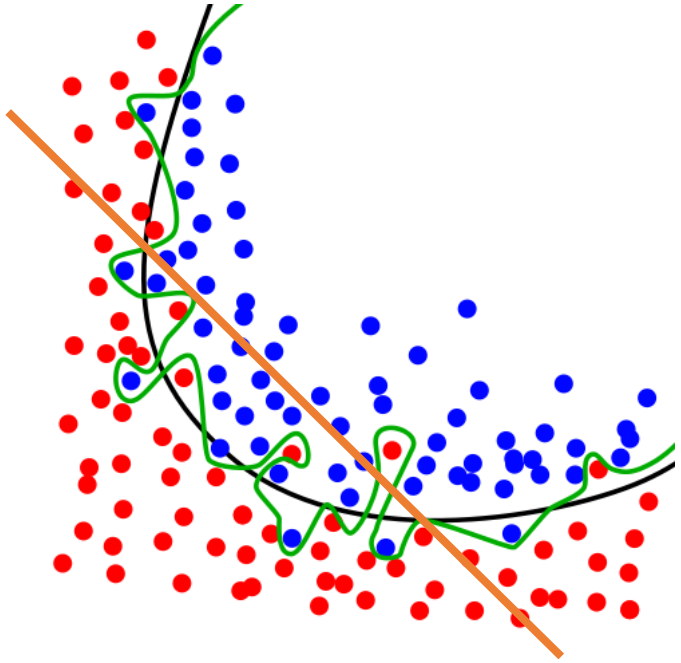
Modelos de Clasificación

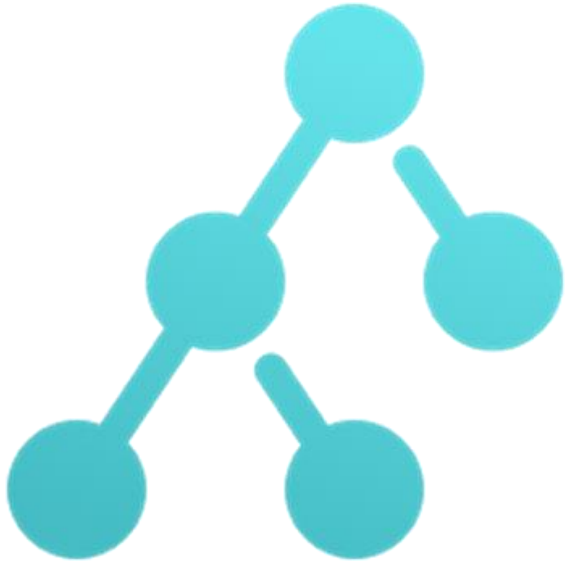
Que son los Modelos de Clasificación?

Los modelos de Clasificación supervisada, permiten asignar a cada registro de datos (observaciones) una clase pre establecida(categoría o estado).



2 categorías:

- Azul (No Moroso)
- Rojo (Moroso)



