

# **One shot Modelling 4 modelos Machine Learnig**

# AGENDA

- Revisión de caso de negocio venta de seguros
- Ranking de variables importantes
- Entendimiento comparativa de métodos de modelamiento
  - Regresión Logística
  - Árbol de Clasificación
  - Naive Bayes
  - KNN



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

# METODO DE ENSEÑANZA

CRISP DM

Muestreo

Ranking Var

Models

Estrategia  
Comercial

Not like this....



1



2



3



4

Like this!



1



2



3



4



5

Perfilamiento

One-shot

Modelling

Modelling

Estrategia  
Comercial

Modelling

Intermedio

Avanzado





# FOCO DEL PROBLEMA

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



**Cliente compra o no?**



$$P(A/B) =$$



**Conozco características del cliente**

Reg. Logística  
Naive Bayes  
KNN  
SVM  
Redes Neuro

Arboles de Decisión  
Random Forest  
Gradient Boosting

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$$

Diagram illustrating the components of a linear regression model:

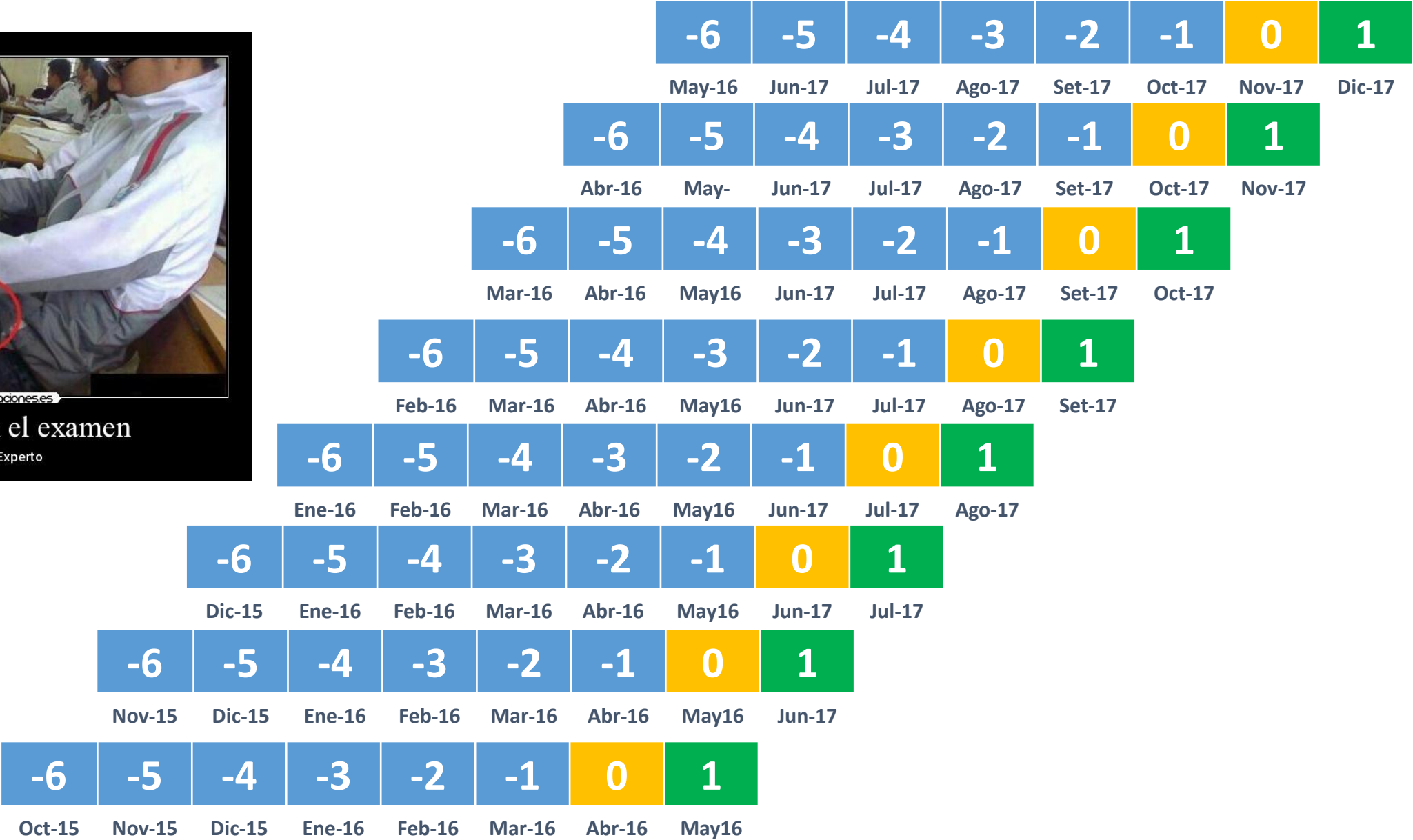
- Coefficients:**  $\beta_0, \beta_1, \beta_2, \dots, \beta_n$
- Explanatory Variables:**  $X_1, X_2, \dots, X_n$
- Random Error Term/Residuals:**  $\epsilon$

