{1-E(y)}\right)$$

$$g(p) = Ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$$

$$Var(E(y)) = Var(p) = p(1-p)$$

Predicción

Atributos de la nueva observación

Parámetros estimados

$$\hat{p}(x_1 \cdots x_k) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k}}$$

Probabilidad estimada que la nueva observación sea 1

Ejemplo Regresión Logística:

Prediciendo el Sexo en base al pulso

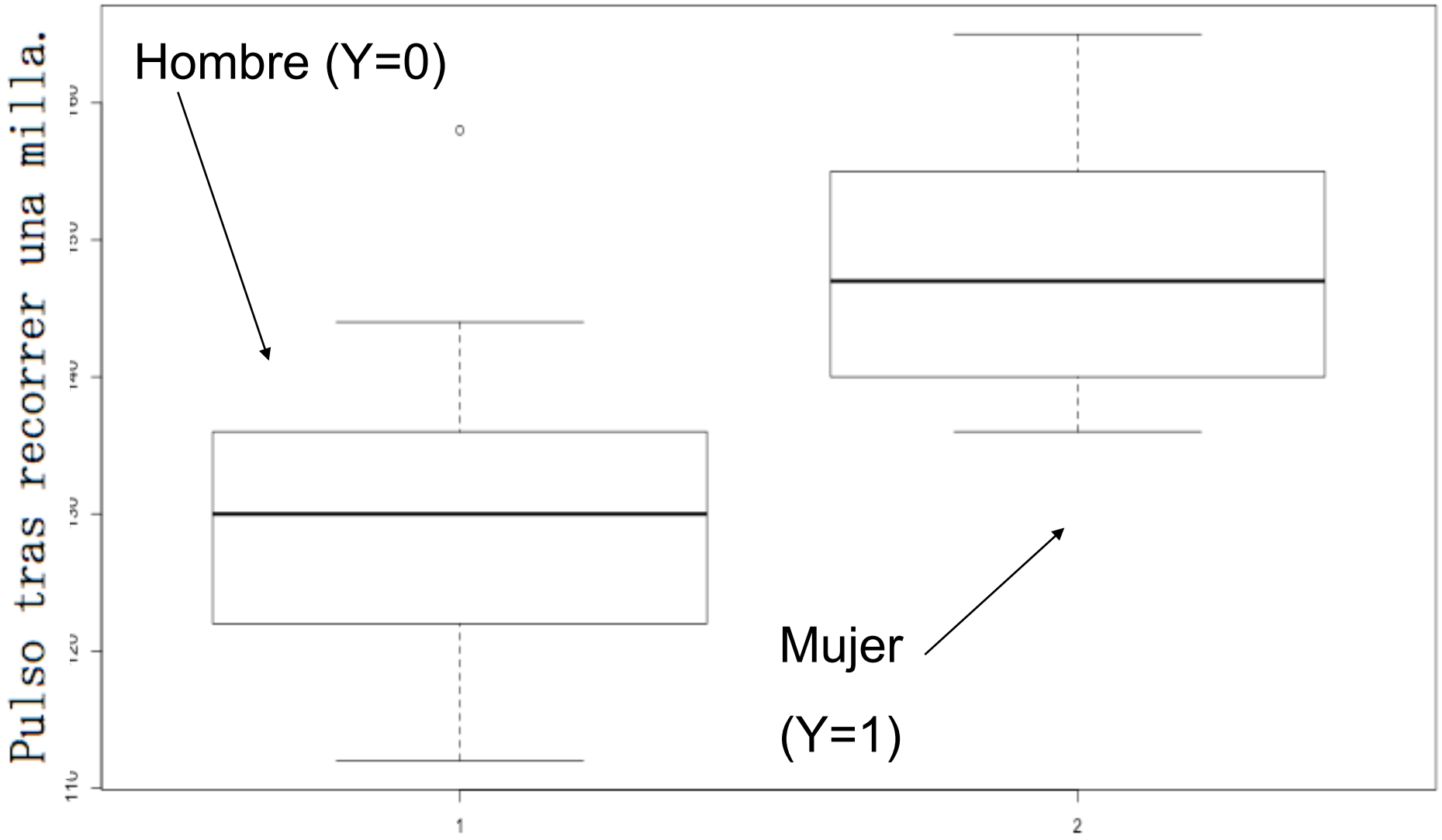
Mujer

Hombre

Sexo	Fumar	Pulso1	Pulso2	Sexo	Fumar	Pulso1	Pulso2
1	1	62	126	1	1	70	122
2	1	78	154	1	2	80	136
1	2	64	128	2	1	76	148
2	2	96	155	2	2	78	148
1	1	66	128	2	2	76	136
2	1	96	165	2	2	80	158
1	2	68	120	1	2	68	116
2	2	72	138	1	2	70	120
2	1	88	160	1	1	68	126
1	1	90	144	1	1	70	144
2	2	82	140	2	2	86	144
1	2	74	134	1	2	72	126
2	1						
2	2						
1	1						
1	2						
1	2	76	158	1	2	74	116
2	2	86	146	1	1	90	138
2	1	88	156	1	2	66	142
1	1	66	132	1	2	70	132

- Pulso 1: pulso en reposo
- Pulso 2: Pulso tras recorrer una milla.

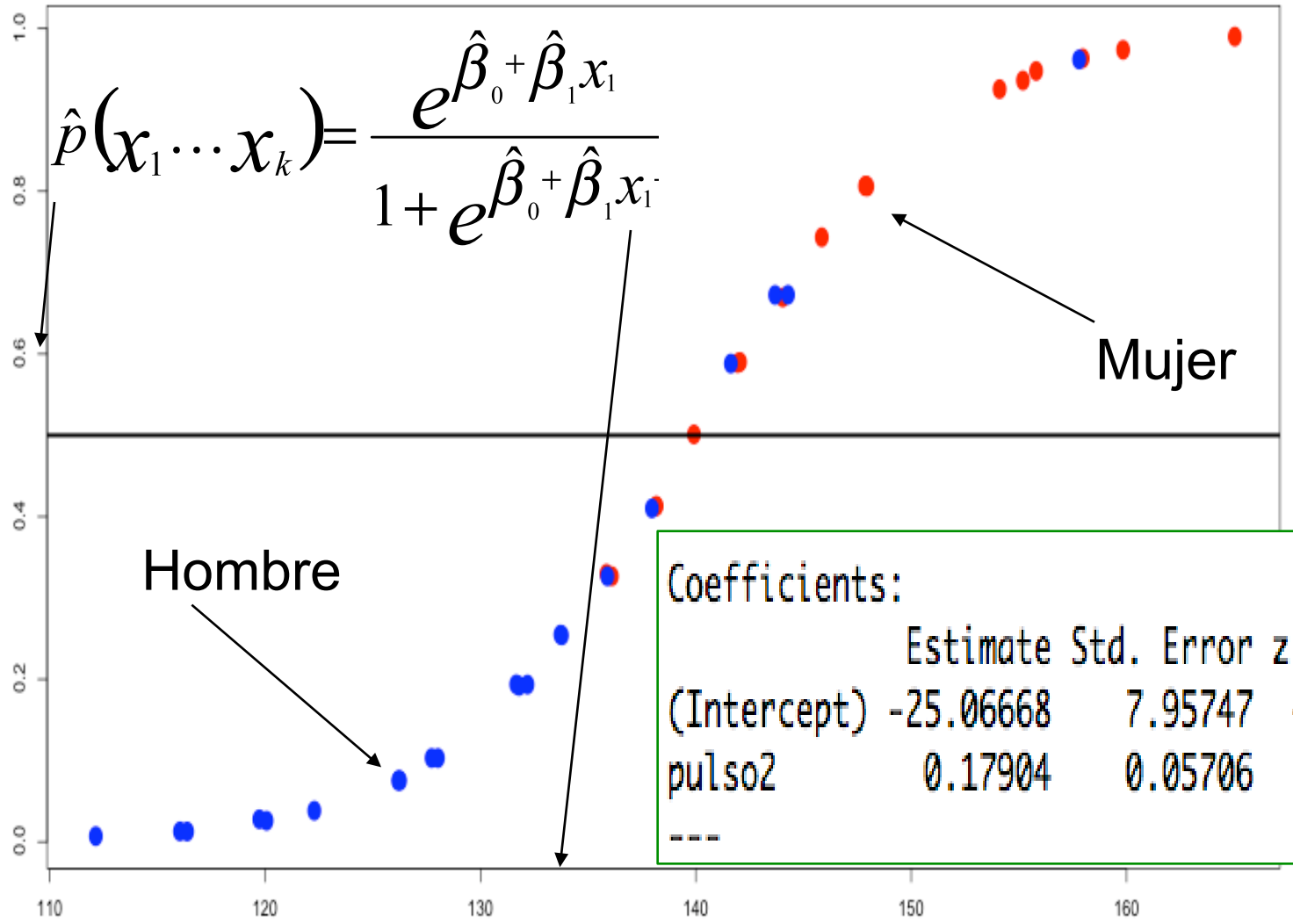
Boxplots



Ejemplo Regresión Logística:

Prediciendo el Sexo en base al pulso

$$\hat{p}(x_1 \cdots x_k) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1}}$$



