Modelo de Clasificación Nivel 1: Regresión Logística y Arboles de Clasificación

AGENDA

- Esquema de aprendizaje de la Regresión Logística

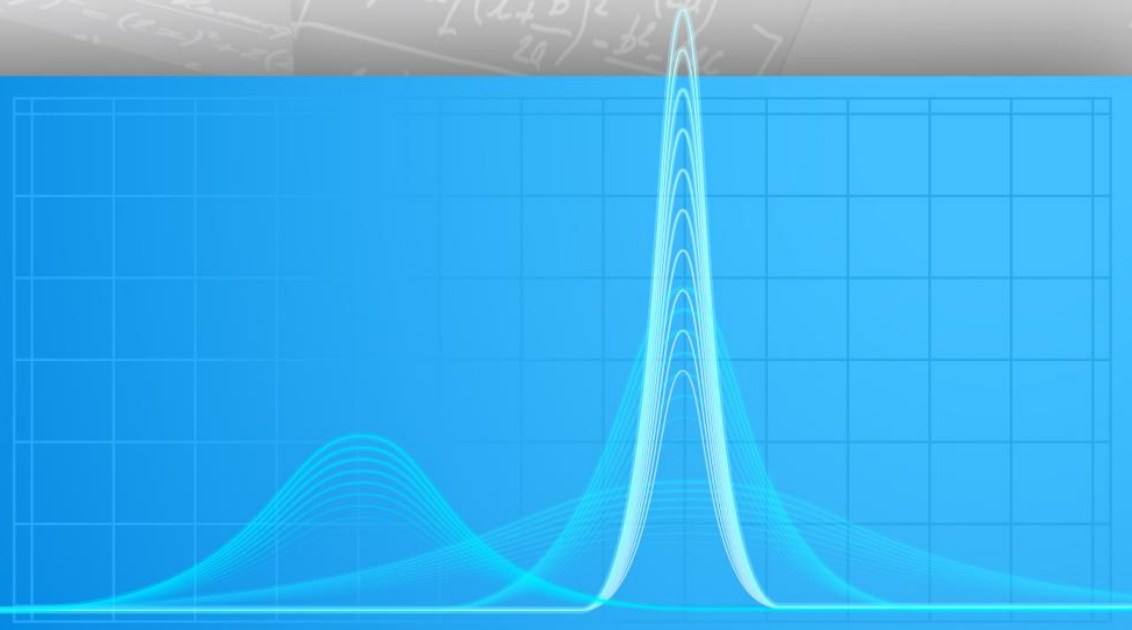
R AnalyticFlow Regresión Logística

- Esquema de aprendizaje de la Árbol de Clasificación

R AnalyticFlow Árbol de Clasificación

- **Caso de negocio**: Incorporación de modelo de venta en campaña de adquisición de clientes en seguros salud

Algoritmo Machine Learning:
Regresión Logística



FOCO DEL PROBLEMA

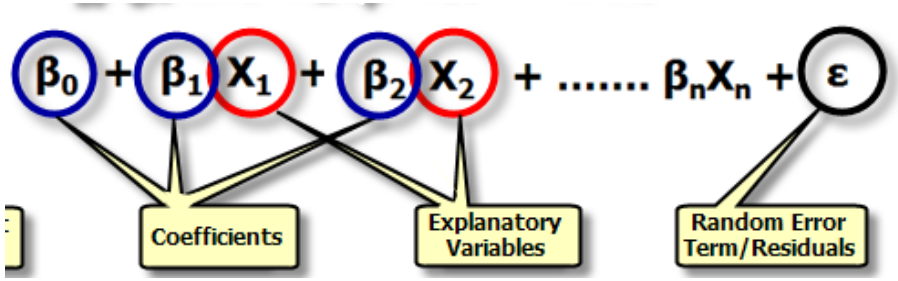
Business Insight Center

Manejar la incertidumbre si un cliente nos compra o no según sus características del cliente

Cliente compra o no?