Arboles de Clasificación

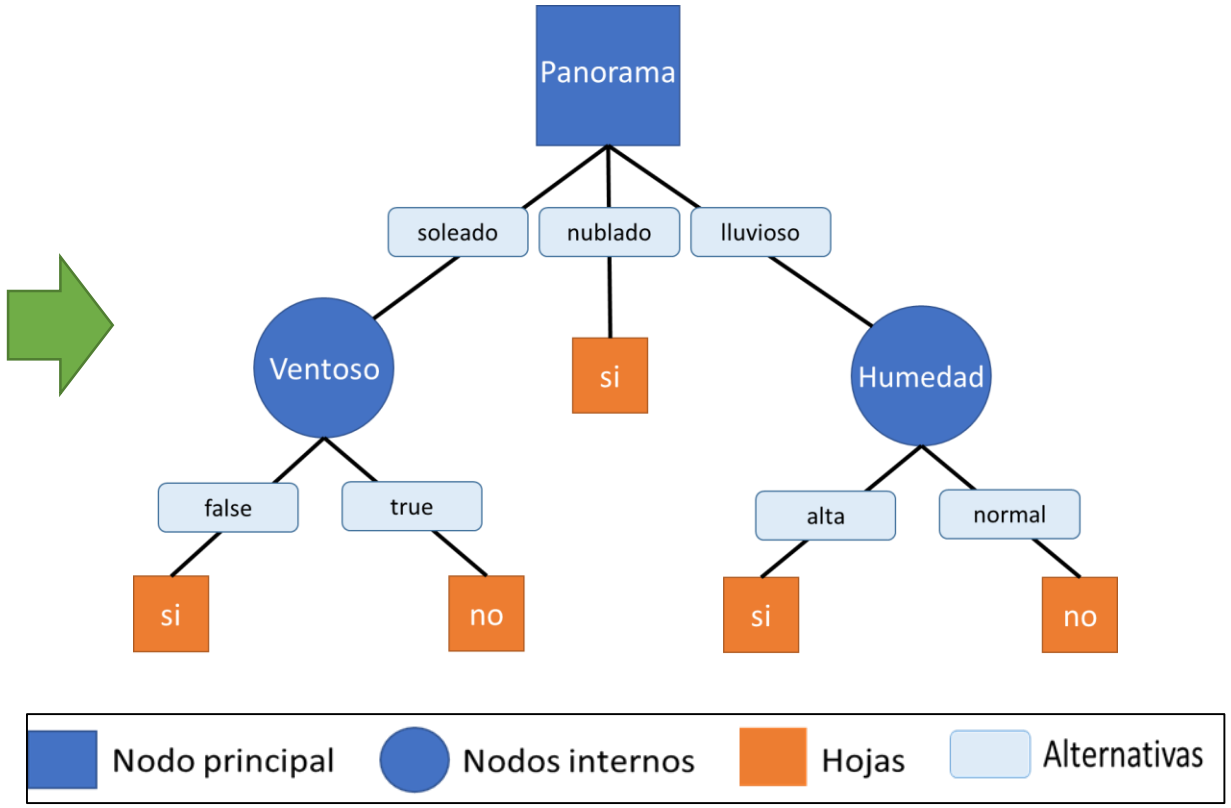
Arboles de Clasificación

- Es un conjunto de Reglas organizadas en una estructura jerárquica.

Panorama	Temperatura	Humedad	Ventoso	Paseo
lluvioso	caliente	alta	false	no
lluvioso	caliente	alta	true	no
nublado	caliente	alta	false	si
soleado	templado	alta	false	si
soleado	frio	normal	false	si
soleado	frio	normal	true	no
nublado	frio	normal	true	si
lluvioso	templado	alta	false	no
lluvioso	frio	normal	false	si
soleado	templado	normal	false	si
lluvioso	templado	normal	true	si
nublado	templado	alta	true	si
nublado	caliente	normal	false	si
soleado	templado	alta	true	no