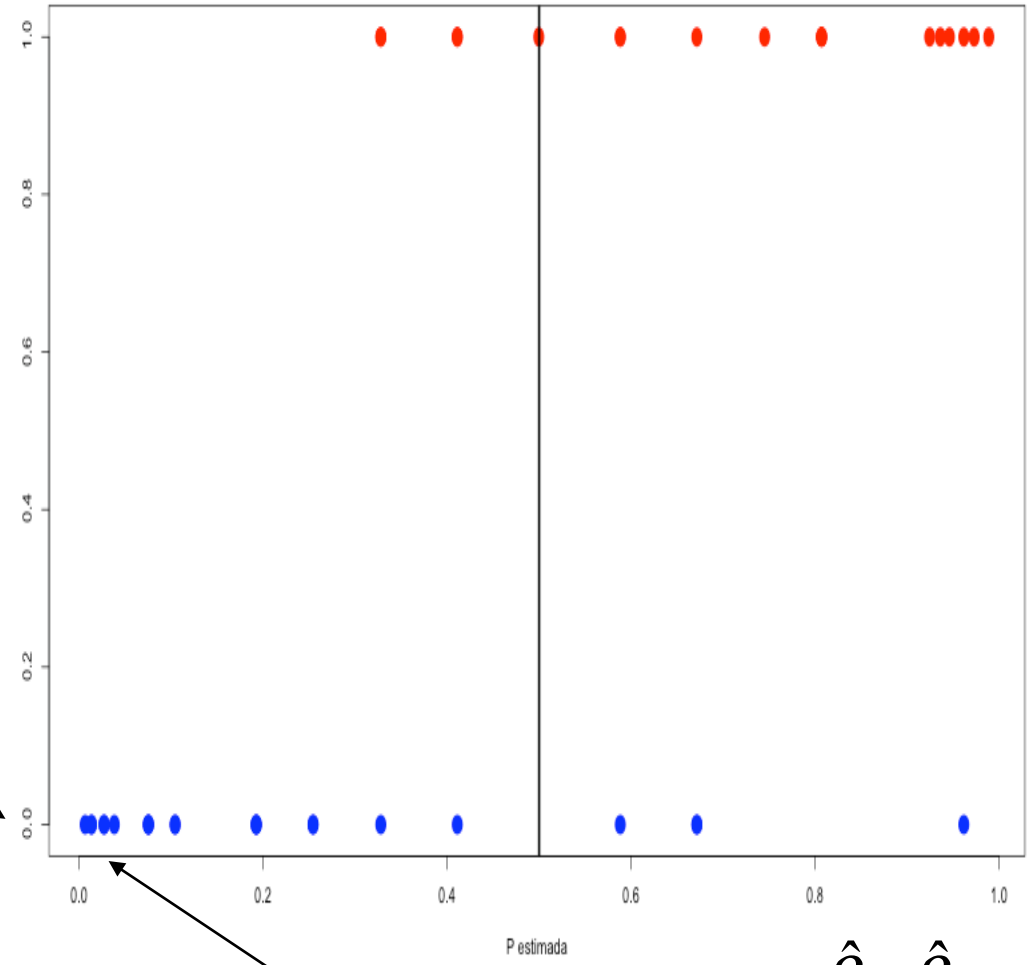

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -25.06668    7.95747  -3.150  0.00163 **
pulso2        0.17904    0.05706   3.138  0.00170 **
---
```

Ejemplo Regresión Logística:

Prediciendo el Sexo en base al pulso

Mujer

Hombre

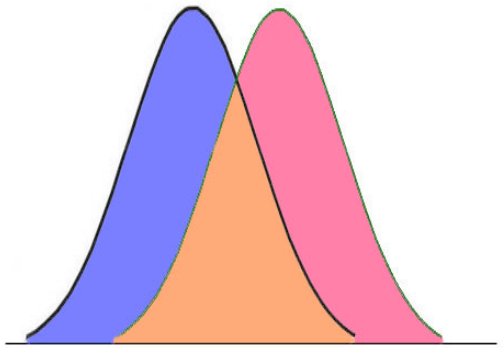


$$\hat{p}(x_1 \cdots x_k) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1}}$$

Matriz de Confusión Ingenua

Clasificación correta

Clasificación erronea



Clasificados

Reales	H	M	Total
H	18	4	22
M	5	13	18
Total	23	17	40

Tasa de error global = (5 + 4) / 40 = 0.225

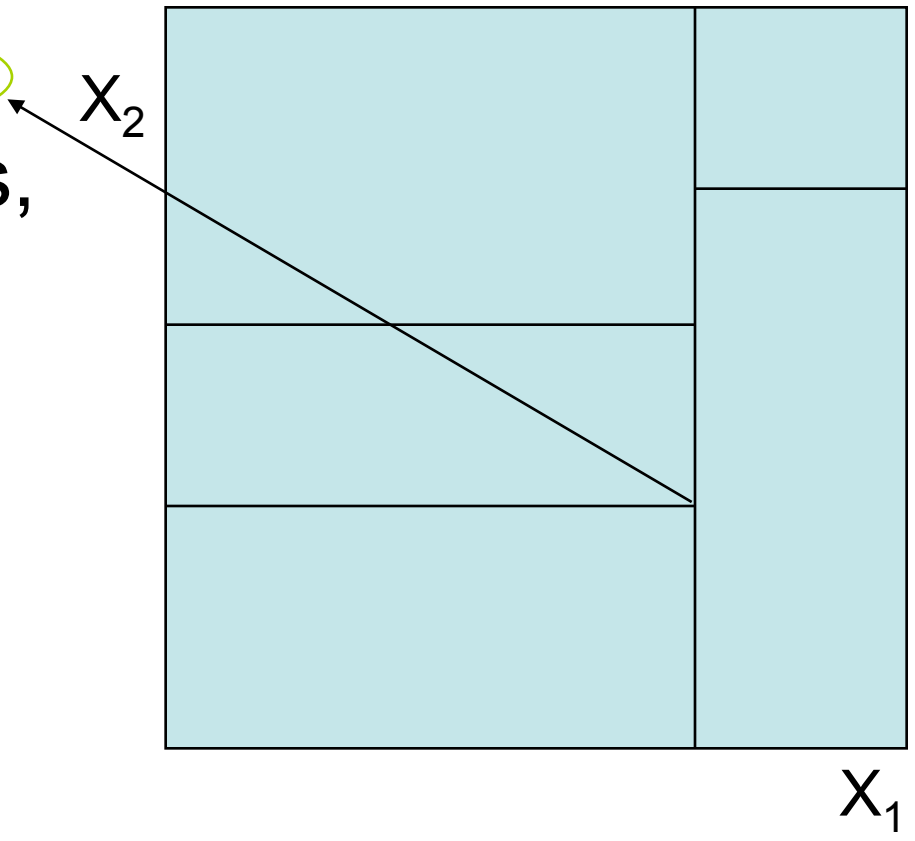
CART

C.A.R.T. (Árboles)

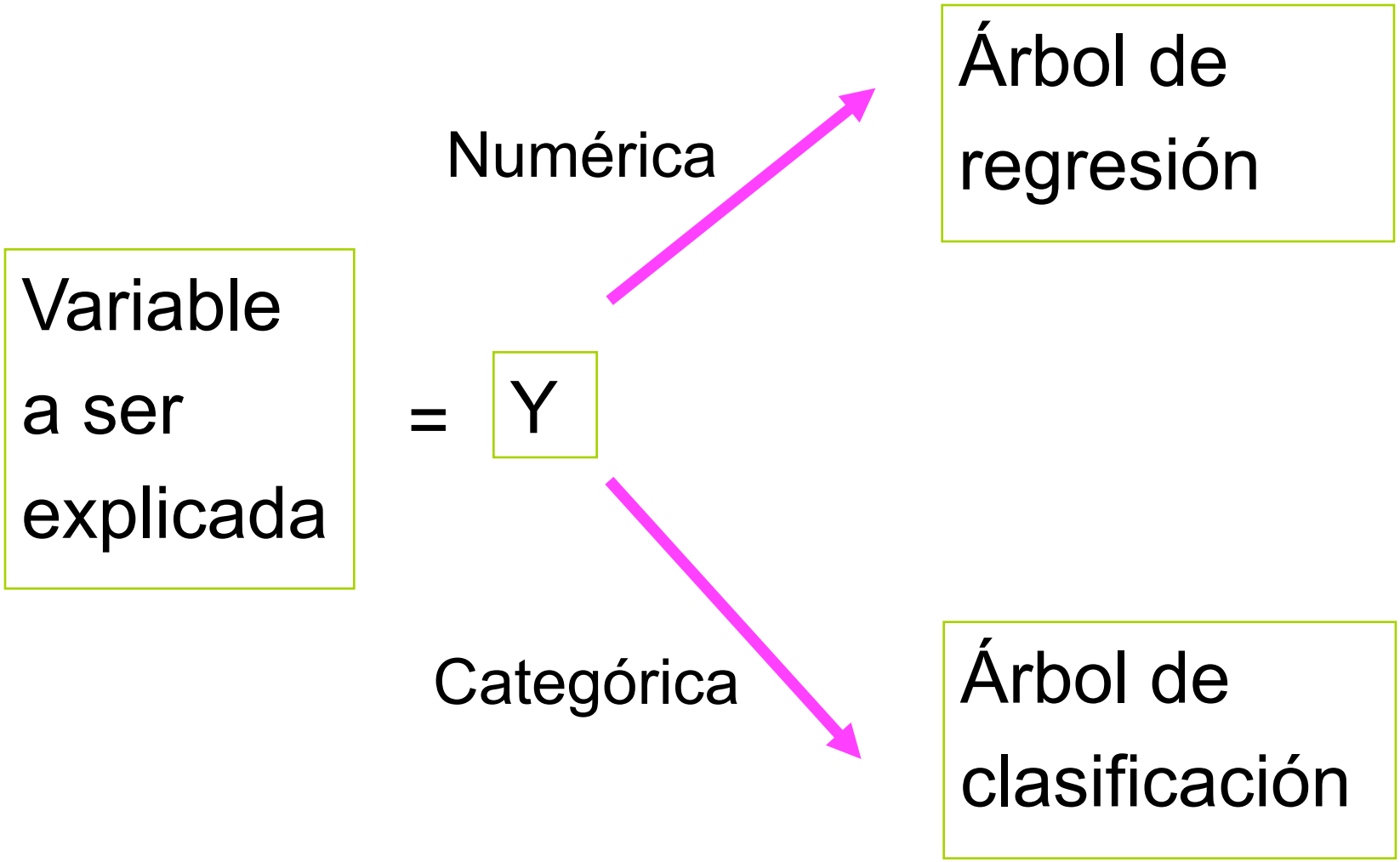
- Es una técnica exploratoria supervisada que permite la detección de estructuras en los datos. Permiten:
  - Clasificar observaciones.
  - Establecer reglas de decisión.