- R1: IF (age>38.5) AND (years-in-job>2.5) THEN  $y = 0.8$
- R2: IF (age>38.5) AND (years-in-job≤2.5) THEN  $y = 0.6$
- R3: IF (age≤38.5) AND (job-type='A') THEN  $y = 0.4$
- R4: IF (age≤38.5) AND (job-type='B') THEN  $y = 0.3$
- R5: IF (age≤38.5) AND (job-type='C') THEN  $y = 0.2$



**Reducir el error**

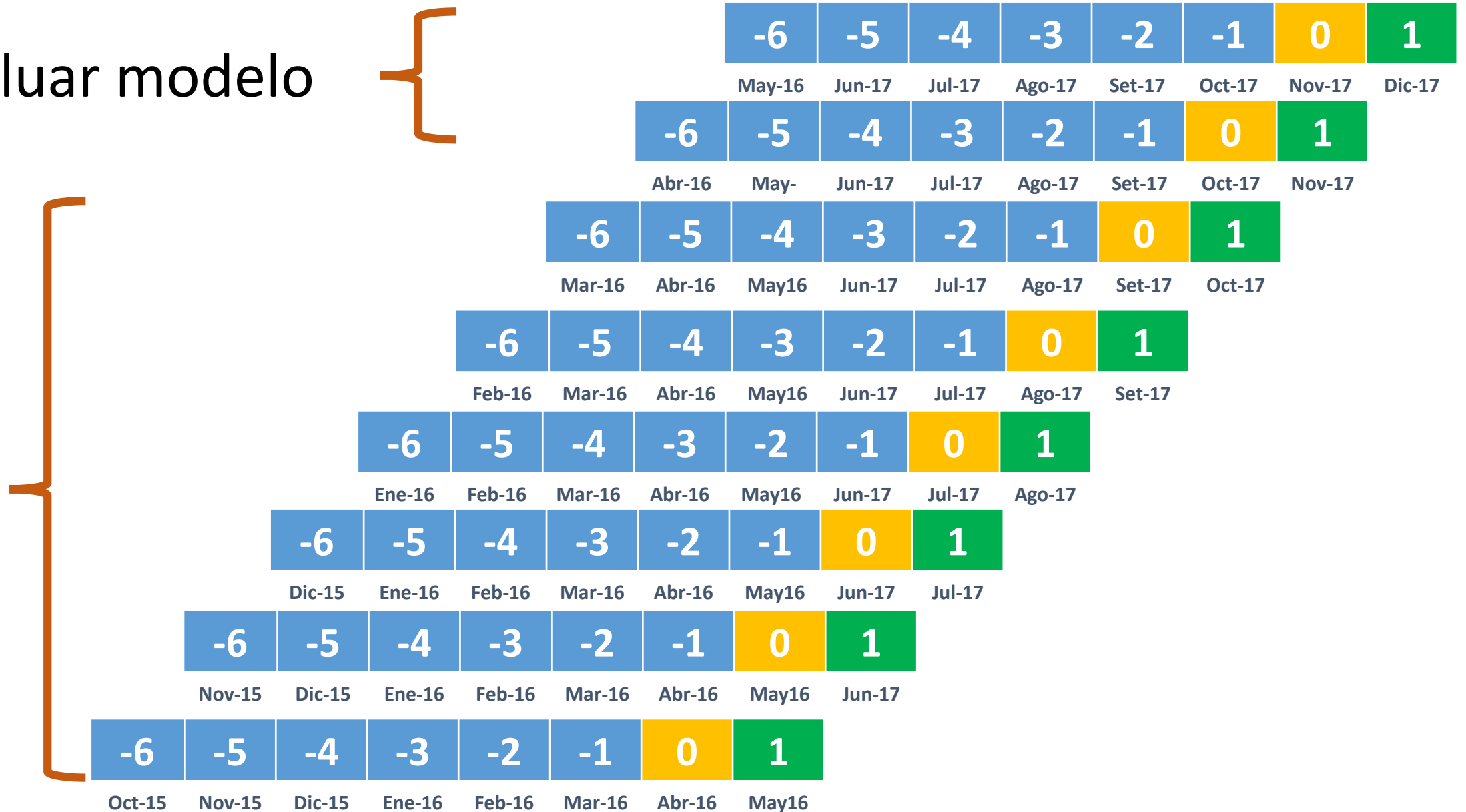
# APRENDER DE TODA LA HISTORIA?