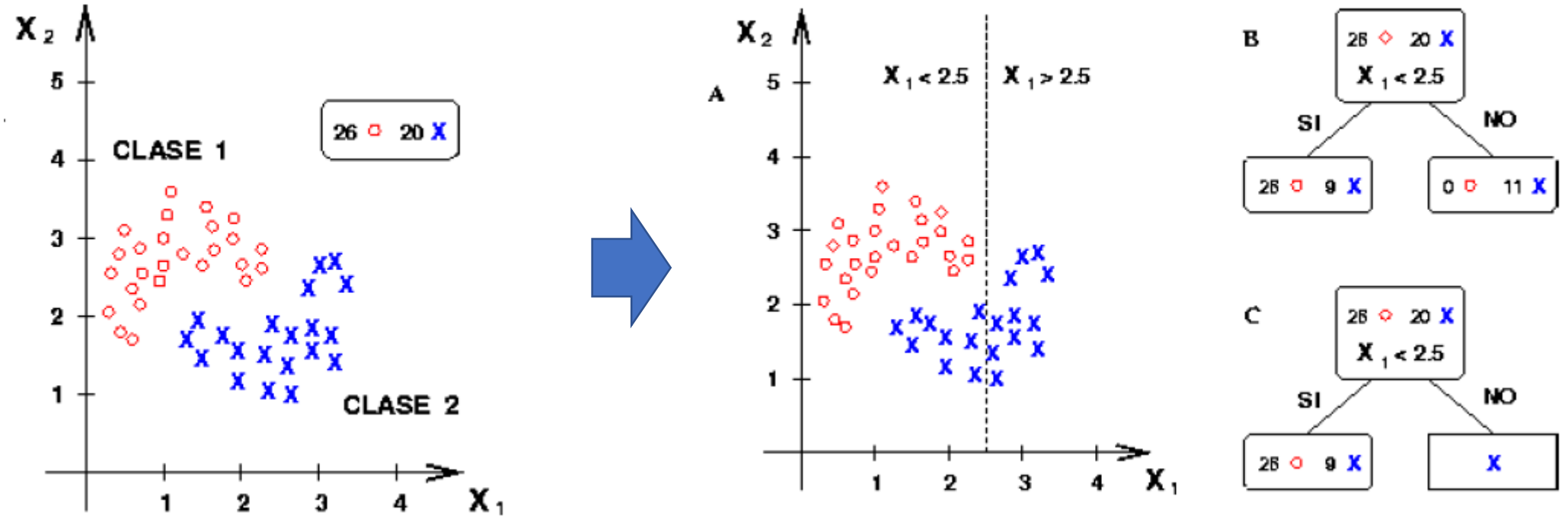
Predictores

Target



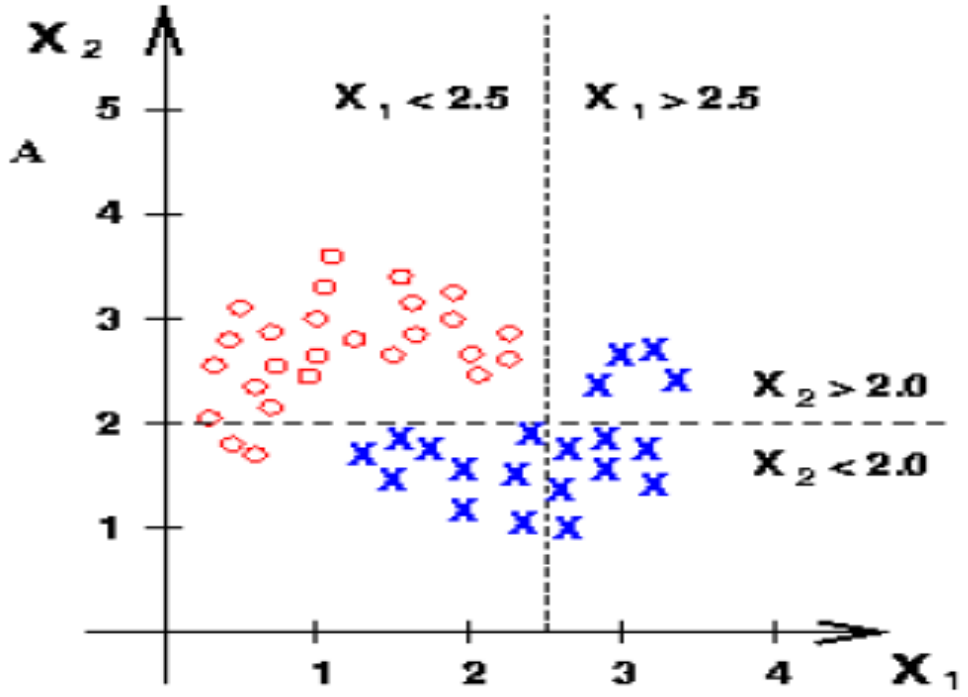
Arboles de Clasificación – Fundamento Cualitativo

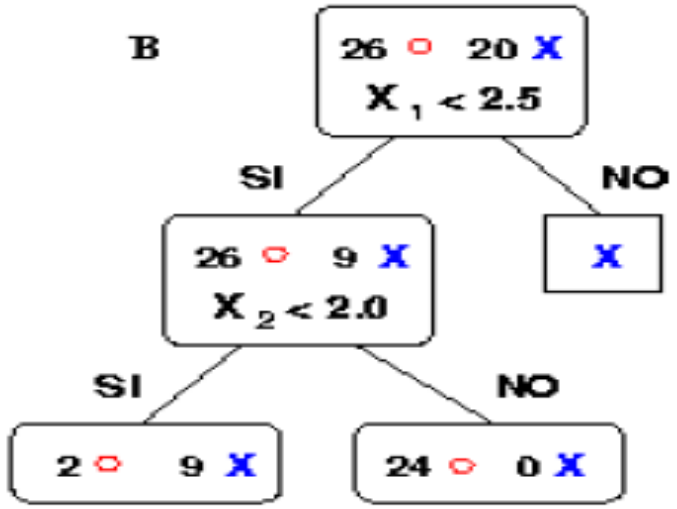
Se busca determinar cortes en las variables que permitan maximizar la proporción de clasificados de un solo tipo.