  - Resumir grandes bases de datos
  - Seleccionar variables de interés
  - Detectar interacción entre variables
  - Captar efectos no aditivos

C.A.R.T. (Árboles): Mas especificamente

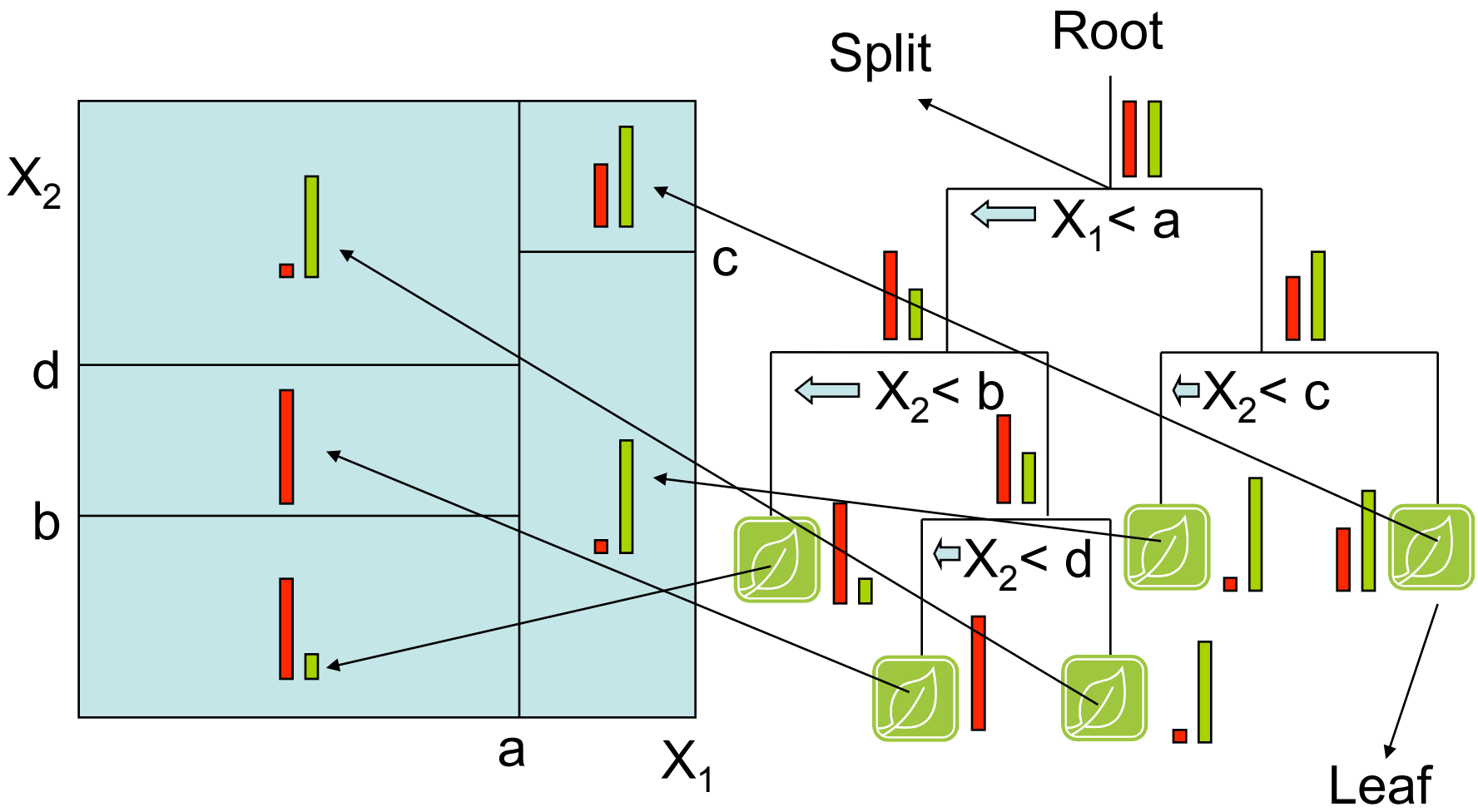
Técnica exploratoria supervisada que busca una partición ortogonal del espacio de atributos, de modo tal que en cada elemento de la partición se “ajuste” adecuadamente a la variable de interes.



Árboles de Regresión y Clasificación



Estructura de un árbol de Clasificación



Medidas de desigualdad de un nodo

Regresión: Varianza (a minimizar)

Nodo

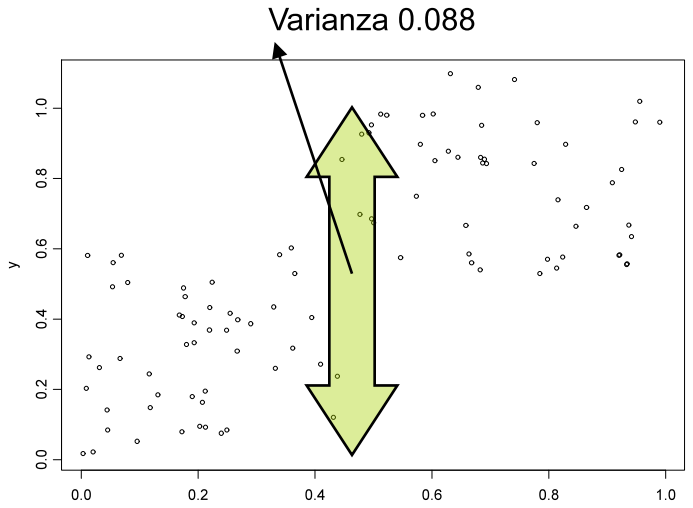
$$Q(L) = \frac{\sum_{x \in L} (x - \bar{x}_L)^2}{n_L}$$

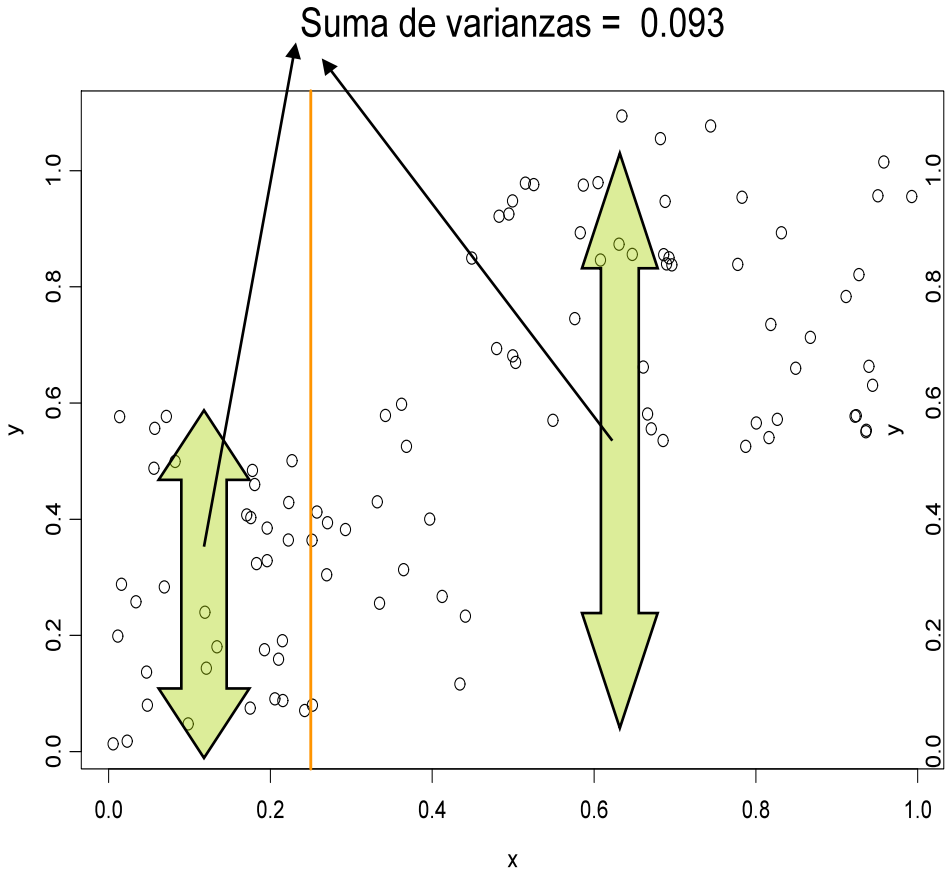
Tamaño del Nodo

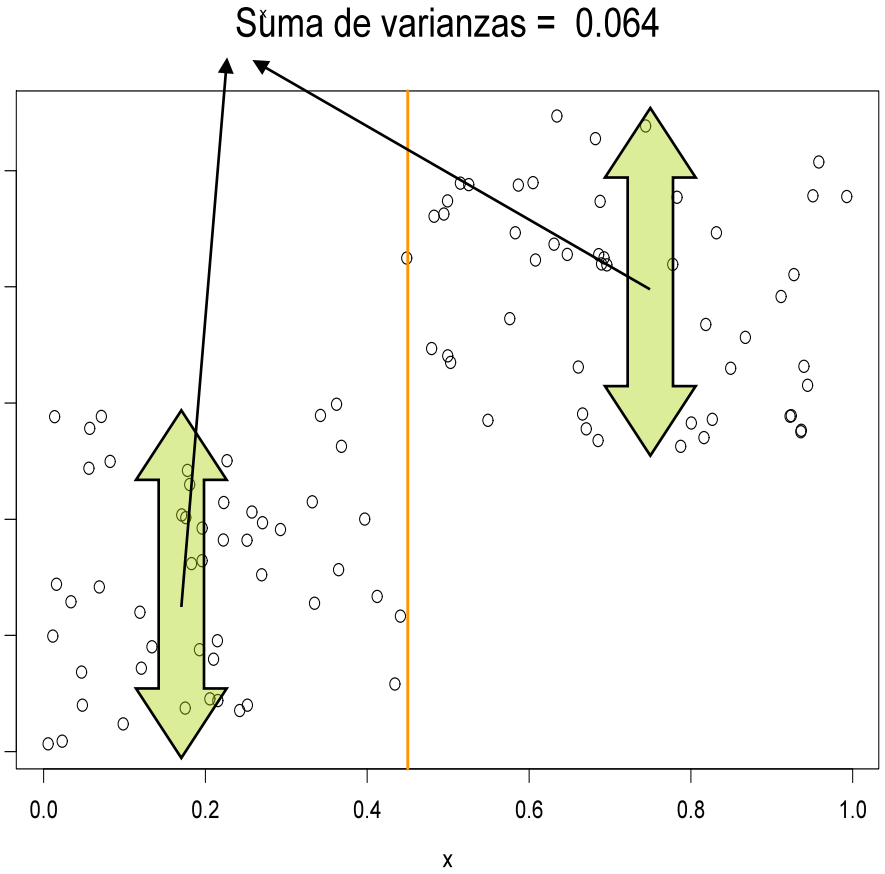
Clasificación: Entropía (a minimizar)

$$Q(L) = -\sum_{k=1}^{K} \hat{p}_k^{n_L}(L) * Ln\left(\hat{p}_k^{n_L}(L)\right)$$