$$P(A/B) =$$

Reg. Logistica

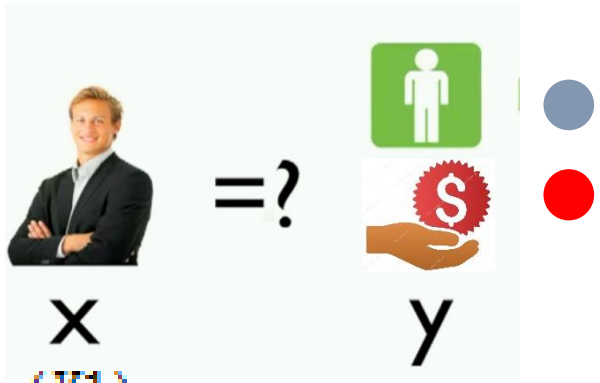


Reducir el error

Conozco características del cliente

ALGORITMO REGRESIÓN LOGÍSTICA

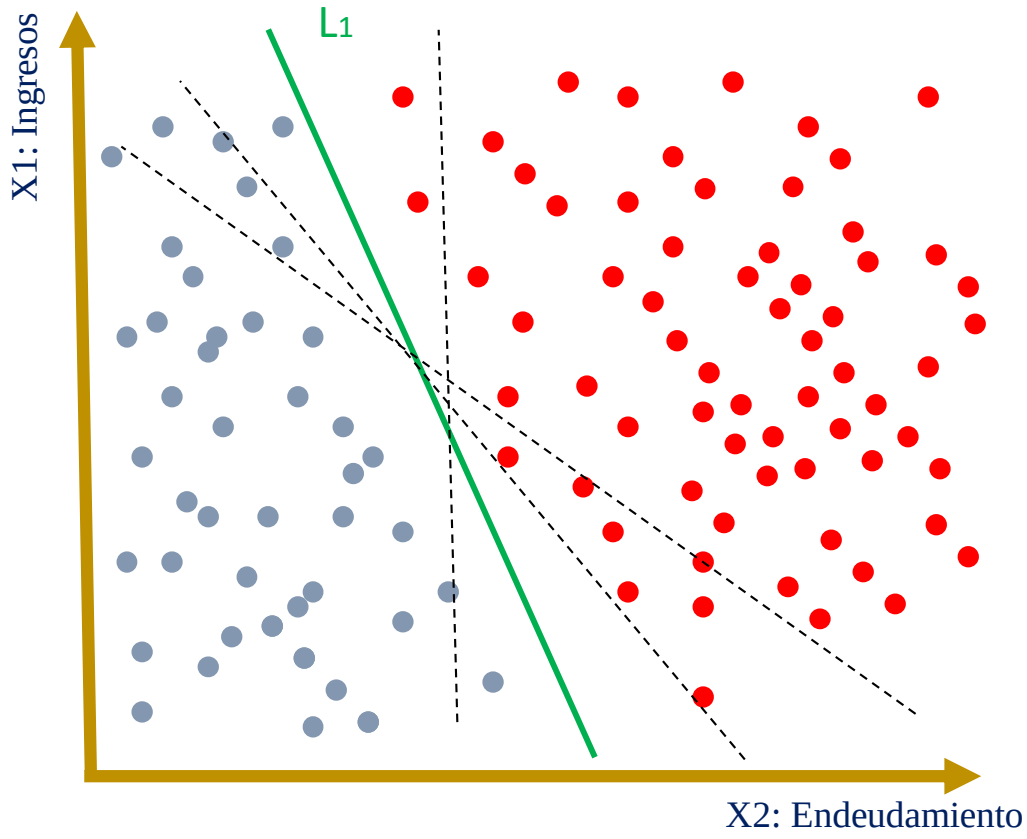
El problema



$$\begin{pmatrix} X1 \\ X2 \\ .. \\ Xp \end{pmatrix}$$

Data histórica

Representación gráfica



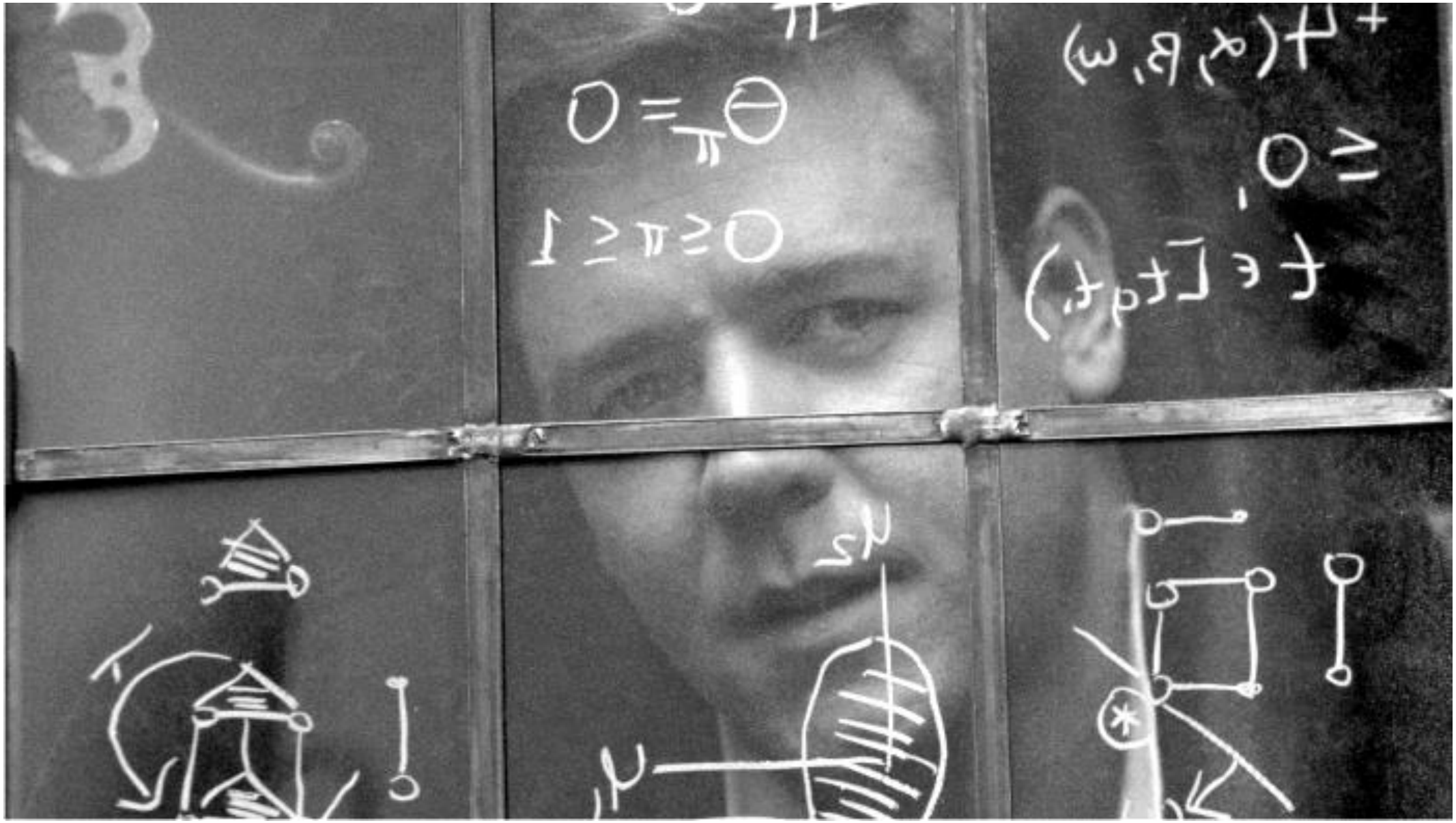
Para el ejemplo visual definamos :

Variable Y

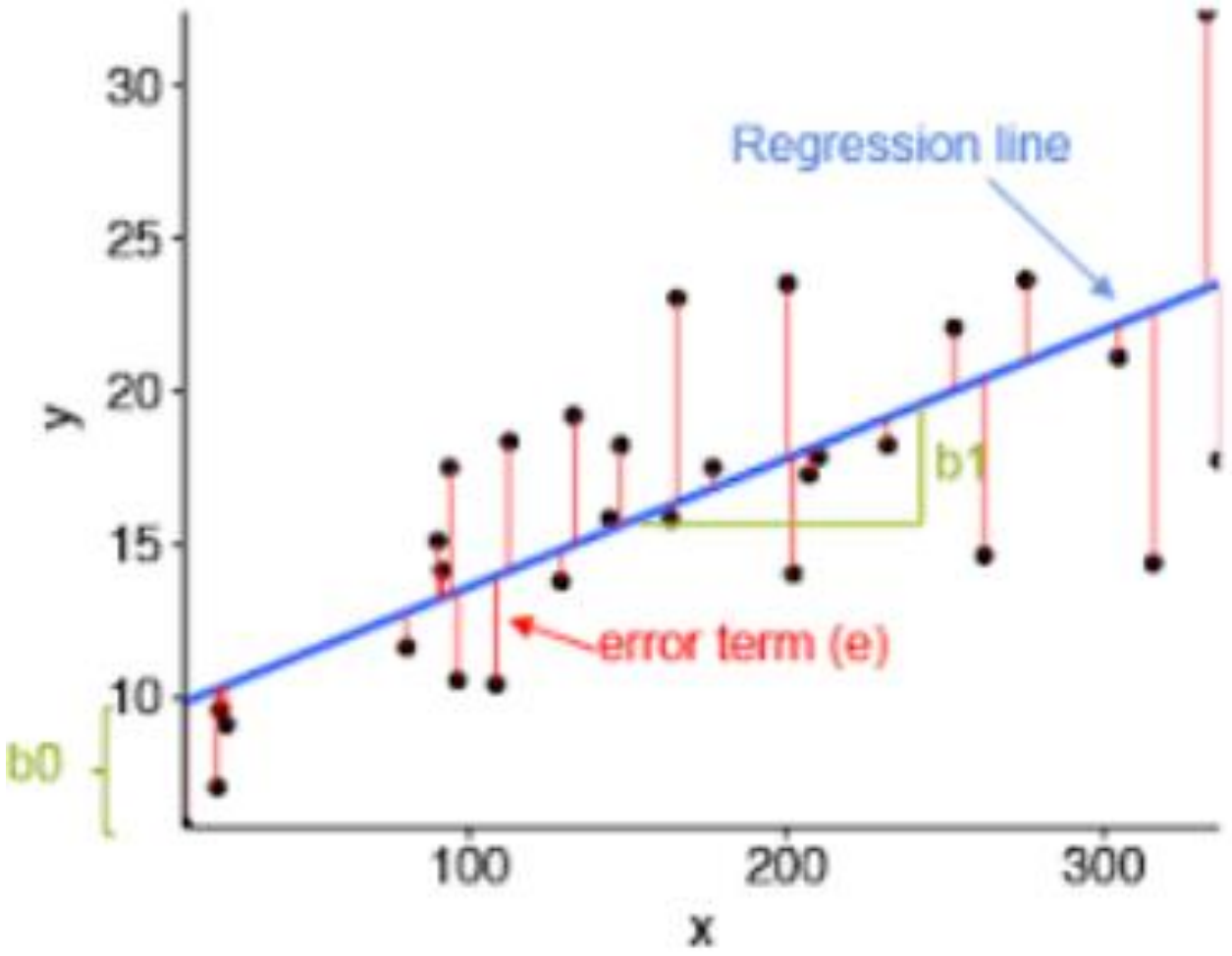
$$\begin{matrix} X1 & Ingresos\ mesuales \\ X2 & Endeudamient\ SSFF \end{matrix} \begin{pmatrix} 245 \\ 3450 \end{pmatrix} \begin{pmatrix} 345 \\ 4532 \end{pmatrix} \begin{pmatrix} 234 \\ 3452 \end{pmatrix} \begin{pmatrix} 45 \\ 1234 \end{pmatrix} \cdots \begin{pmatrix} 345 \\ 1232 \end{pmatrix} \begin{pmatrix} 345 \\ 123 \end{pmatrix}$$

$$L(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}.$$

QUÉ MATEMÁTICA-ESTADÍSTICA HAY DETRÁS?



APRENDER DE TODA LA HISTORIA CON REGRESION LINEAL



ahorro

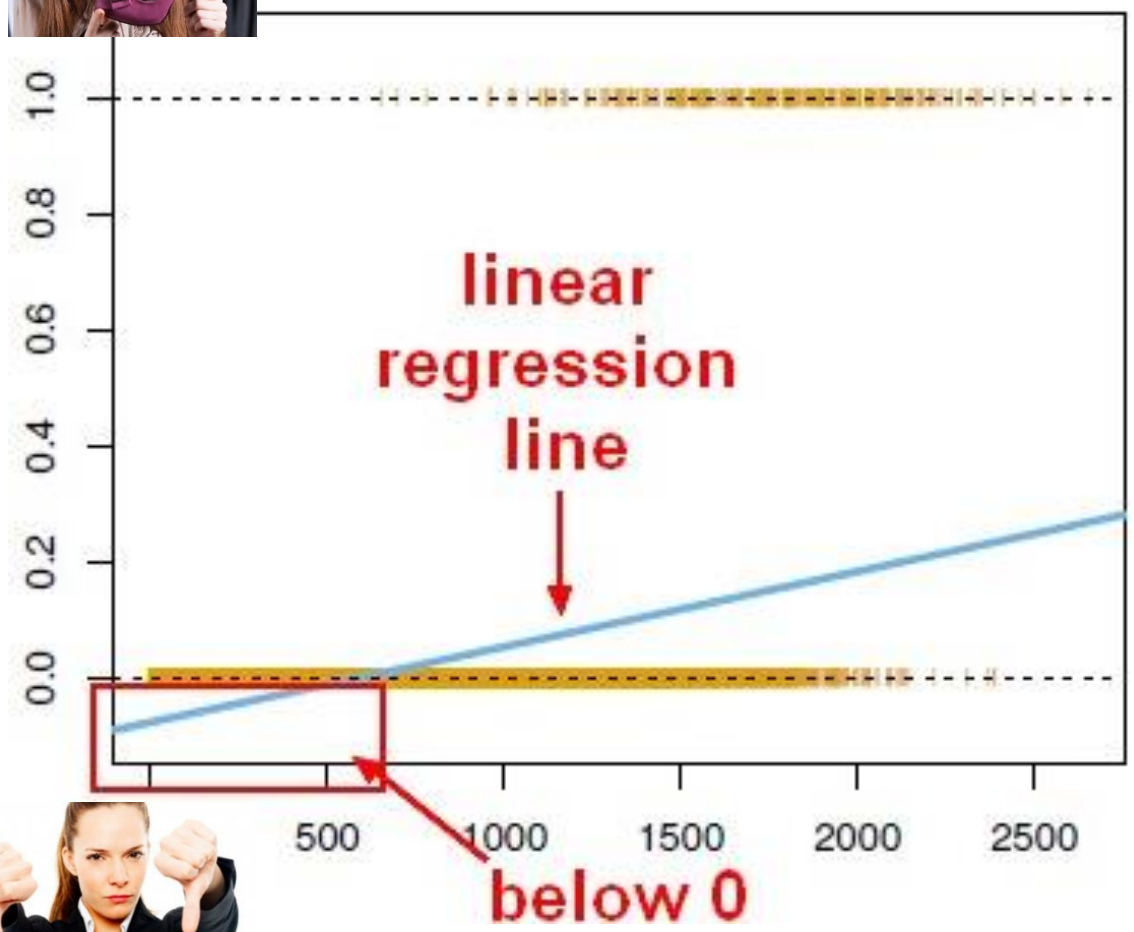
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Linear component