Arboles de Clasificación – Fundamento Cualitativo

El conjunto de cortes en un orden especifico determinan el patrón para nuevos casos





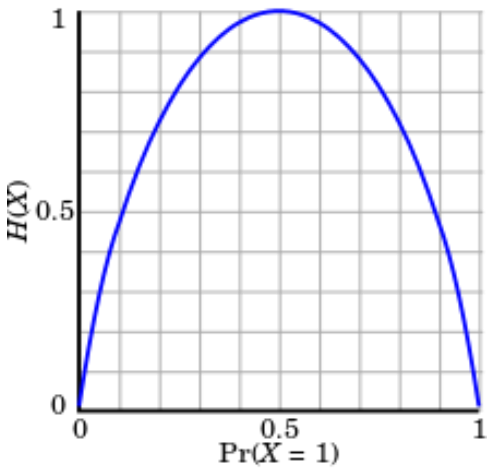
Arboles de Clasificación - Fundamento Cuantitativo

- Orden de las variables a evaluar.
  - Se debe escoger la variable que permite clasificar en forma **mas certera** las observaciones, es decir genera ramas mas homogéneas (no tenemos dudas cual es la clasificación).
  - **El nivel de certeza** la medimos matemáticamente a través un indicador llamado **Ganancia de Información**, debemos escoger la variable con mayor ganancia como nodo de decisión.
  - **La Ganancia de Información** consiste en el decremento de la incertidumbre. La incertidumbre se mide matemáticamente a través de la **Entropia**.
- Condición de parada
  - Se da cuando se tiene la certeza total, es decir cuando la Entropía es cero.
  - Una **rama** con **entropía cero** se convierte en una hoja (nodo-respuesta), ya que representa una muestra completamente homogénea, en la que todos los ejemplos tienen la misma clasificación. Si no es así, la rama debe seguir subdividiéndose con el fin de clasificar mejor sus nodos

Arboles de Clasificación - Fundamento Cuantitativo

- Entropía de Shannon:
  - Sea S un conjunto de valores de la variable X, que se puede dividir en C clases, donde S_i es el sub conjunto de valores de la clase C_i.
  - La Entropía del conjunto S, esta en función de la proporción de los distintos valores que puede tomar X: $E(S) = f(p_i)$
  - p_i es la proporción de ocurrencias en la clase Ci del conjunto S : $p_i = \frac{|S_i|}{|S|}$
  - Finalmente la entropía de la variable S es:

$$E(S) = \sum_{i \in C} (-p_i \, log_2 p_i)$$



Arboles de Clasificación – Hands On

```r
#traninig and test sets
library(caTools)
set.seed(123)
split<-sample.split(ds$Purchased,SplitRatio=0.7)
training_set<-subset(ds,split==TRUE)
test_set<-subset(ds,split==FALSE)
#modelar
library(rpart)
arbol<-rpart(formula = Purchased ~ .,
             data = training_set,
             method = "class")
summary(arbol) # Variable importance Age (62%) EstimatedSalary (38%)
print(arbol)
#mostrando el arbol
library(rpart.plot)
rpart.plot(arbol)

#Evaluando el modelo (y_pred vs test_set$Purchased)
y_pred<-predict(arbol,newdata = test_set,type = "class")
cm<-table(y_pred,test_set$Purchased)
```