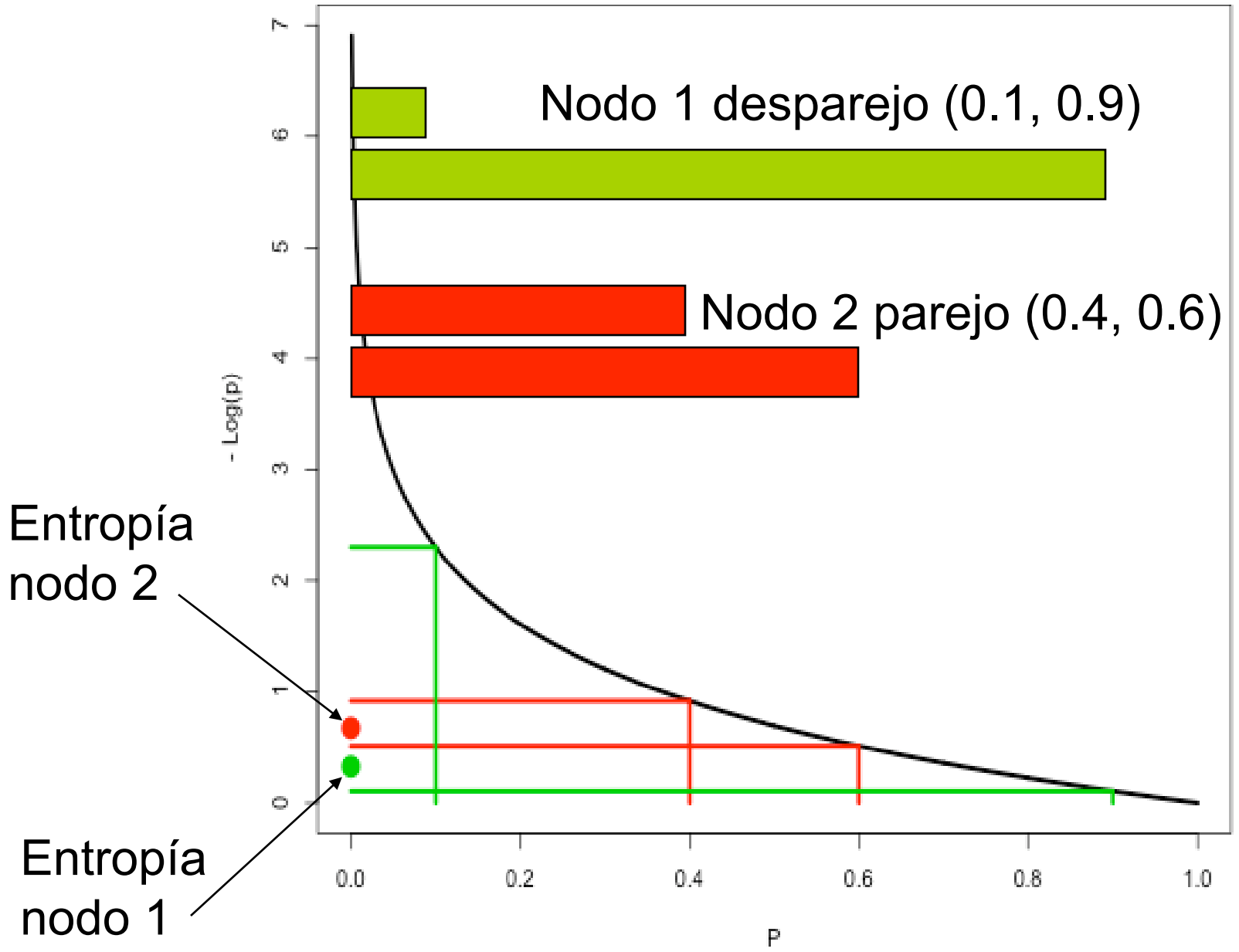
El mejor corte (regresión)





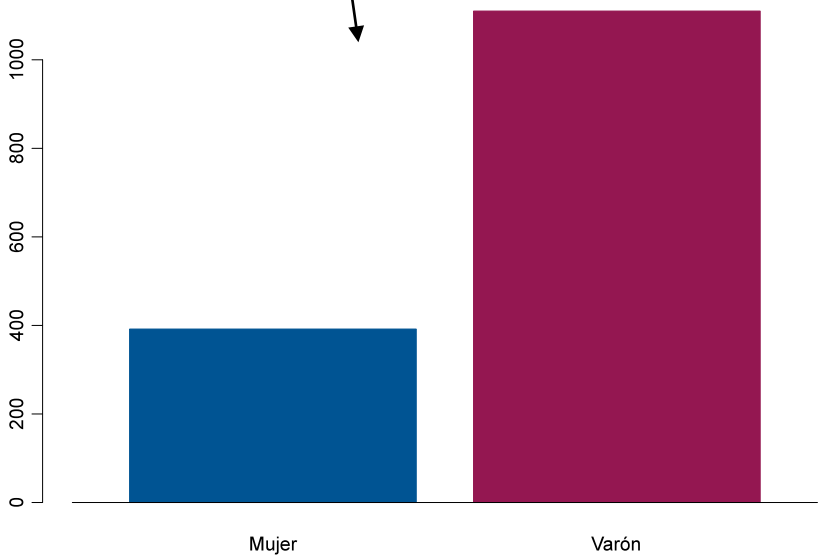


Porque funciona la Entropía



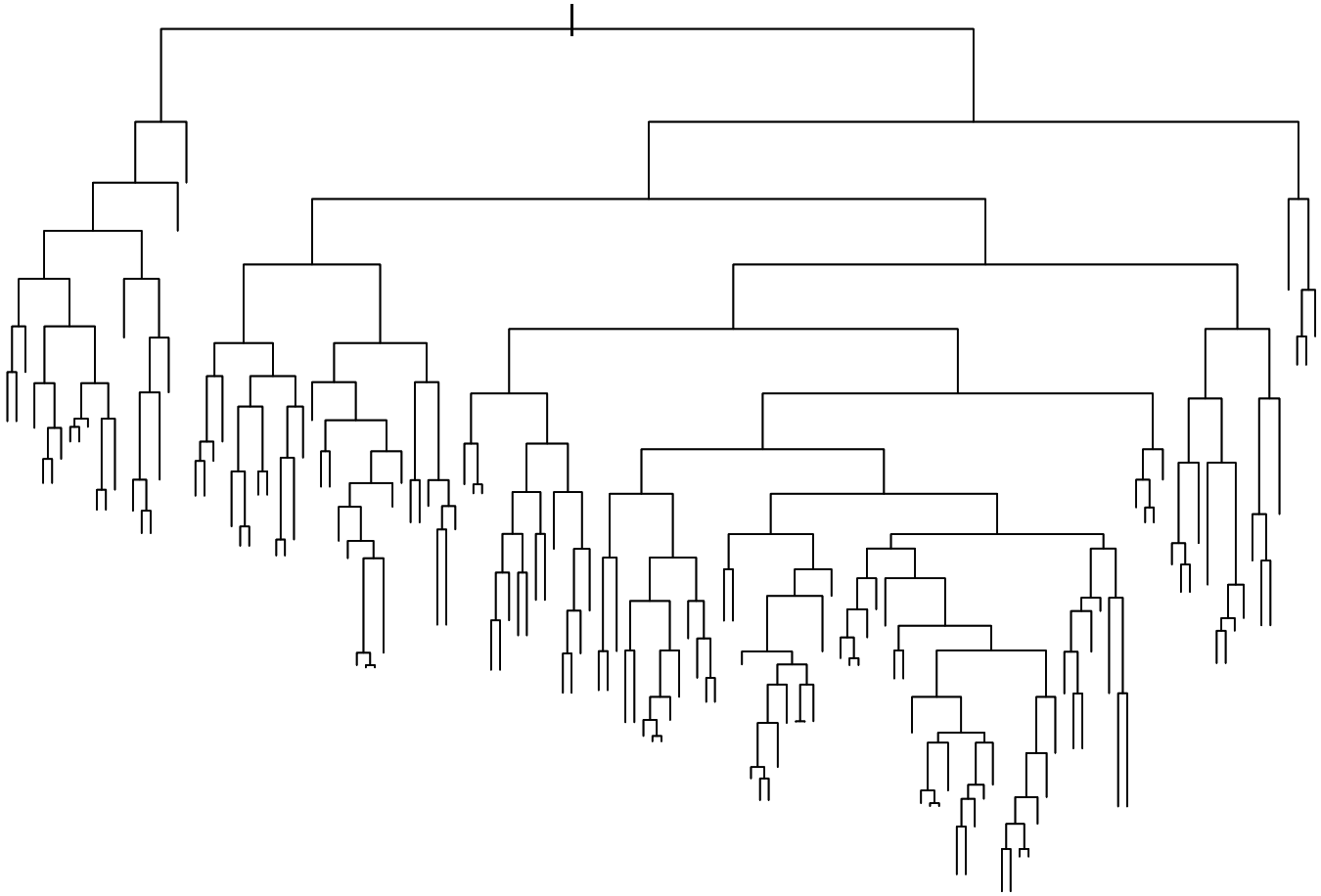
Ejemplo: Prediciendo el Sexo en base a “lo que se tiene en cuenta”

	NO	SI	NA
precio.encuenta	197	831	474
exterior.encuenta	212	947	343
interior.encuenta	295	1108	99
confort.encuenta	196	993	313
seguridad.encuenta	425	967	110
consumo.encuenta	540	910	52
potencia.encuenta	526	891	85
espacio.encuenta	338	957	207
posmanejo.encuenta	519	968	15
relprepro.encuenta	272	1159	71
tiempoentrega.encuenta	426	1037	39
atenconc.encuenta	406	1037	59
promocion.encuenta	1075	414	13
condpagos.encuenta	538	861	103
postventa.encuenta	608	880	14
garantia.encuenta	454	1038	10
costorep.encuenta	773	704	25
costomant.encuenta	590	891	21



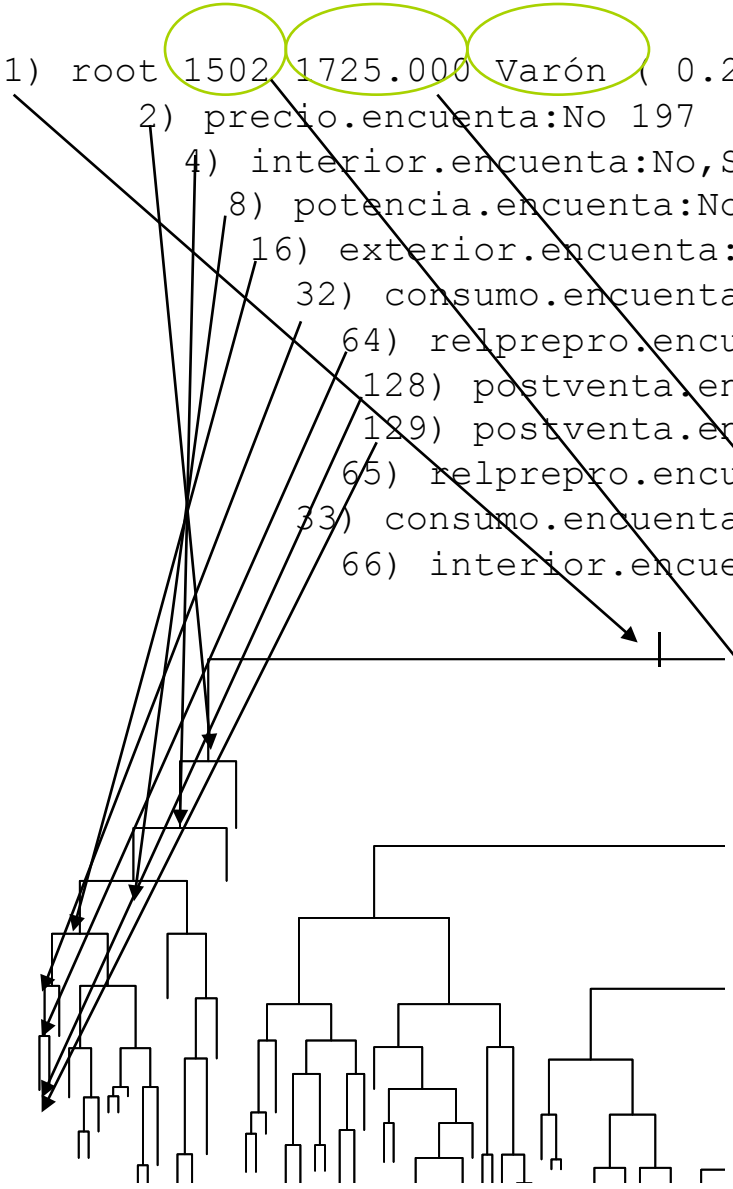
Árbol completo