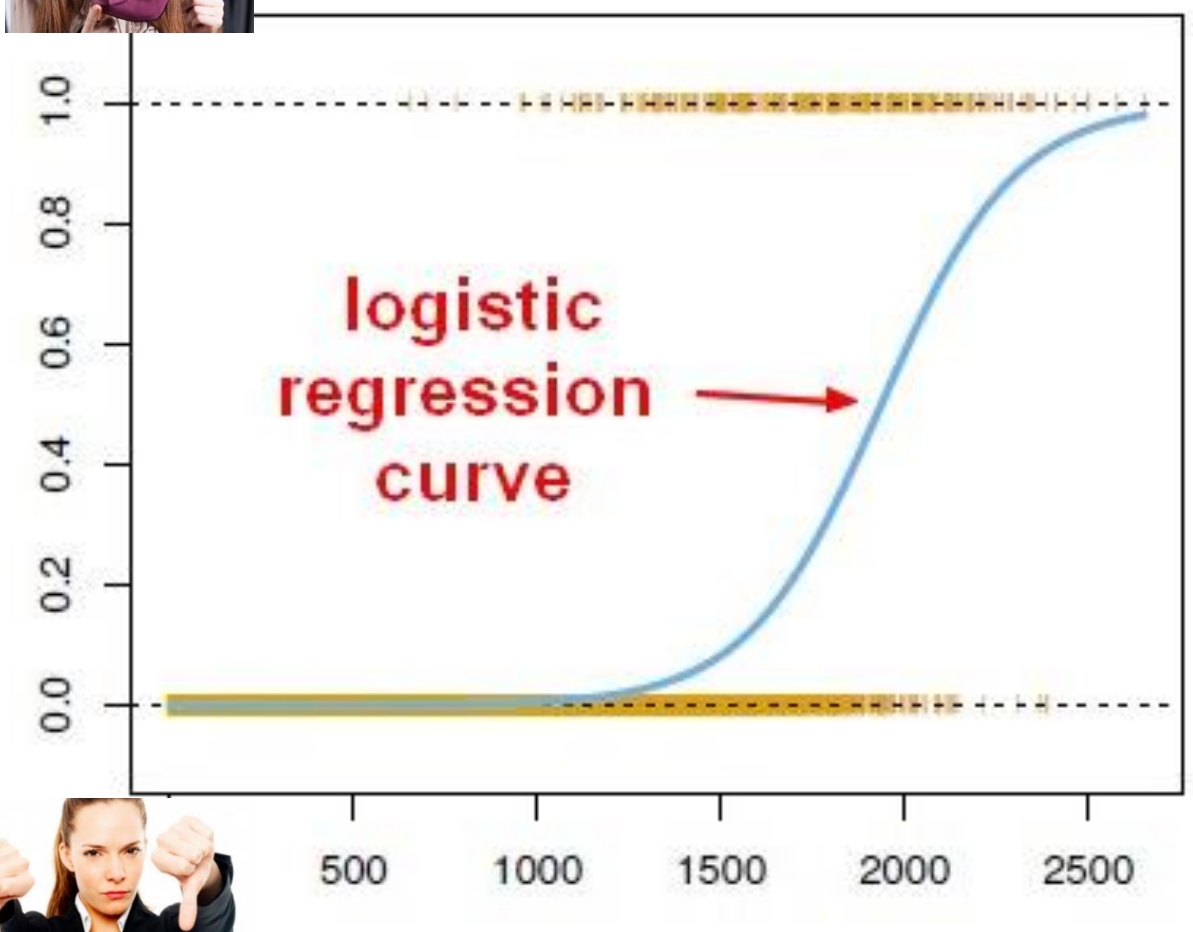
Random Error component

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

APRENDER DE TODA LA HISTORIA CON REGRESION LINEAL - CLASIFICACIÓN



Sueldo de un cliente



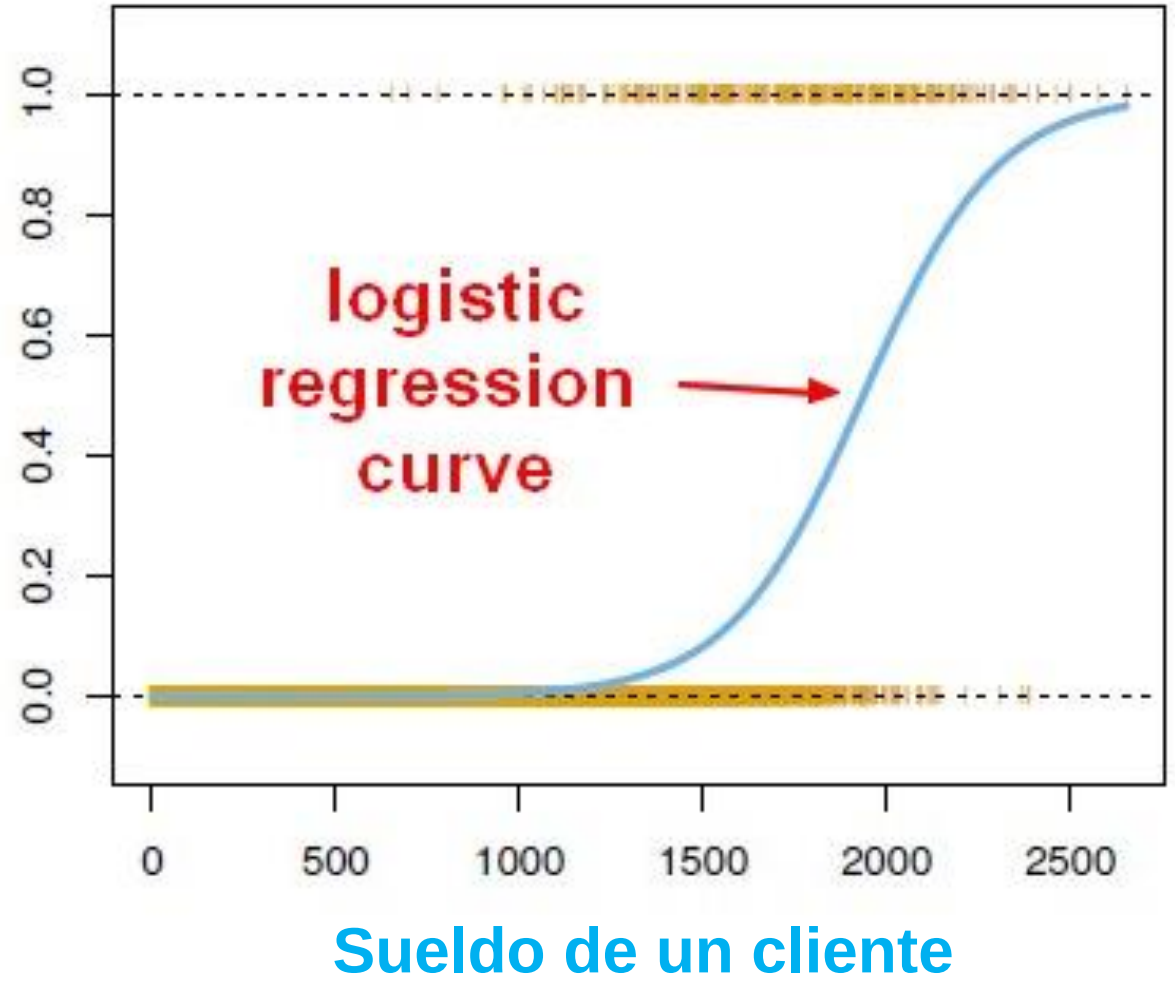
Sueldo de un cliente

REGRESIÓN LOGÍSTICA TÉRMINOS DE NEGOCIO

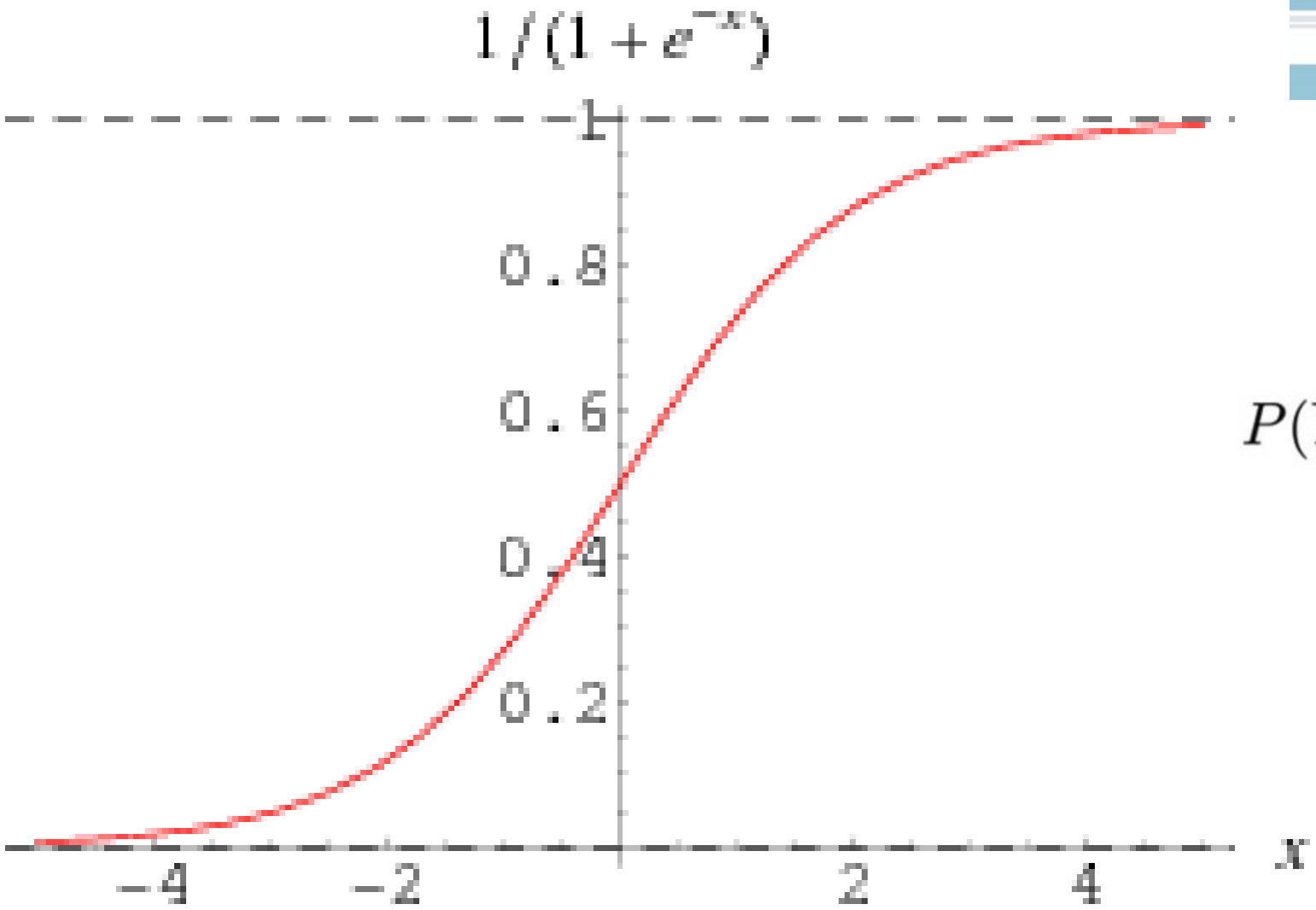
1. Cada cliente solo tiene 2 opciones de decisión y sigue una distribución Binomial

2. La decisión de cada cliente son Independientes entre si

3. Se relaciona la ponderación de las variables del cliente y la probabilidad de compra



FUNCIÓN SIGMOIDE LOGIT

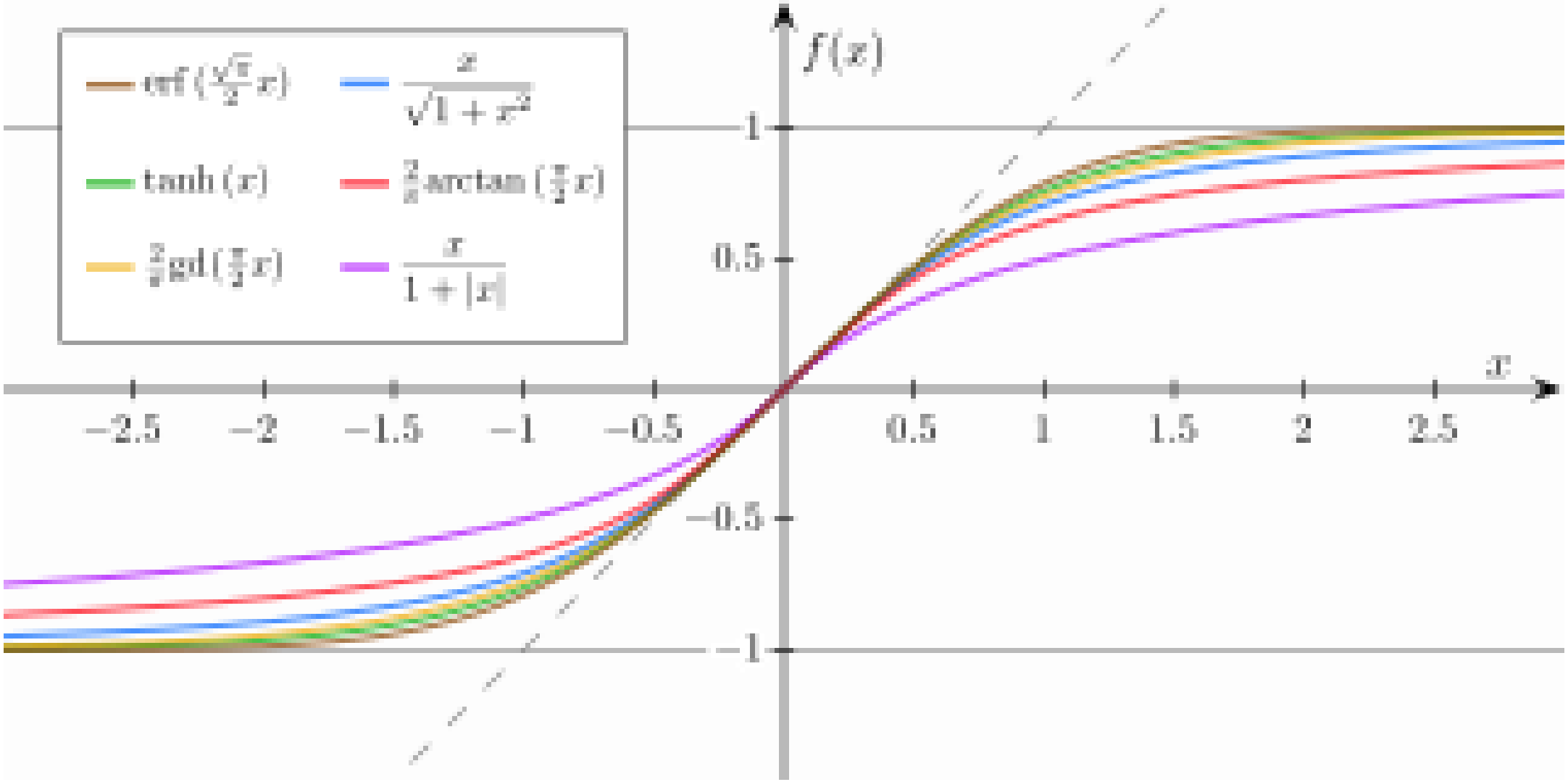


Cliente compra o no?

$$P(Y=1|X)=\frac{e^{(\beta_o+\beta_1 x)}}{e^{(\beta_o+\beta_1 x)}+1}$$

Conozco característic del cliente

OTRAS FUNCIONES SIGMOIDE



OTRAS FUNCIONES SIGMOIDE