# APRENDER DE TODA LA HISTORIA?

Evaluar modelo

Aprender y  
probar

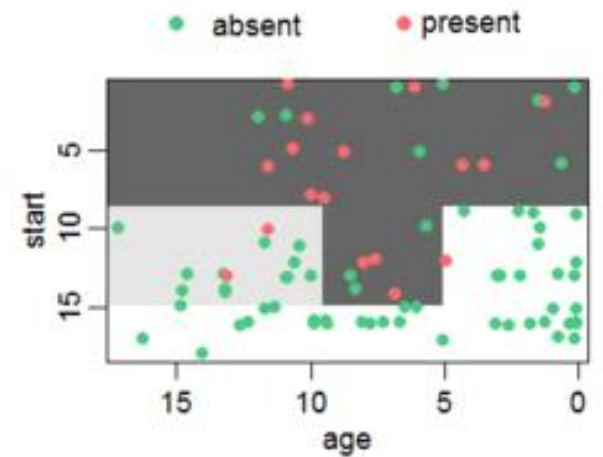
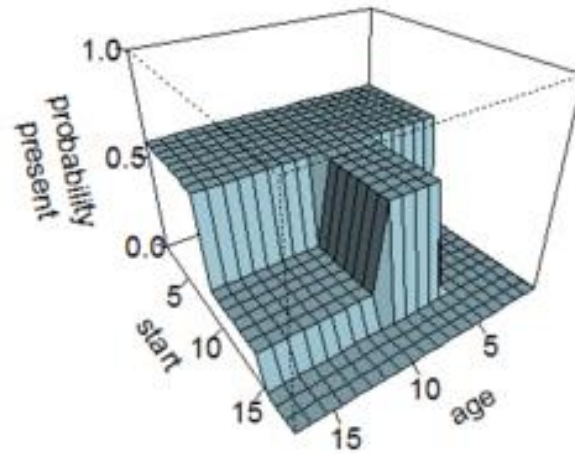
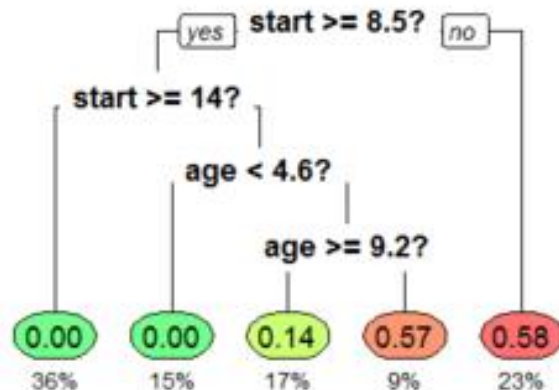


# ALGORITMO ARBOL DE DECISION

Por qué  
Árbol  
Clasificación?