```
Classification tree:
tree(formula = sexo ~ ., data = ENCUENTA, na.action =
na.tree.replace.all)
Number of terminal nodes:  147
Residual mean deviance:  0.8939 = 1211 / 1355
Misclassification error rate: 0.1964 = 295 / 1502
```



Una rama del árbol

```
1) root 1502 1725.000 Varón ( 0.26100 0.7390 )
     2) precio.encuenta:No 197  187.400 Varón ( 0.18270 0.8173 )
       4) interior.encuenta:No,Sí 185  182.300 Varón ( 0.19460 0.8054 )
         8) potencia.encuenta:No,Sí 176  178.300 Varón ( 0.20450 0.7955 )
          16) exterior.encuenta:No,Sí 131  118.600 Varón ( 0.16790 0.8321 )
            32) consumo.encuenta:Sí 47   27.360 Varón ( 0.08511 0.9149 )
              64) relprepro.encuenta:Sí 30   23.560 Varón ( 0.13330 0.8667 )
               128) postventa.encuenta:No 9   11.460 Varón ( 0.33330 0.6667 ) *
               129) postventa.encuenta:Sí 21    8.041 Varón ( 0.04762 0.9524 ) *
              65) relprepro.encuenta:No,NA 17    0.000 Varón ( 0.00000 1.0000 ) *
            33) consumo.encuenta:No,NA 84   87.290 Varón ( 0.21430 0.7857 )
              66) interior.encuenta:No 37   25.350 Varón ( 0.10810 0.8919 )
```

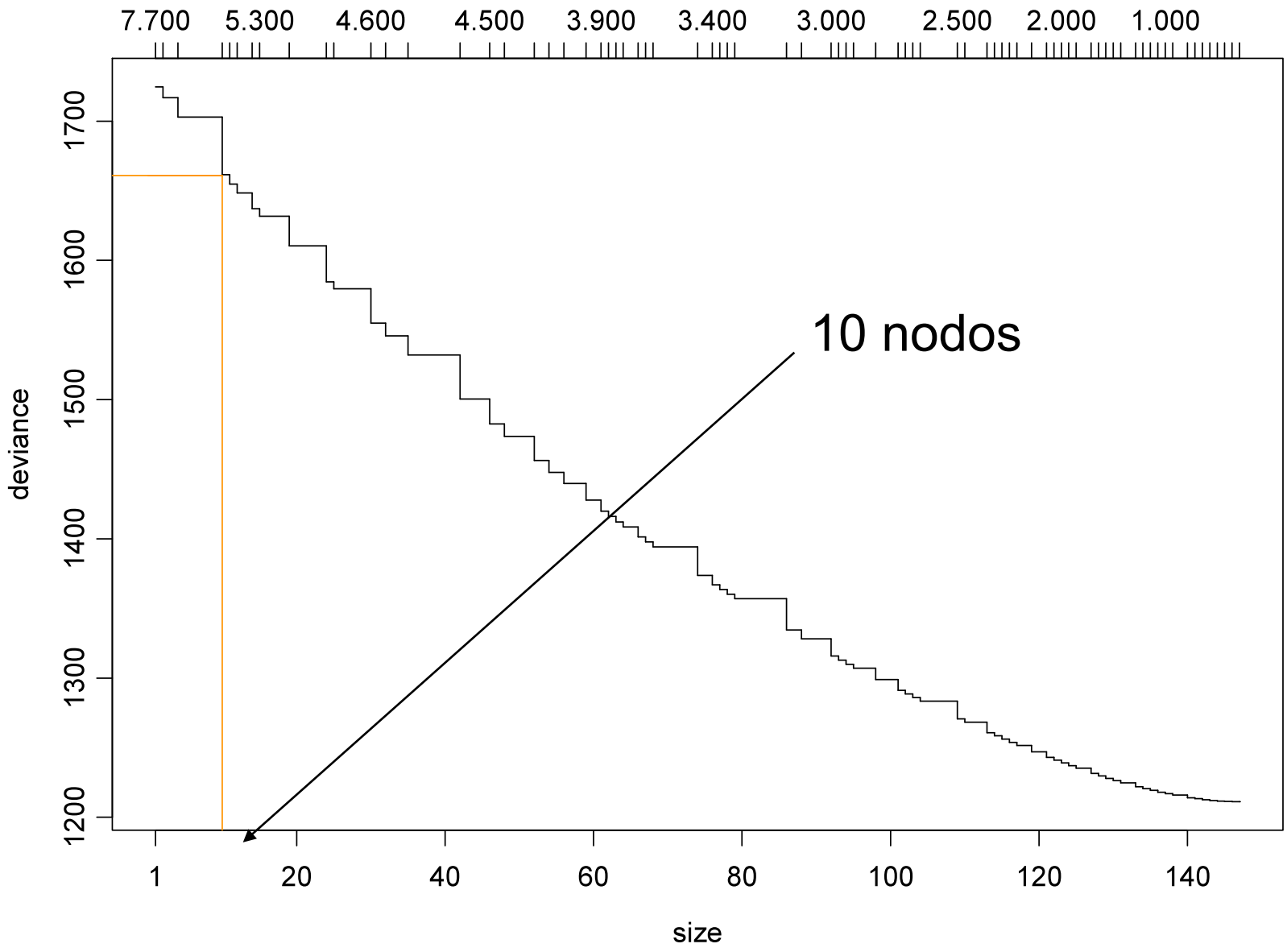


Deviance

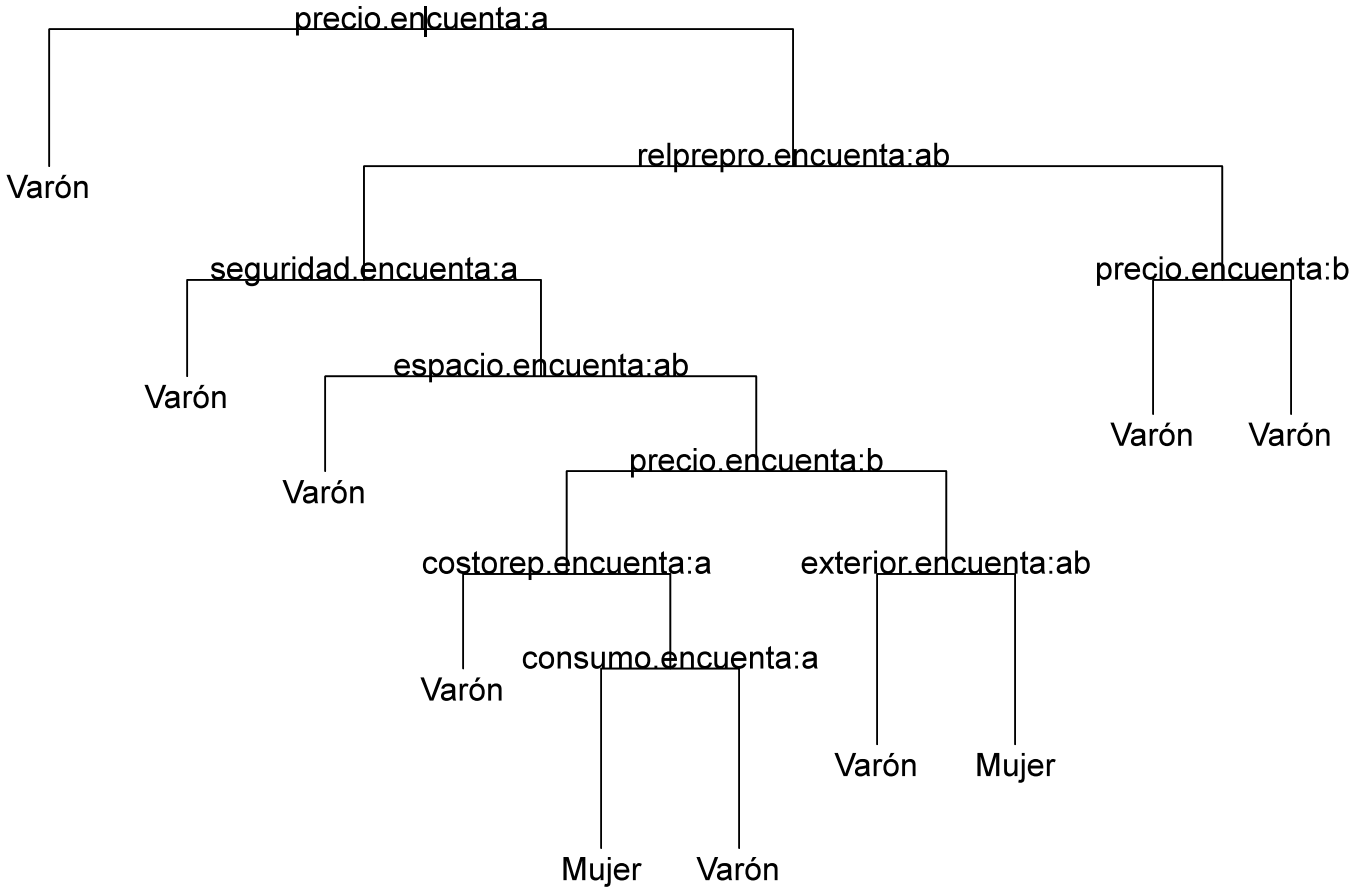
Cantidad de individuos

Nodos terminales

Poda del árbol



Sub-árbol elegido (10 nodos)



Sub-árbol elegido (10 nodos)

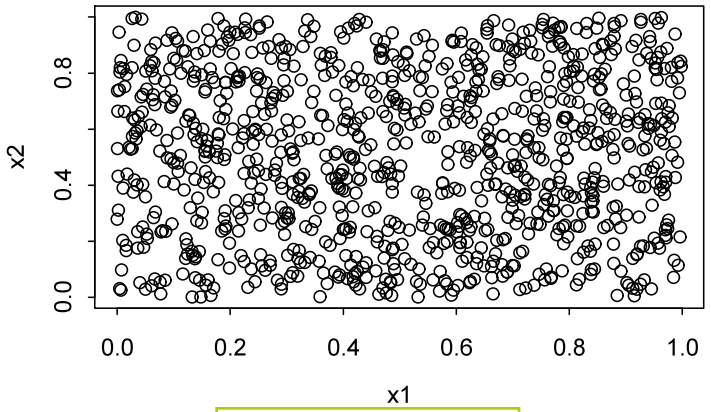
```
1) root 1502 1725.000 Varón ( 0.2610 0.7390 )
   2) precio.encuenta:No 197  187.400 Varón ( 0.1827 0.8173 ) *
   3) precio.encuenta:Sí,NA 1305 1530.000 Varón ( 0.2728 0.7272 )
     6) relprepro.encuenta:No,Sí 1242 1471.000 Varón ( 0.2794 0.7206 )
      12) seguridad.encuenta:No 337  364.700 Varón ( 0.2315 0.7685 ) *
      13) seguridad.encuenta:Sí,NA 905 1101.000 Varón ( 0.2972 0.7028 )
        26) espacio.encuenta:No,Sí 778  926.700 Varón ( 0.2828 0.7172 ) *
        27) espacio.encuenta:NA 127  169.400 Varón ( 0.3858 0.6142 )
          54) precio.encuenta:Sí 93  117.000 Varón ( 0.3226 0.6774 )
           108) costorep.encuenta:No 47   48.650 Varón ( 0.2128 0.7872 ) *
           109) costorep.encuenta:Sí 46   62.980 Varón ( 0.4348 0.5652 )
             218) consumo.encuenta:No 9    6.279 Mujer ( 0.8889 0.1111 ) *
             219) consumo.encuenta:Sí,NA 37   46.630 Varón ( 0.3243 0.6757 ) *
          55) precio.encuenta:NA 34   46.660 Mujer ( 0.5588 0.4412 )
           110) exterior.encuenta:No,Sí 27   37.100 Varón ( 0.4444 0.5556 ) *
           111) exterior.encuenta:NA 7    0.000 Mujer ( 1.0000 0.0000 ) *
     7) relprepro.encuenta:NA 63   51.670 Varón ( 0.1429 0.8571 )
      14) precio.encuenta:Sí 40   15.880 Varón ( 0.0500 0.9500 ) *
      15) precio.encuenta:NA 23   28.270 Varón ( 0.3043 0.6957 ) *
```

Ejemplo de Arbol de Regresión

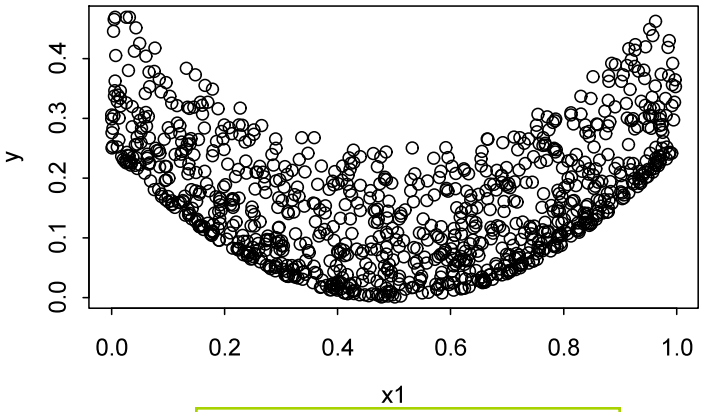
1000 datos, modelo real:

$$Y = (x_1 - 1/2)^2 + (x_2 - 1/2)^2$$

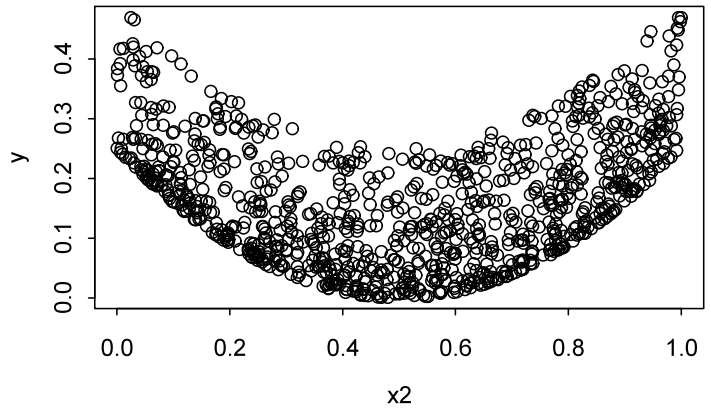
X_2 vs. X_1



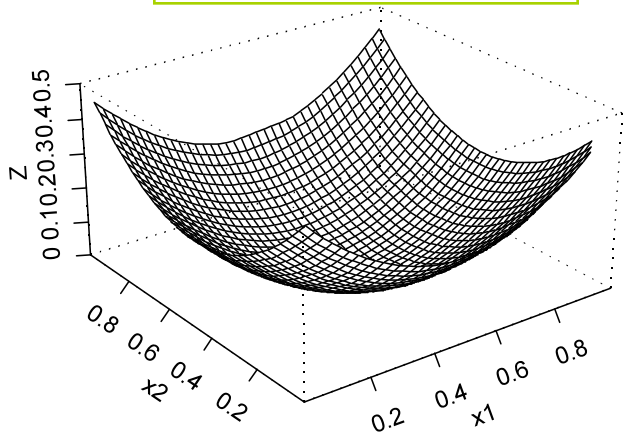
Y vs. X_1



Y vs. X_2



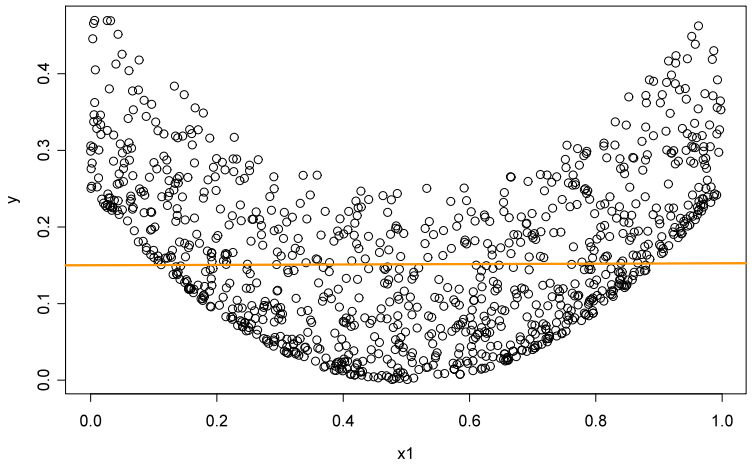
Y vs. X_1 X_2

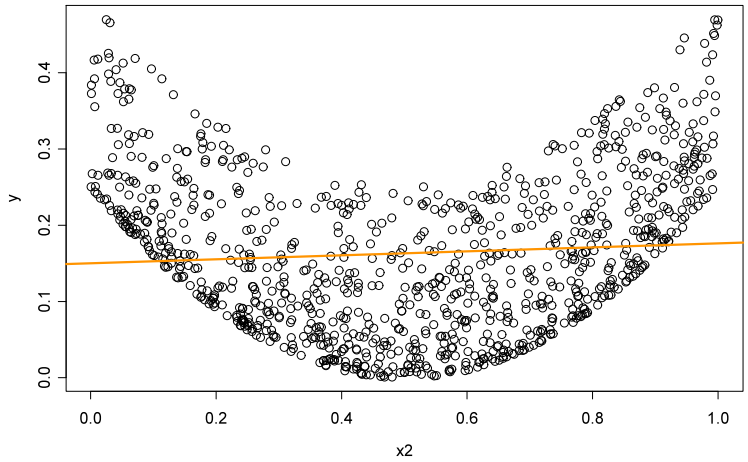


Ajuste lineal (erróneo)

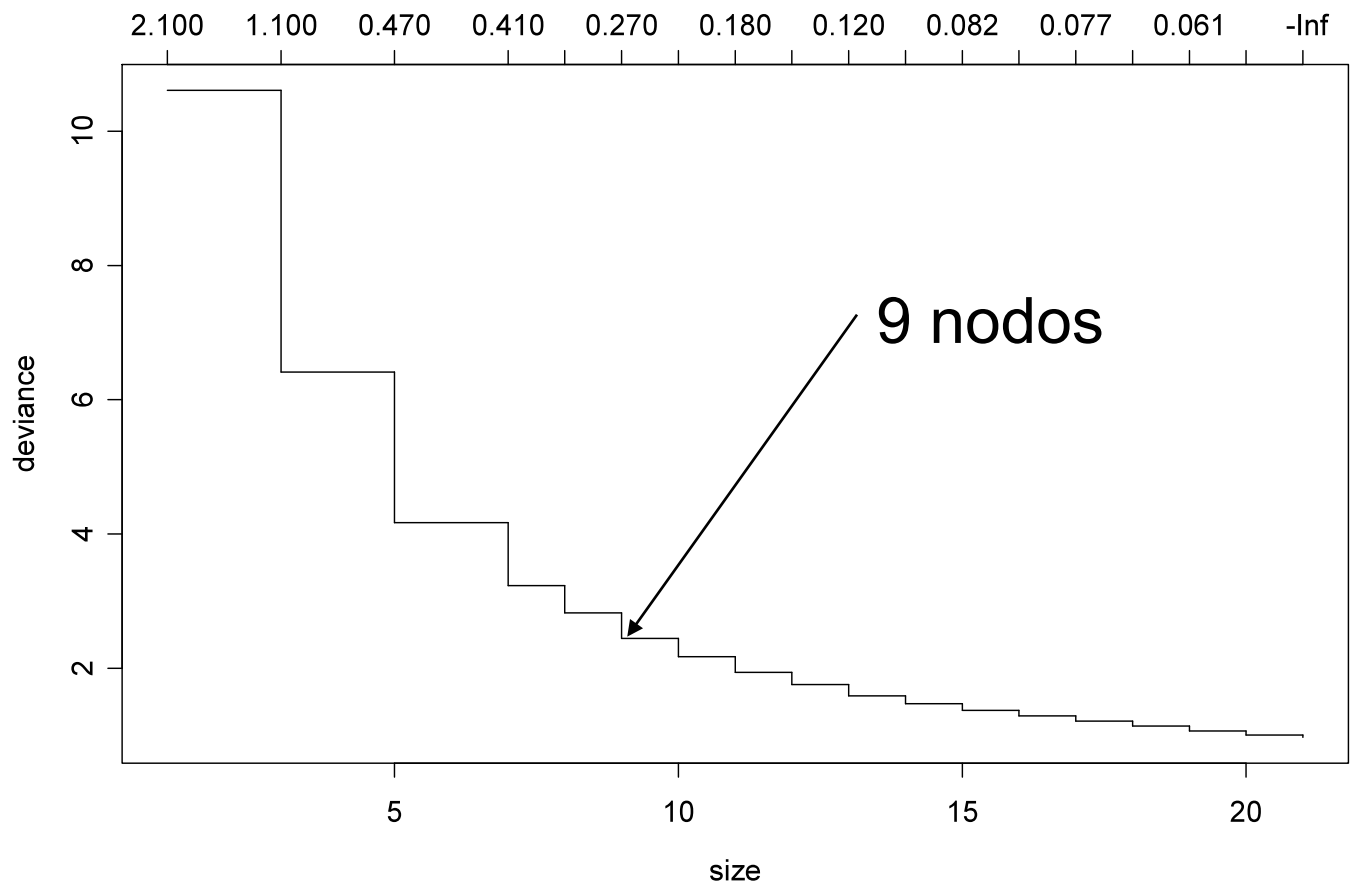
```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept)  0.1501  0.0088    17.0944  0.0000
         x1  0.0024  0.0112     0.2102  0.8336
         x2  0.0261  0.0115     2.2765  0.0230