Evaluación de Desempeño

- Matriz de Confusión

A1	A2	A3	clase	predicción
19	15	2	SI	SI
13	7	21	NO	SI
20	24	24	NO	SI
13	9	7	NO	NO
5	3	8	NO	SI
5	20	21	NO	NO
13	4	18	NO	NO
20	23	13	NO	SI
18	8	10	NO	SI
10	8	14	NO	SI
4	15	6	NO	NO
19	11	12	NO	NO
17	15	15	NO	SI
15	3	2	SI	NO
23	13	18	NO	NO
3	21	1	NO	NO
1	16	1	NO	SI
15	22	16	NO	NO
6	14	1	NO	NO
19	6	11	SI	SI
10	11	14	SI	NO
3	15	23	NO	SI
20	7	14	NO	SI
22	20	14	NO	SI
9	6	16	NO	SI

$$error = \frac{FP + FN}{total}$$

Caso mal clasificados

$$exito = \frac{VP + VN}{total}$$

Caso bien clasificados

$$sensibilidad = \frac{VP}{VP + FN}$$

Probabilidad de clasificar correctamente a un individuo con el valor de interés (+)

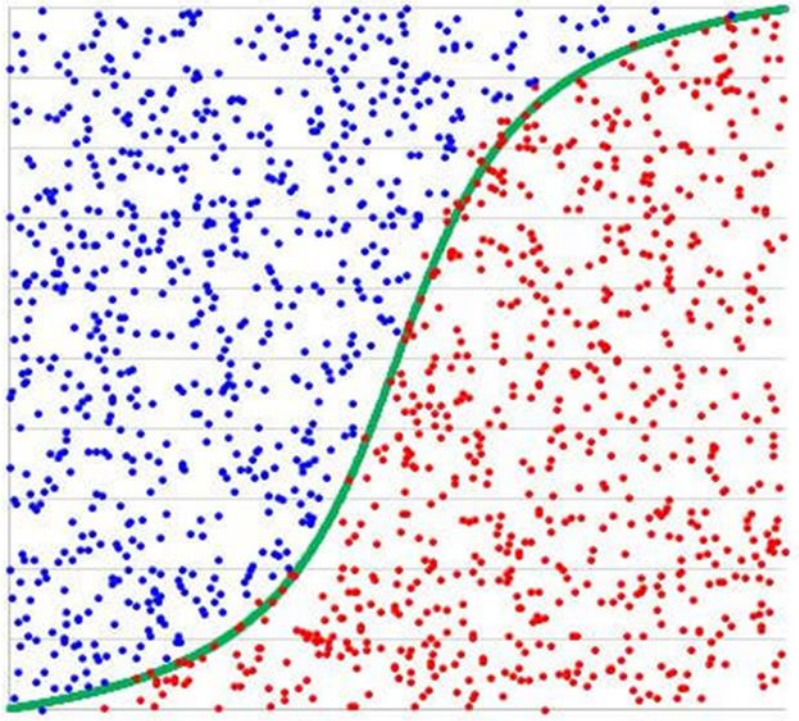
$$especificidad = \frac{VN}{VN + FP}$$

Probabilidad de clasificar correctamente a un individuo sin el valor de interés (-)