Se puede interpretar la suma ponderada de características del cliente?

$$\beta_0 + \beta_1 x$$

$$p(X) = \frac{e^{(\beta_o+\beta_1 x)}}{e^{(\beta_o+\beta_1 x)} + 1}$$

$$p(e^{(\beta_o+\beta_1 x)} + 1) = e^{(\beta_o+\beta_1 x)}$$

$$p.e^{(\beta_o+\beta_1 x)} + p = e^{(\beta_o+\beta_1 x)}$$

$$p = e^{(\beta_o+\beta_1 x)} - p.e^{(\beta_o+\beta_1 x)}$$

$$p = e^{(\beta_o+\beta_1 x)}(1-p)$$

$$\frac{p}{1-p} = e^{(\beta_o+\beta_1 x)}$$

$$\ln(\tfrac{p}{1-p}) = \beta_0 + \beta_1 x$$

EVALUACIÓN DE LA REGRESION LOGISTICA

Business Insight Center

Criterio de AIC

Akaike Information Criteria

Indicador importante de evaluación de modelo similar a

R-ajustado

Evita el sobre ajuste

$$\mathrm{SSE} = \sum_i \left(y_{exp}(t_i) - y_{mod}(t_i)\right)^2$$

$$\mathrm{AIC} = n_t \cdot \ln\left(\frac{\mathrm{SSE}}{n_t}\right) + 2 \cdot (n_p + 1) + \frac{2 \cdot (n_p + 1) \cdot (n_p + 2)}{n_t - n_p - 2}$$