Residual standard error: 0.1029 on 997 degrees of
   freedom
Multiple R-Squared: 0.005226
F-statistic: 2.619 on 2 and 997 degrees of freedom,the
   p-value is 0.07341
```





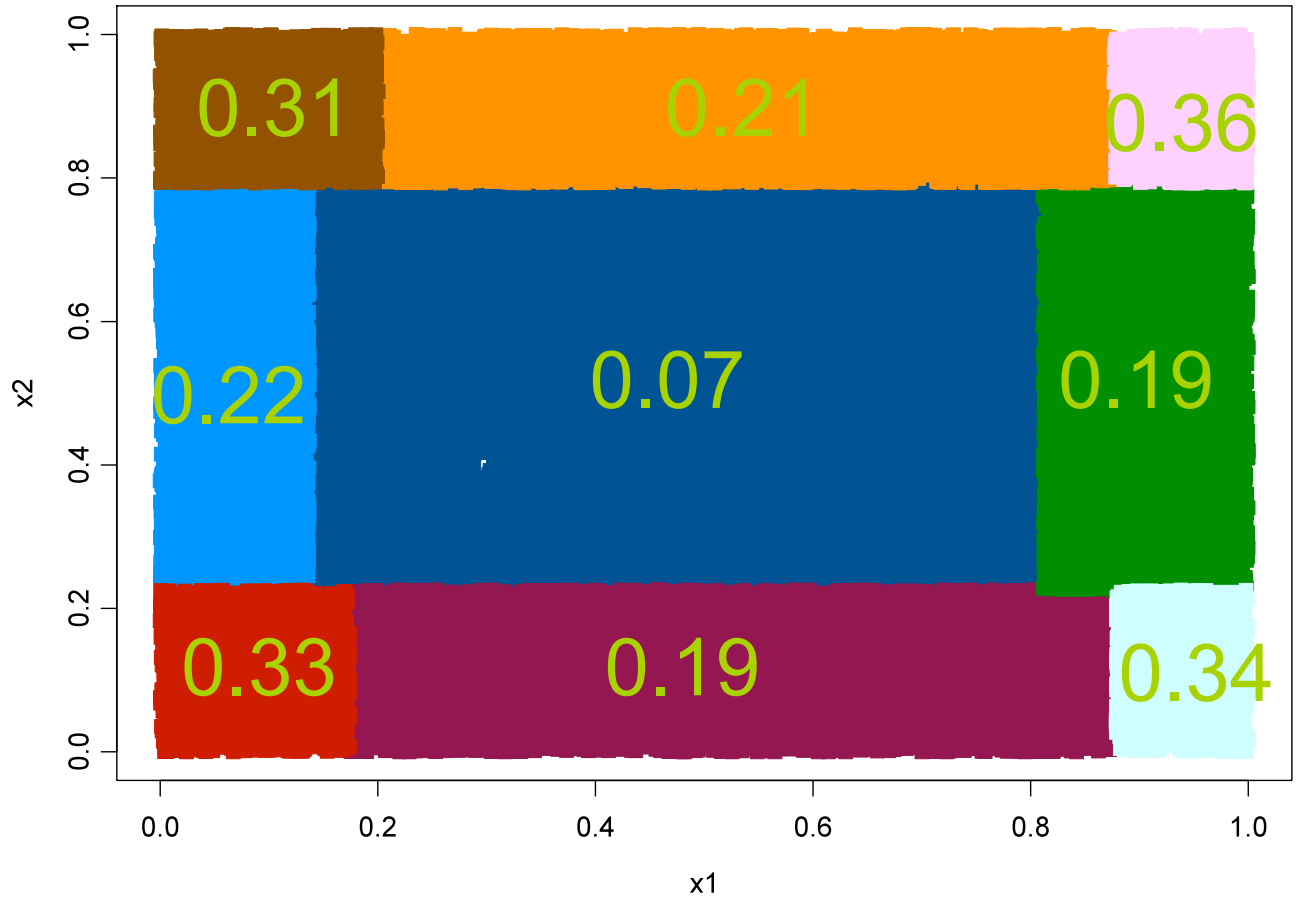
Curva costo-complejidad



Árbol (9 nodos)