Clasificador \ Valor real	(+)	(−)
(+)	VP	FP
(−)	FN	VN

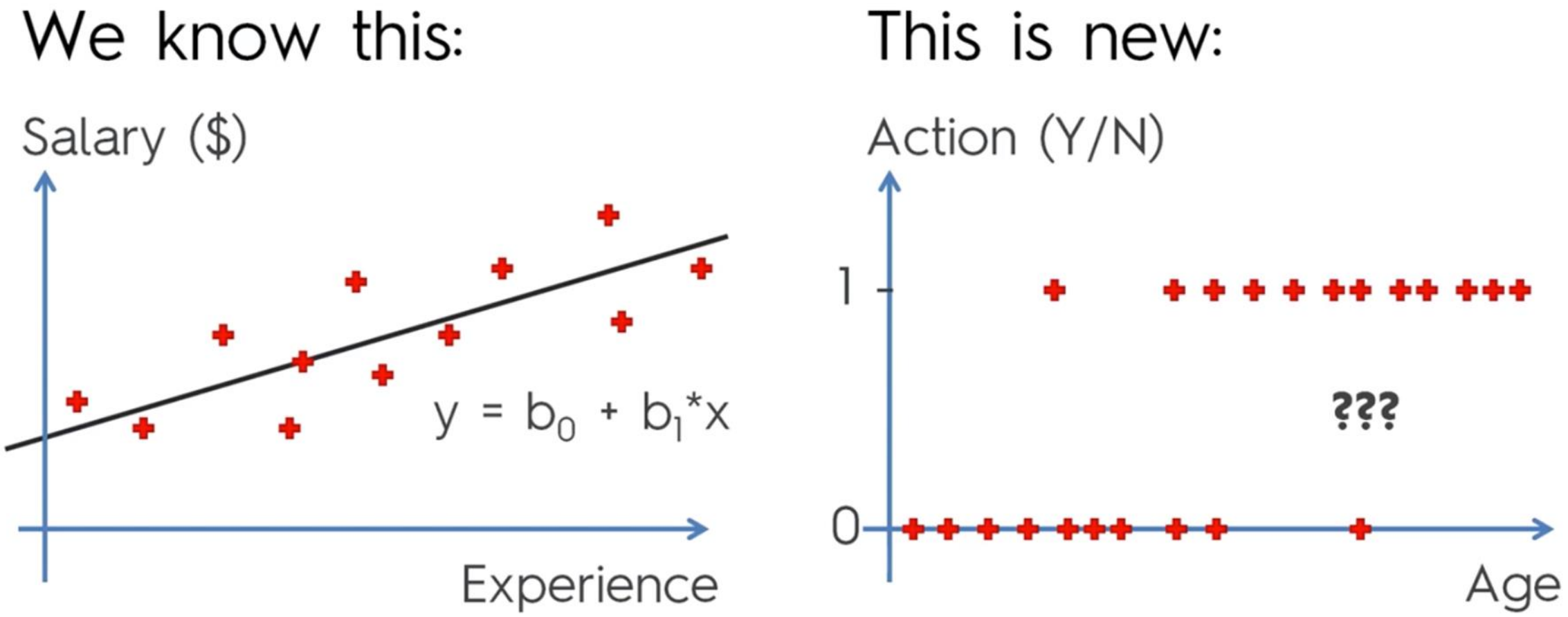
FP = error de tipo I (α)
Muy costoso

FN = error de tipo II (β)
β <5%, 20%>



Regresión Logística

Regresión Logística



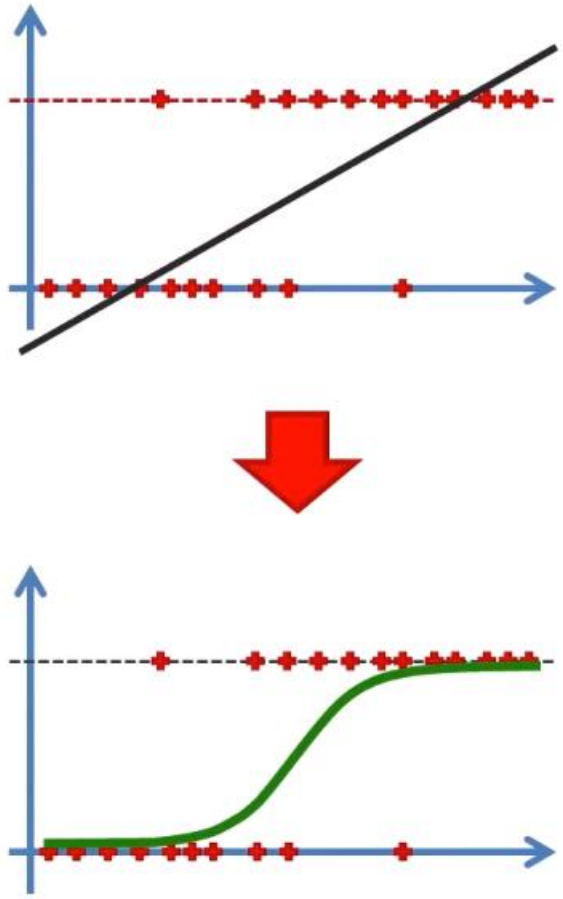
Regresión Logística

$$y = b_0 + b_1 * x$$

Sigmoid Function

$$p = \frac{1}{1 + e^{-y}}$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * x$$



Regresión Logística