- ✓ Simple visualización y explicación
- ✓ Identificación de relaciones no lineales
- ✓ No requiere revisar supuestos

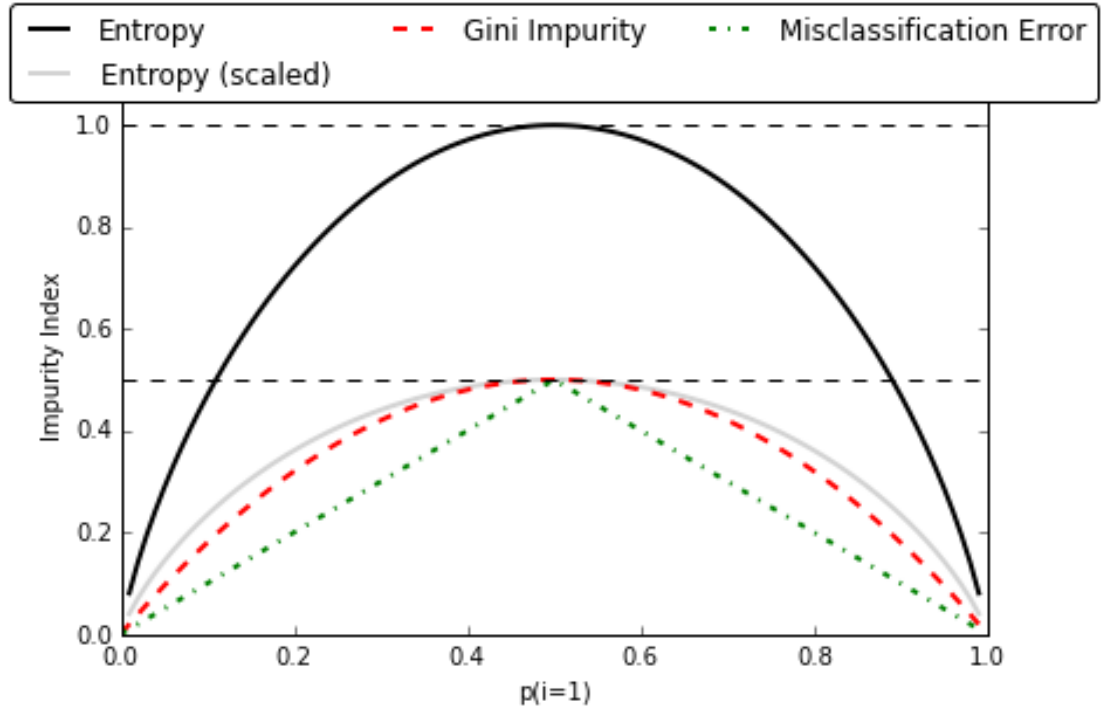
Qué es  
Árbol de  
Clasificación?