```
1) root 1000 10.61000 0.16460
   2) x2<0.79303 790  7.13900 0.14180
     4) x2<0.225821 200  1.52200 0.23320
       8) x1<0.174377 35  0.14700 0.33190 *
       9) x1>0.174377 165  0.96140 0.21220
        18) x1<0.878404 139  0.37830 0.18790 *
        19) x1>0.878404 26  0.06003 0.34240 *
     5) x2>0.225821 590  3.38100 0.11080
      10) x1<0.137237 72  0.12940 0.21920 *
      11) x1>0.137237 518  2.28900 0.09576
        22) x1<0.811252 405  0.73320 0.06950 *
        23) x1>0.811252 113  0.27510 0.18990 *
   3) x2>0.79303 210  1.50900 0.25050
     6) x1<0.876669 181  1.04000 0.23280
      12) x1<0.200377 44  0.16540 0.31360 *
      13) x1>0.200377 137  0.49450 0.20690 *
     7) x1>0.876669 29  0.06168 0.36050 *
```

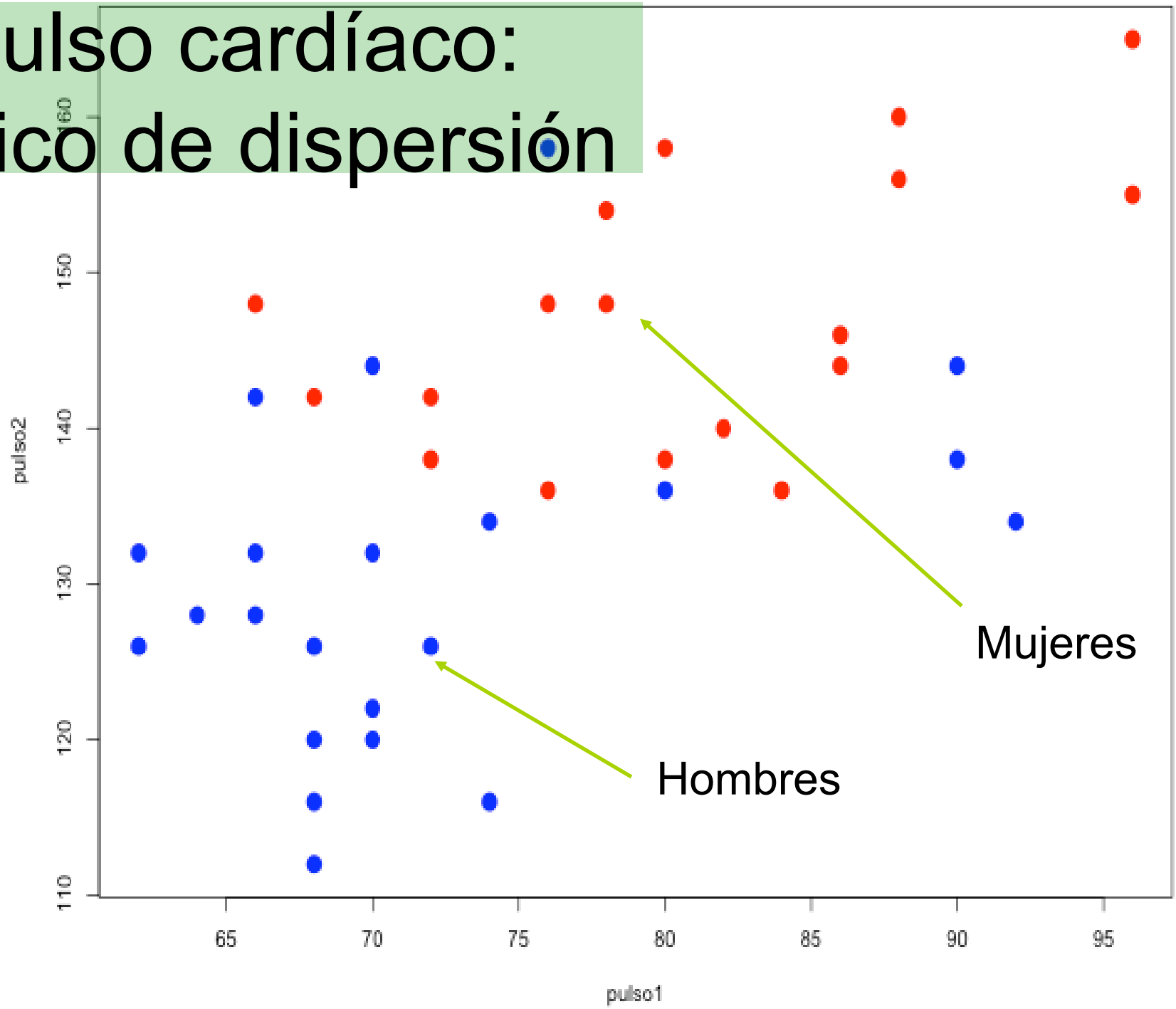
Resultado del ajuste: Árbol de 9 nodos



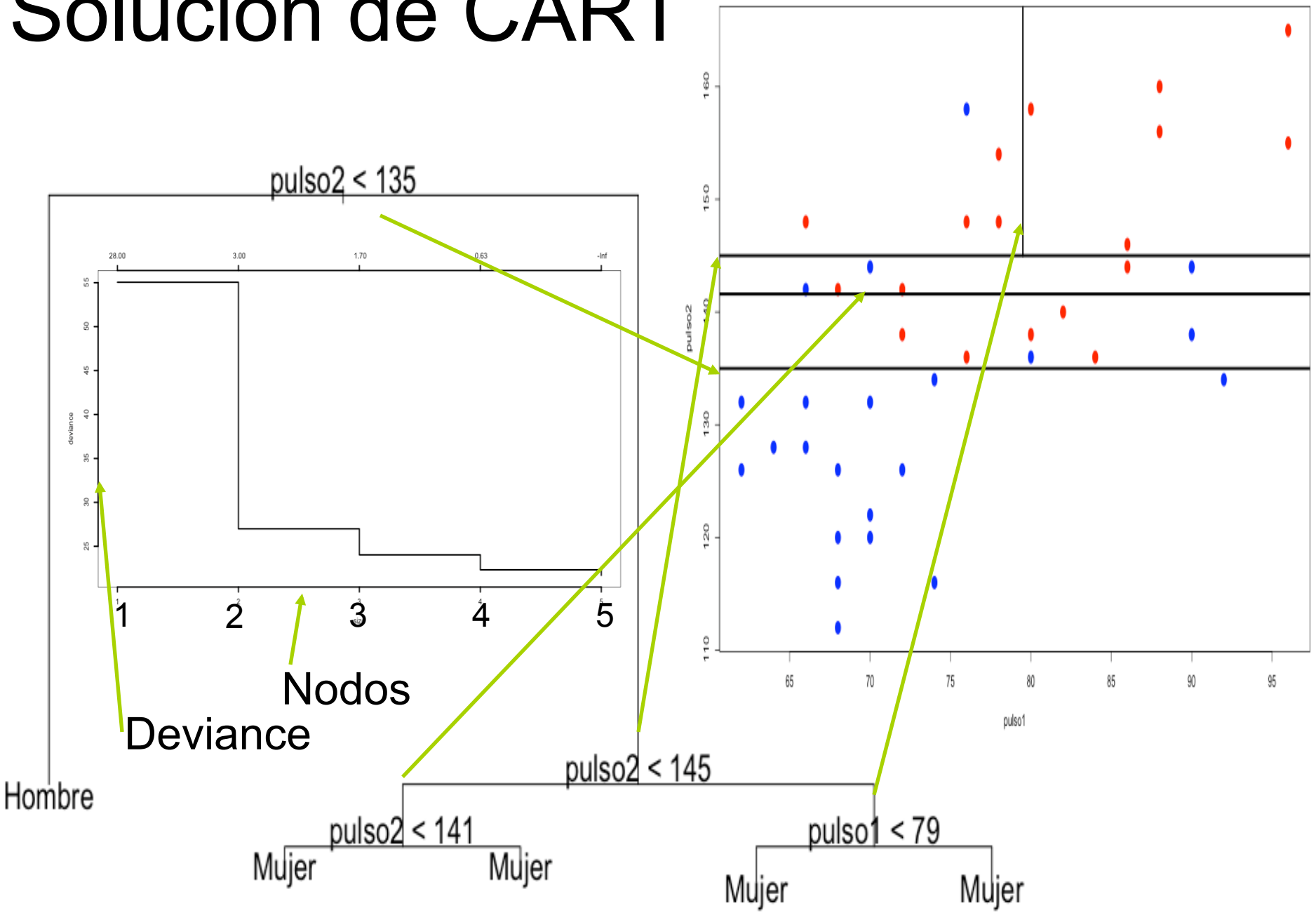
Ejemplo 1: Pulso cardíaco

Sexo	Fumar	Pulso1	Pulso2	Sexo	Fumar	Pulso1	Pulso2
1	1	62	126	1	1	70	122
2	1	78	154	1	2	80	136
1	2	64	128	2	1	76	148
2	2	96	155	2	2	78	148
1	1	66	128	2	2	76	136
2	1	96	165	2	2	80	158
1	2	68	120	1	2	68	116
2	2	72	138	1	2	70	120
2	1	88	160	1	1	68	126
1	1	90	144	1	1	70	144
2	2	82	140	2	2	86	144
1	2	74	134	1	2	72	126
2	1	66	148	2	2	84	136
2	2	68	142	2	2	72	142
1	1	92	134	2	2	80	138
1	2	68	112	1	1	62	132
1	2	76	158	1	2	74	116
2	2	86	146	1	1	90	138
2	1	88	156	1	2	66	142
1	1	66	132	1	2	70	132

Pulso cardíaco: gráfico de dispersión

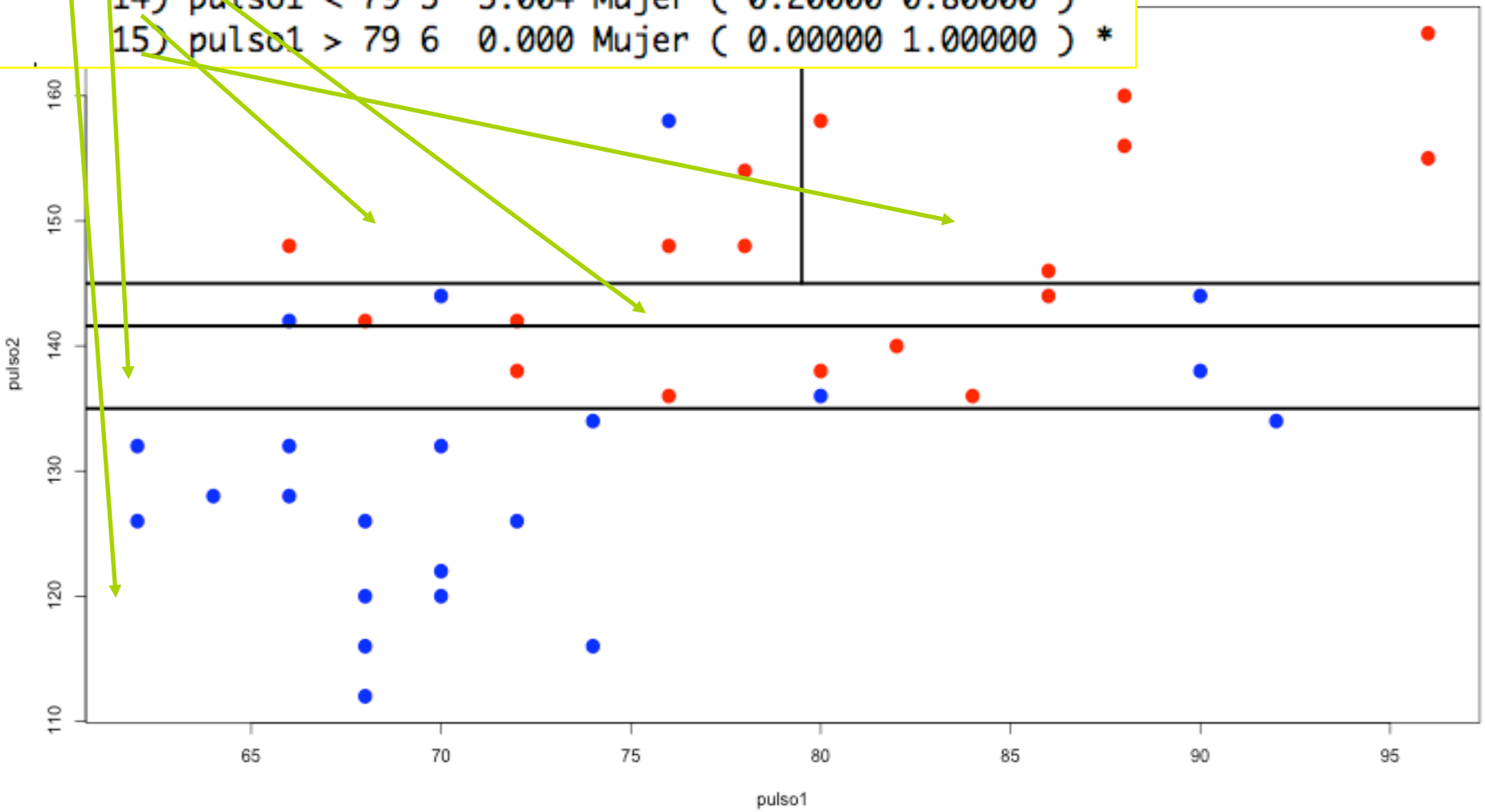


Solución de CART