Matriz de confusión

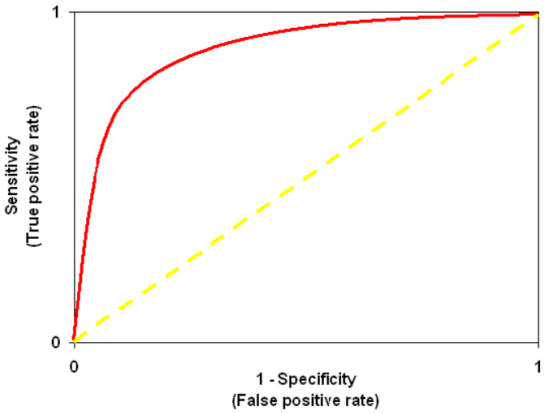
Tabla cruzada entre pronósticos

y decisiones reales de los clientes. Se derivan métricas como: Precisión

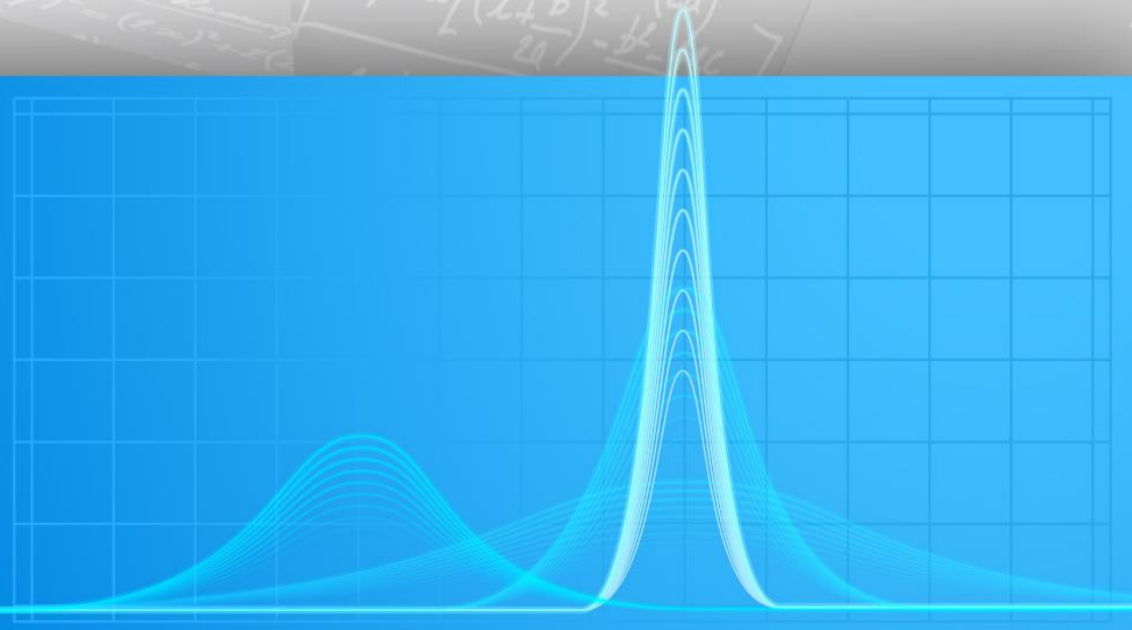
	1 (Predicted)	0 (Predicted)
1 (Actual)	True Positive	False Negative
0 (Actual)	False Positive	True Negative

ROC

Probabilidad de acierto del modelo evitando falsas alarmas

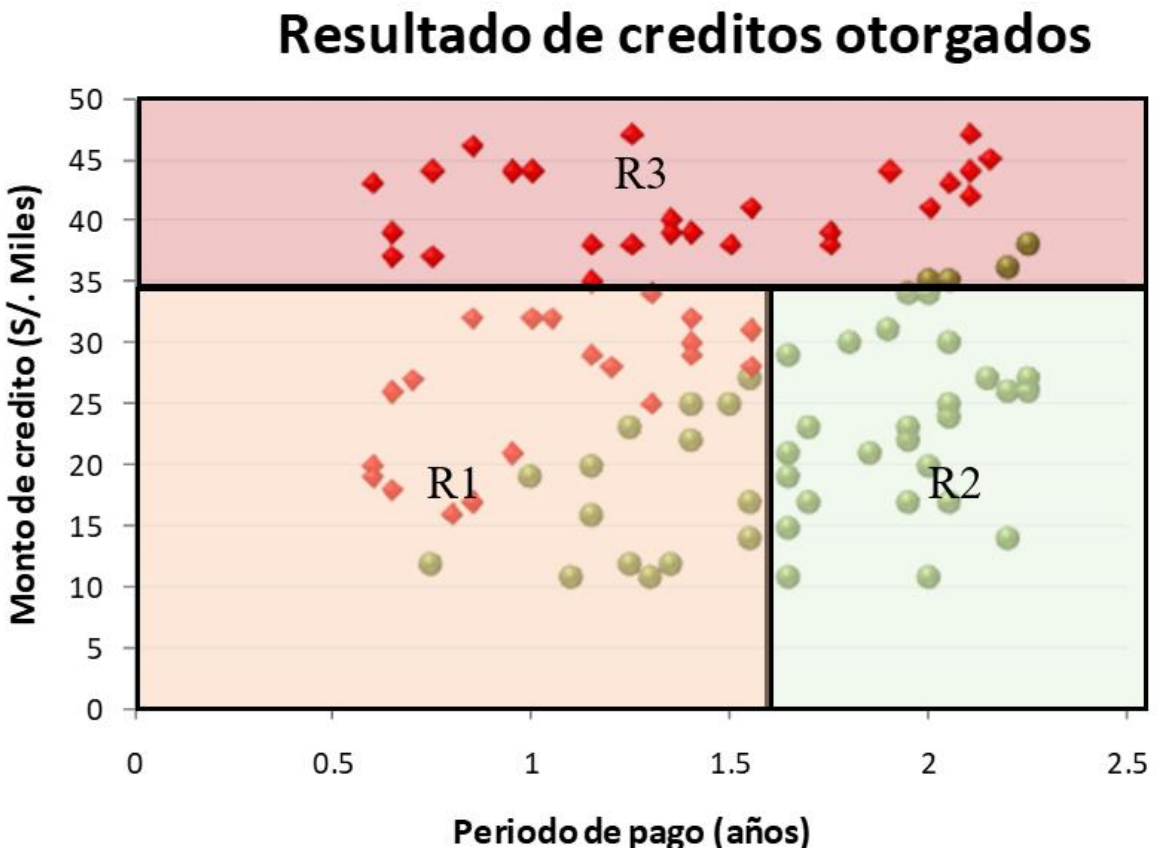


Algoritmo Machine Learning:
Árbol de Clasificación



ALGORTIMO ARBOL DE DECISION

Business Insight Center



Impureza inicial: 0.25

Clientes morosos: 46 (51%)

Tamaño de muestra: 100%

Bondad de Ajuste de cortes:

= 0.25 – (30% 0.07 +0% 0+ 40% 0.24)

= 0.13

Monto de crédito > 35 mil soles

Si

No

Impureza R3: 0.07

Clientes morosos: 25 (93%)

Tamaño de muestra: 30%

Periodo de pago > 1.6 años

Si

No

Impureza R2: 0

Clientes morosos: 0 (0%)