# ALGORITMO ARBOL DE DECISION

## Ganar un mayor indicar de impureza

- Entropía
- Diversidad de Gini
- Error de clasificación

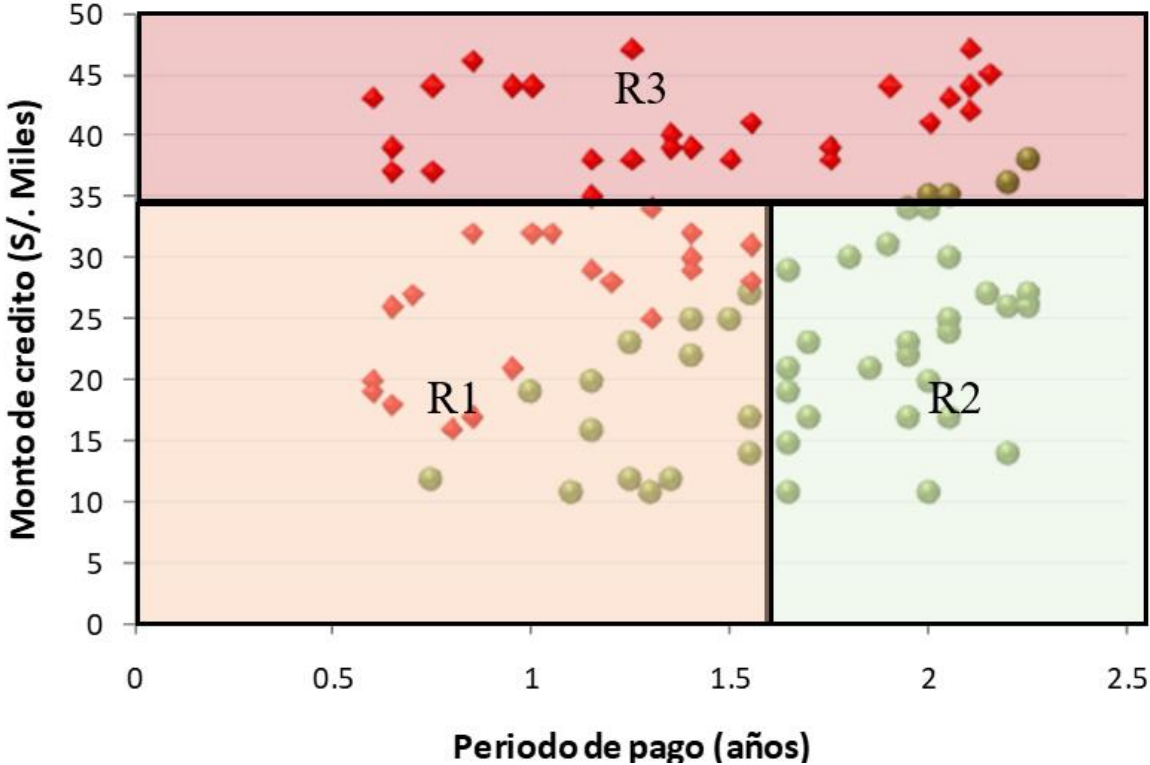




# ALGORITMO ARBOL DE DECISION



Resultado de creditos otorgados



**Impureza inicial: 0.25**

Clientes morosos: 46 (51%)