```
1) root 40 55.050 Hombre ( 0.55000 0.45000 )
  2) pulso2 < 135 16  0.000 Hombre ( 1.00000 0.00000 ) *
  3) pulso2 > 135 24 26.990 Mujer ( 0.25000 0.75000 )
    6) pulso2 < 145 13 17.320 Mujer ( 0.38462 0.61538 )
     12) pulso2 < 141 7  8.376 Mujer ( 0.28571 0.71429 ) *
     13) pulso2 > 141 6  8.318 Mujer ( 0.50000 0.50000 ) *
    7) pulso2 > 145 11  6.702 Mujer ( 0.09091 0.90909 )
     14) pulso1 < 79 5  5.004 Mujer ( 0.20000 0.80000 ) *
     15) pulso1 > 79 6  0.000 Mujer ( 0.00000 1.00000 ) *
```

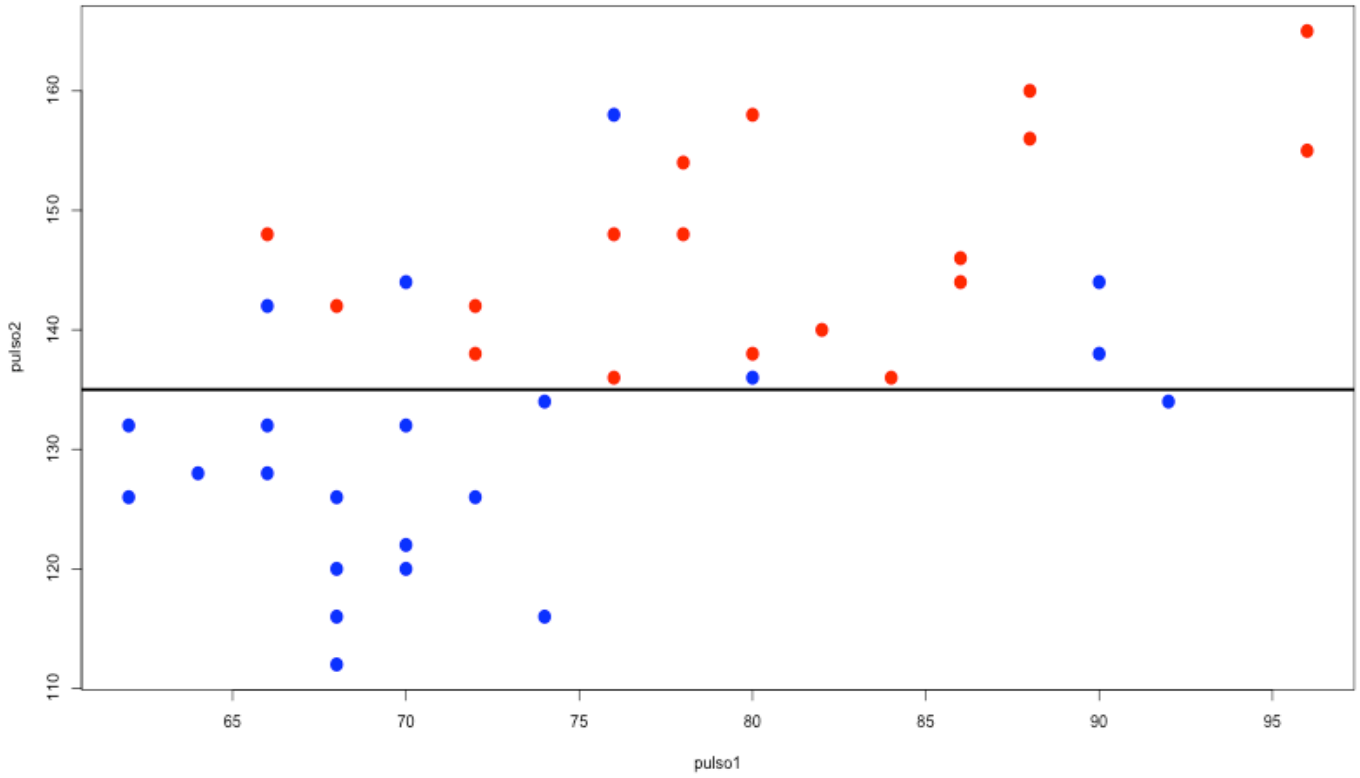
CART



El Mejor Arbol

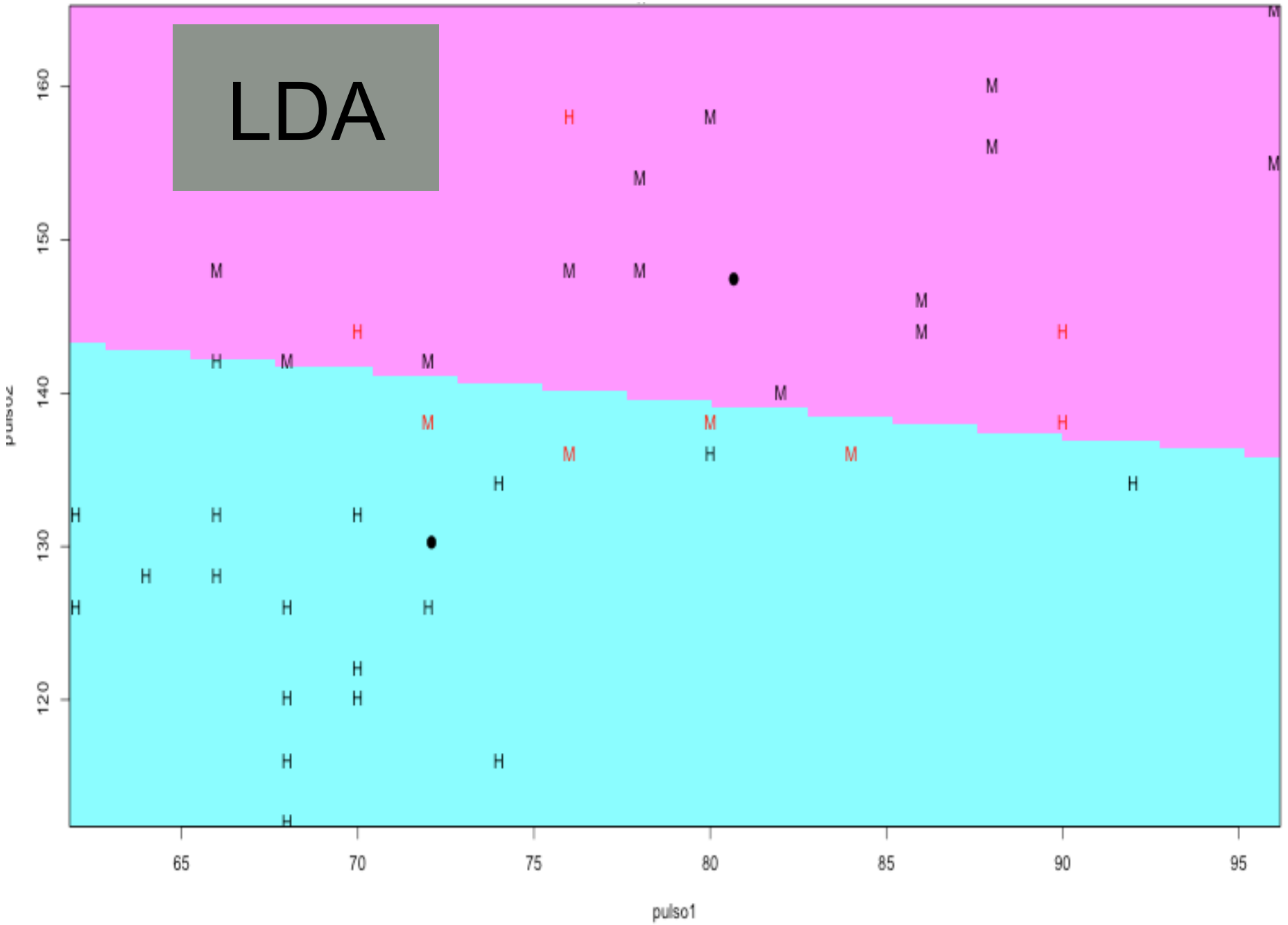
pulso2 < 135

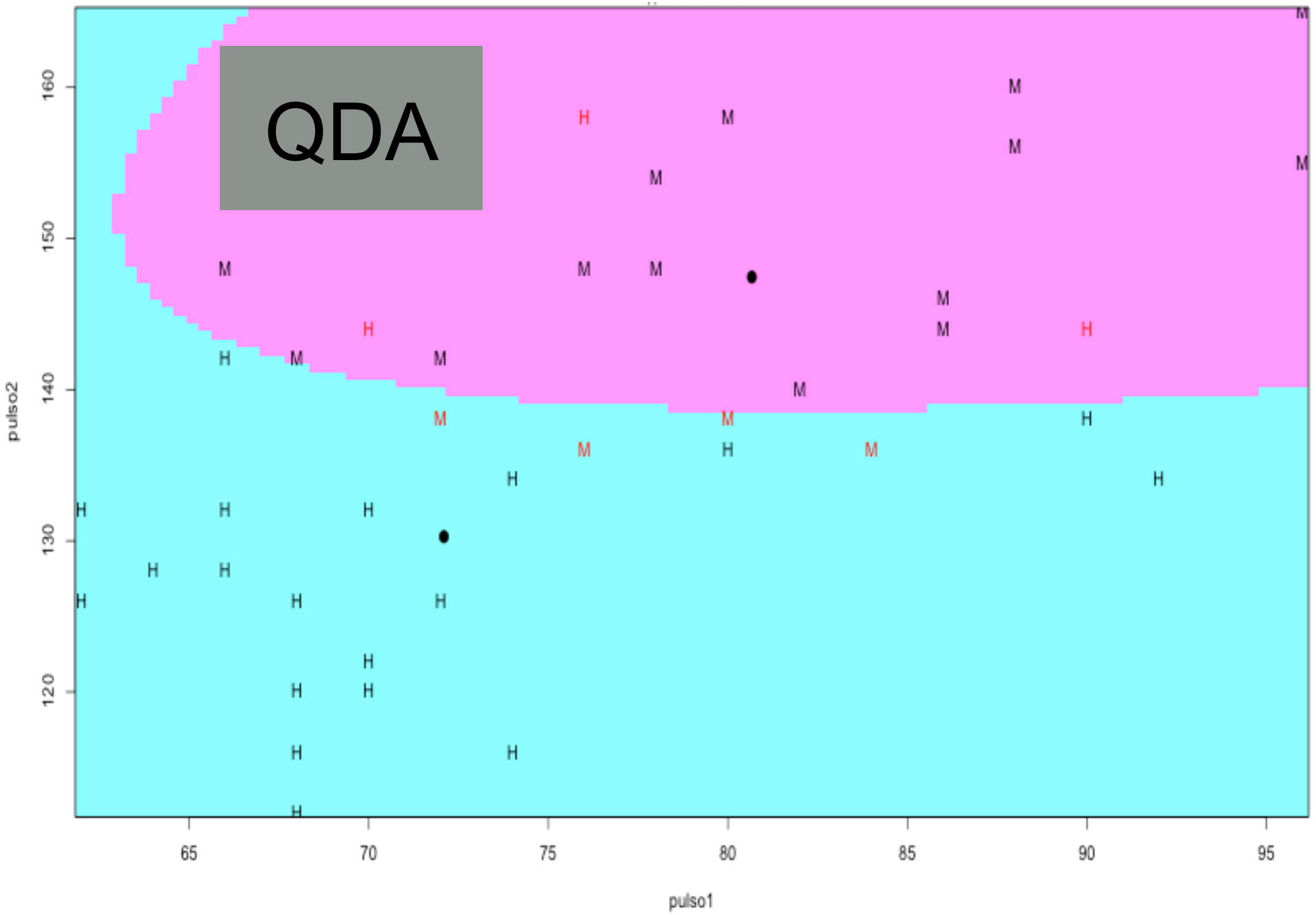
```
1) root 40 55.05 Hombre ( 0.55 0.45 )
  2) pulso2 < 135 16  0.00 Hombre ( 1.00 0.00 ) *
  3) pulso2 > 135 24 26.99 Mujer ( 0.25 0.75 ) *
```



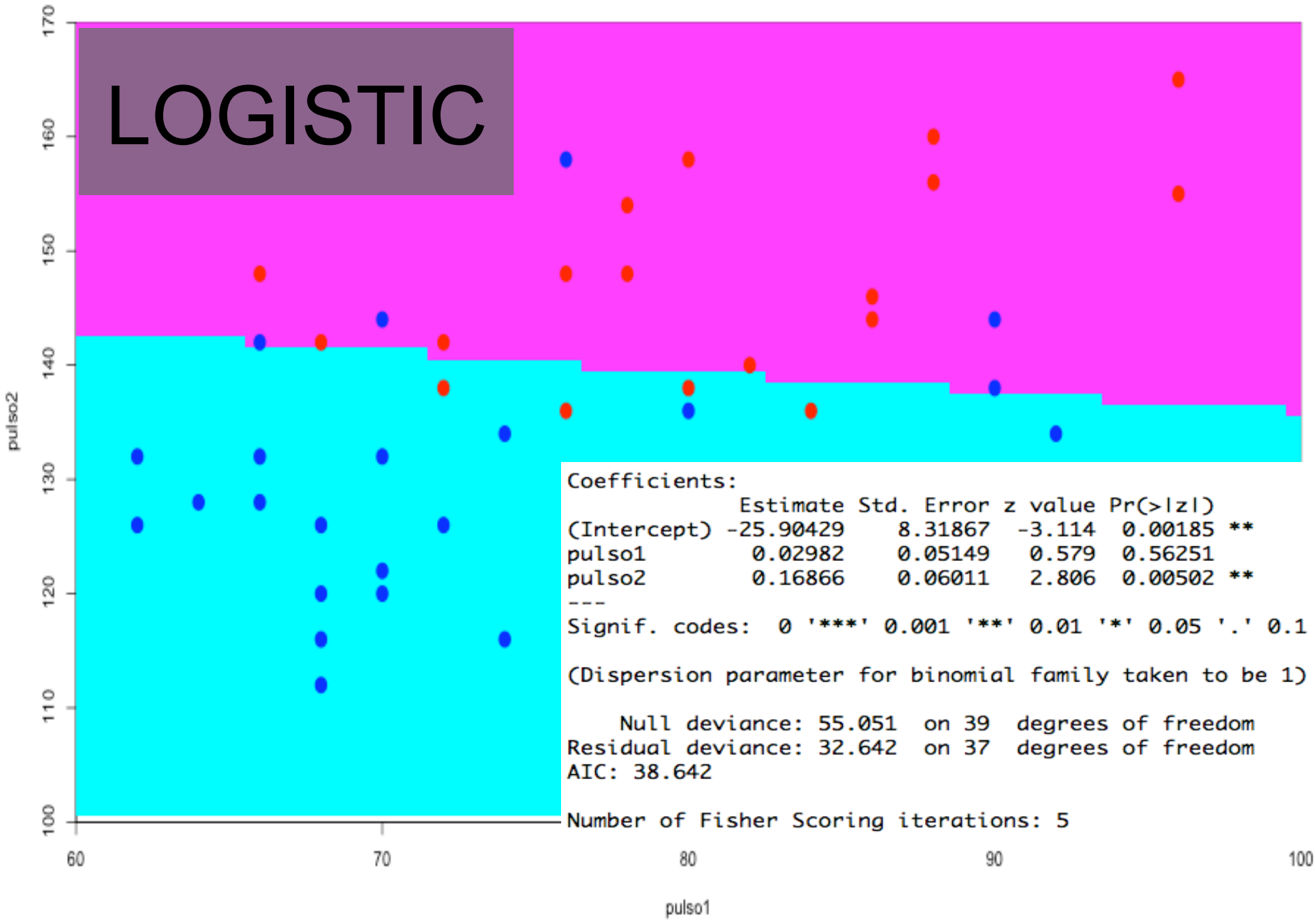
Hombre

Mujer


LDA
160
150
140
130
120
65
70
75
80
85
90
95
pulso1


QDA
pulso2
pulso1
120
130
140
150
160
65
70
75
80
85
90
95

LOGISTIC




```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -25.90429    8.31867  -3.114  0.00185 **
pulso1        0.02982    0.05149   0.579  0.56251
pulso2        0.16866    0.06011   2.806  0.00502 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.051  on 39  degrees of freedom
Residual deviance: 32.642  on 37  degrees of freedom
AIC: 38.642

Number of Fisher Scoring iterations: 5
```