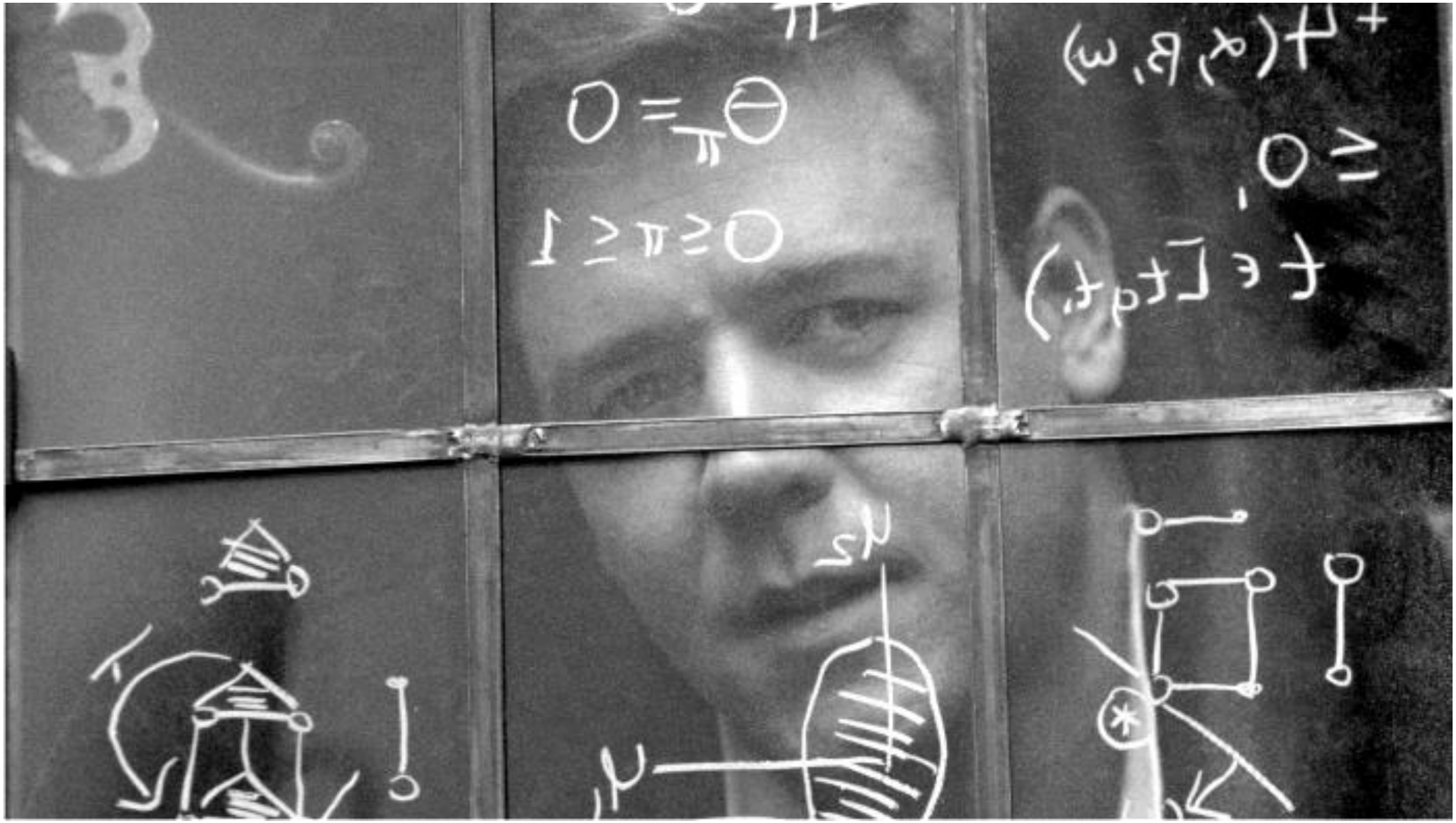
Tamaño de muestra: 31%

Impureza R3: 0.24

Clientes morosos: 21 (58%)

Tamaño de muestra: 40%

QUÉ MATEMÁTICA-ESTADÍSTICA HAY DETRÁS?



ALGORTIMO ARBOL DE DECISION

Business Insight Center

Reducir la impureza

- Medición Error de clasificación o Bayes

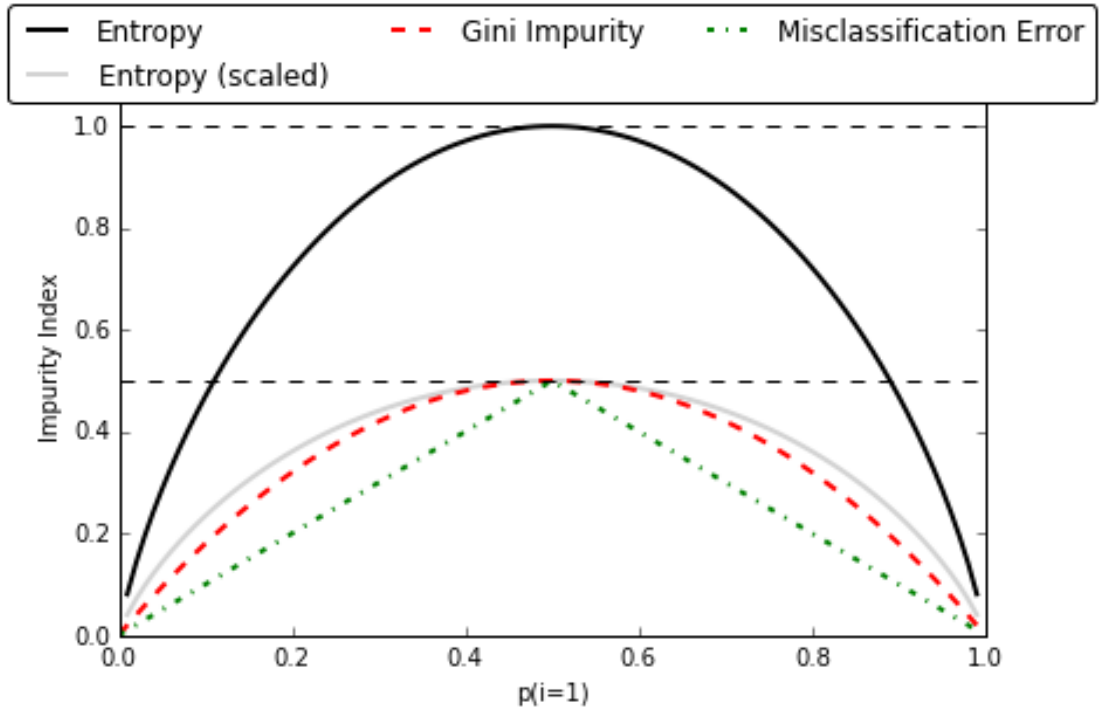
$$\phi(p) = \min(p, 1-p),$$

- Medición Entropia

$$\phi(p) = -p\log(p) - (1-p)\log(1-p),$$

- Medición Indice de diversidad Gini

$$\phi(p) = p(1-p),$$



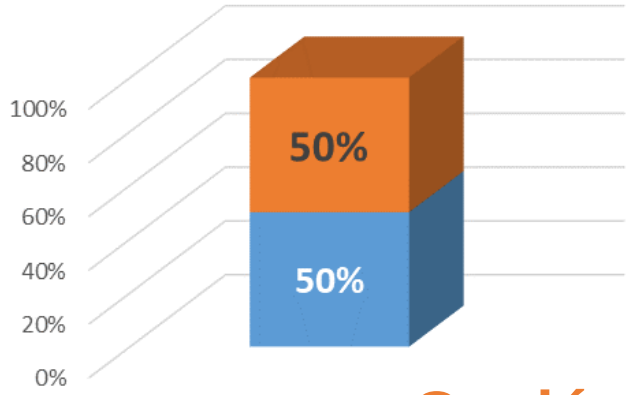
Impureza de 1 nodo

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i \mid t)]^2$$

ALGORTIMO ARBOL DE DECISION

Business Insight Center

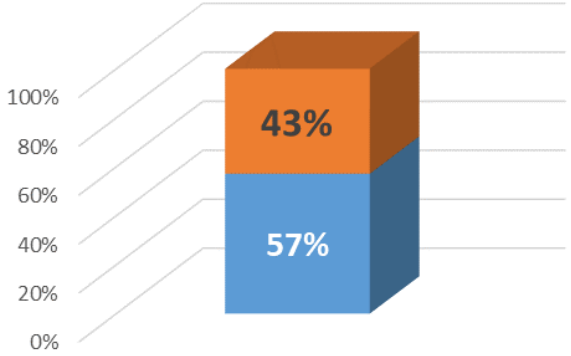
	Nro Cliente	Porcentaje
Compra	6	50%
No Compra	6	50%
	12	**100%**



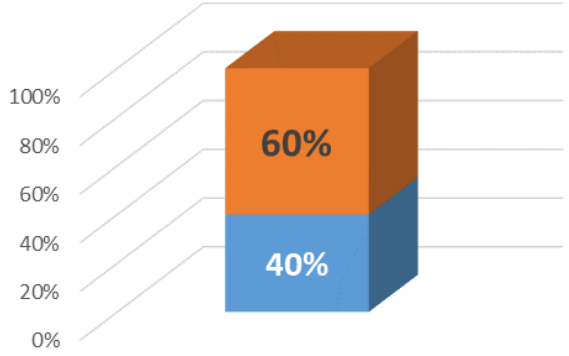
Variable sueldo

Opción de corte A
Sueldo>1500

7 clientes

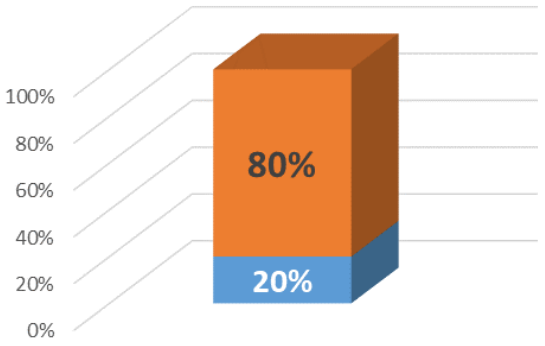


5 clientes

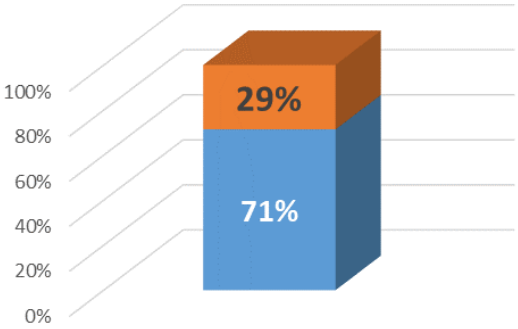


Opción de corte B
Sueldo >2500

5 clientes



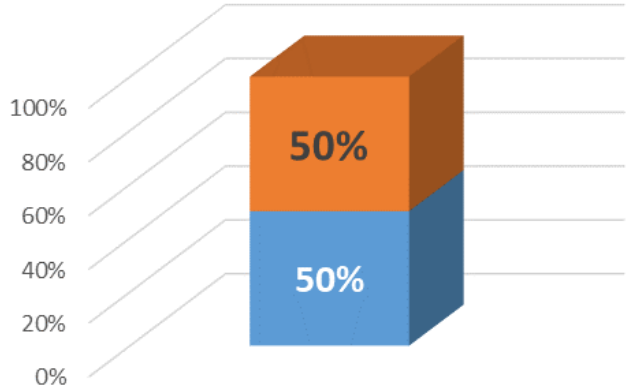
7 clientes



CRITERIO DE CORTE CON GINI

Business Insight Center

	Nro Cliente	Porcentaje
Compra	6	50%
No Compra	6	50%
	12	**100%**



$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i \mid t)]^2$$

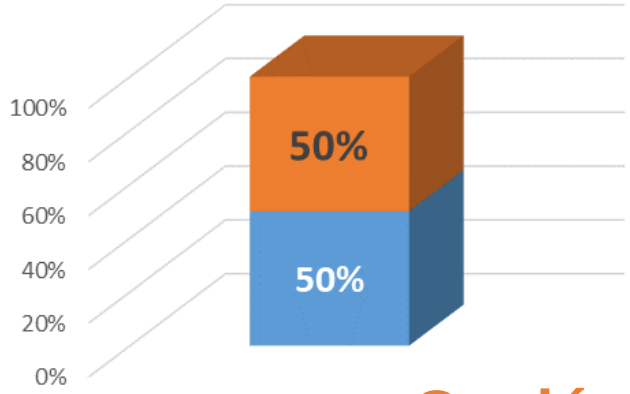
	Parent
C0	6
C1	6
Gini = 0.5	

Gini :
$1 - (6/12)^2 - (6/12)^2$
= 0.5

ALGORTIMO ARBOL DE DECISION

Business Insight Center

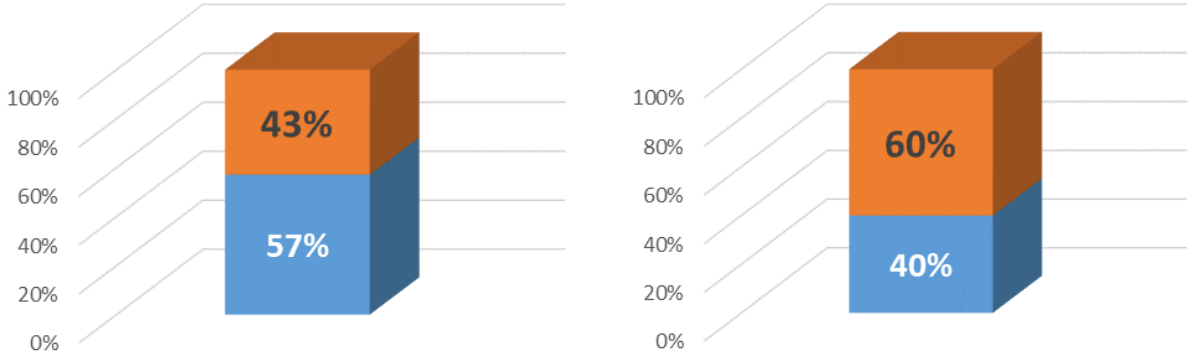
	Nro Cliente	Porcentaje
Compra	6	50%
No Compra	6	50%
	12	**100%**



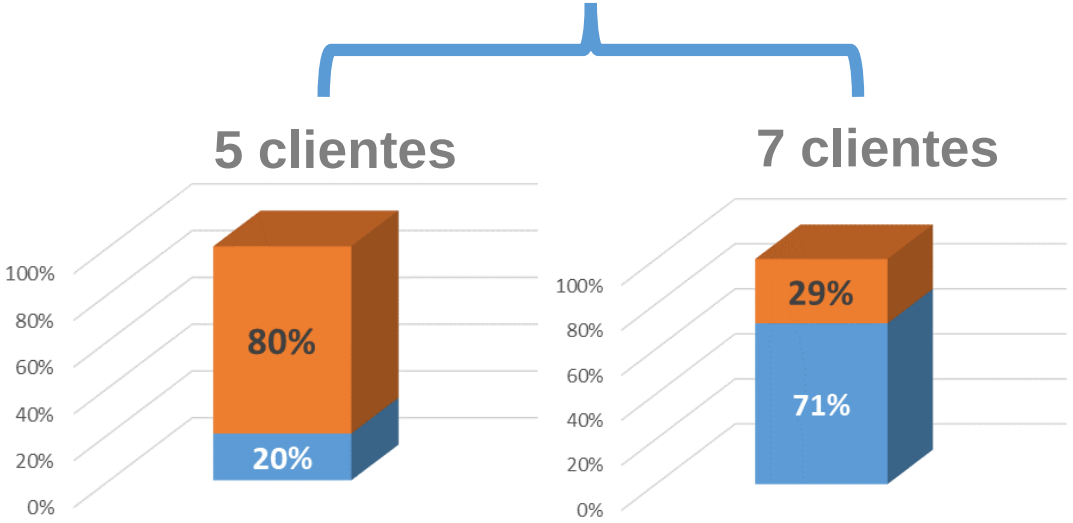
Gini :
$1 - (6/12)^2 - (6/12)^2$
= 0.5

Variable sueldo

Opción de corte A
Sueldo>1500



Opción de corte B
Sueldo >2500



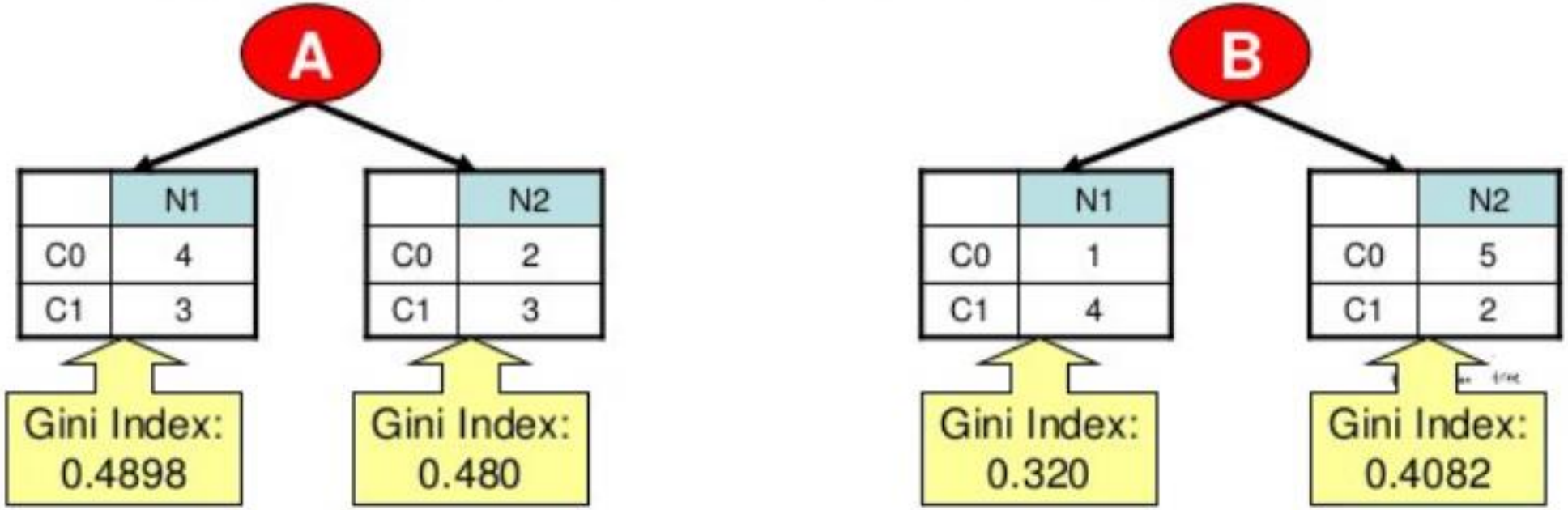
CRITERIO DE CORTE CON GINI

Example :

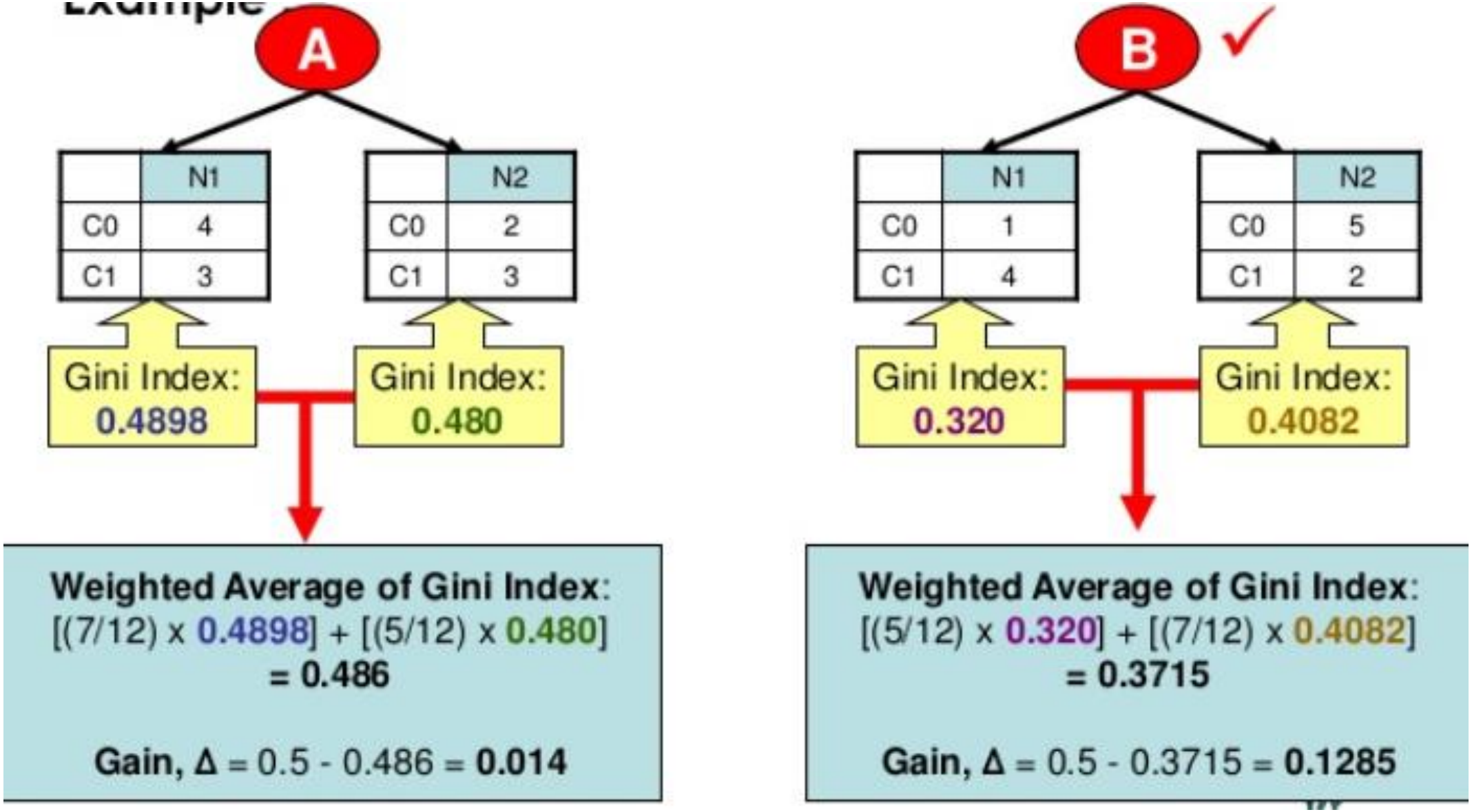
	Parent
C0	6
C1	6
Gini = 0.5	

Gini :
$1 - (6/12)^2 - (6/12)^2$
= 0.5

Suppose there are two ways (A and B) to split the data into smaller subset.



CRITERIO DE CORTE CON GINI

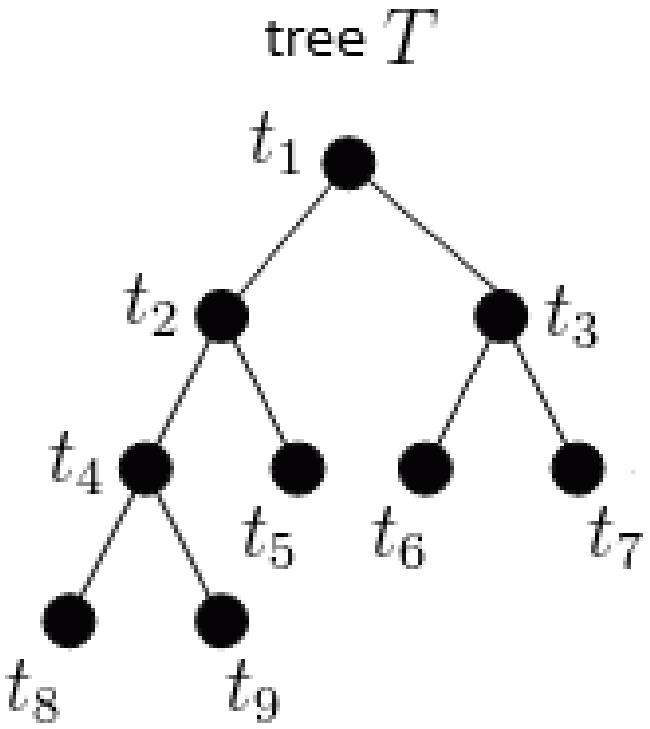


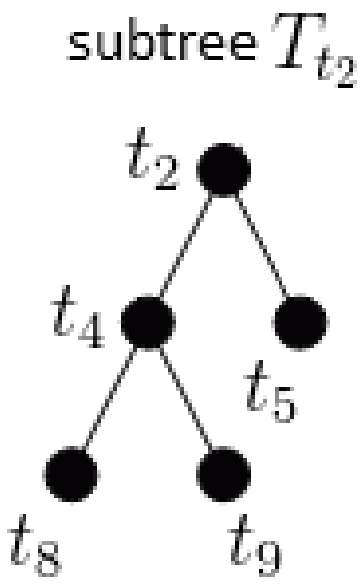
CRITERIO DE COMPLEJIDAD DEL ARBOL

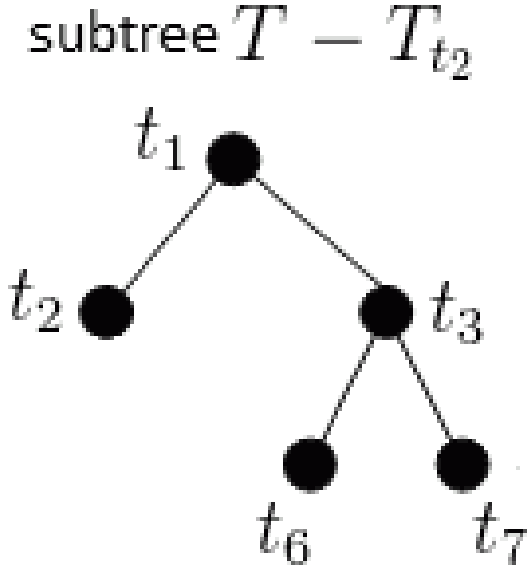
Business Insight Center

Tene un arbol que
En cada corte identifica
Grupos homogenous de clientes

Que tan complejo?







CRITERIO DE COMPLEJIDAD DEL ARBOL

Business Insight Center

Reducir el error

ERRO DE MALA CLASIFICACION

Nro. De nodos terminales

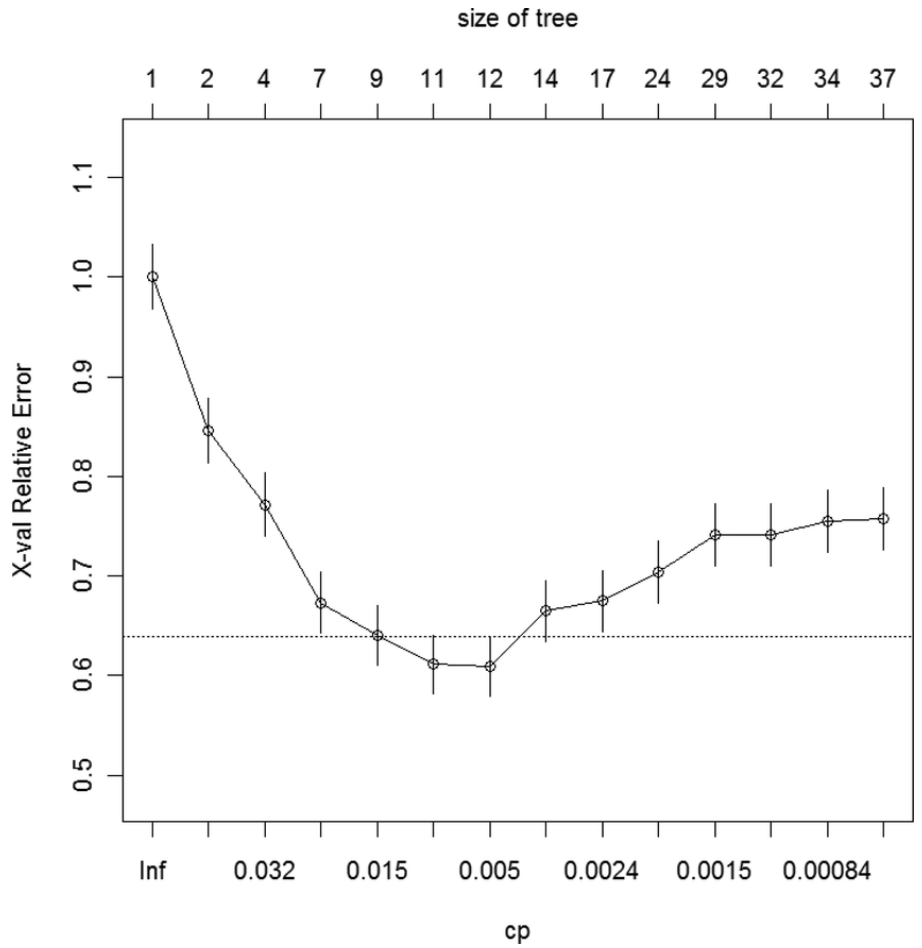
$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

CRITERIO DE COMPLEJIDAD DEL ARBOL

Business Insight Center

Profundidad del arbol

Tasas de Mala clasificación X-Val



Parametro complejidad

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

CROSS VALIDATION 5 FOLDS