Tamaño de muestra: 100%

**Bondad de Ajuste de cortes:**

$$= 0.25 - (30\% \cdot 0.07 + 0\% \cdot 0 + 40\% \cdot 0.24)$$

$$= 0.13$$

**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%

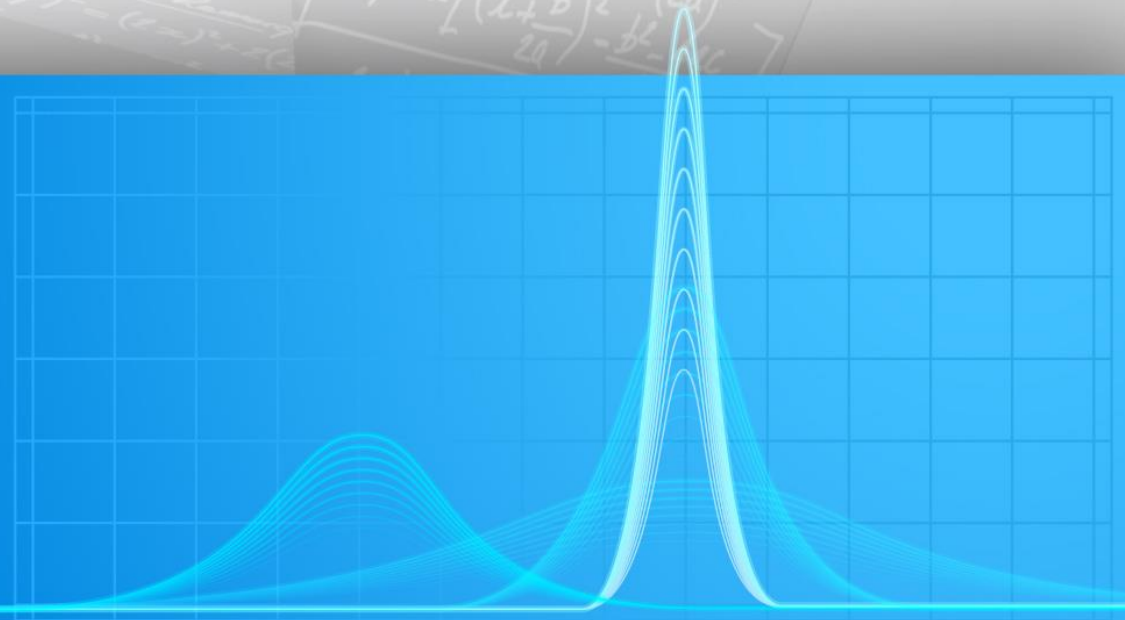
**Impureza R3: 0.24**

Clientes morosos: 21 (58%)

Tamaño de muestra: 40%



## Ranking de variables con Random Forest



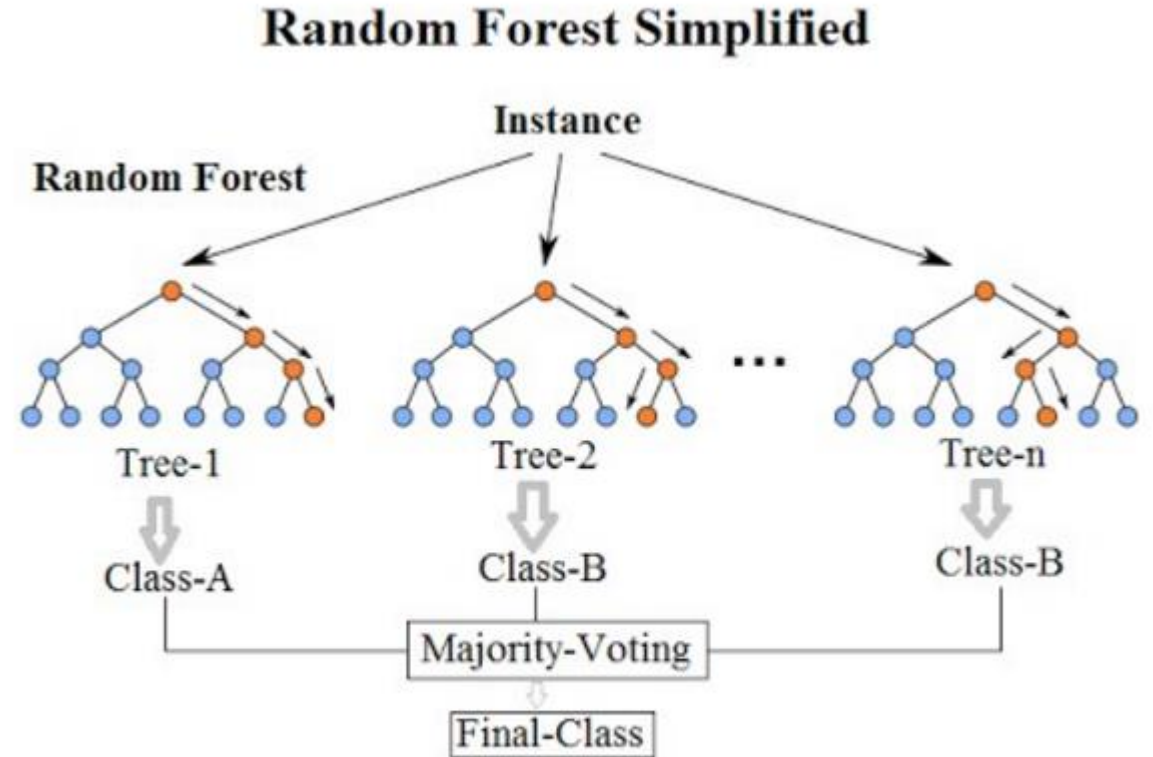
# ALGORITMO RANDOM FOREST

Por qué  
Random Forest?

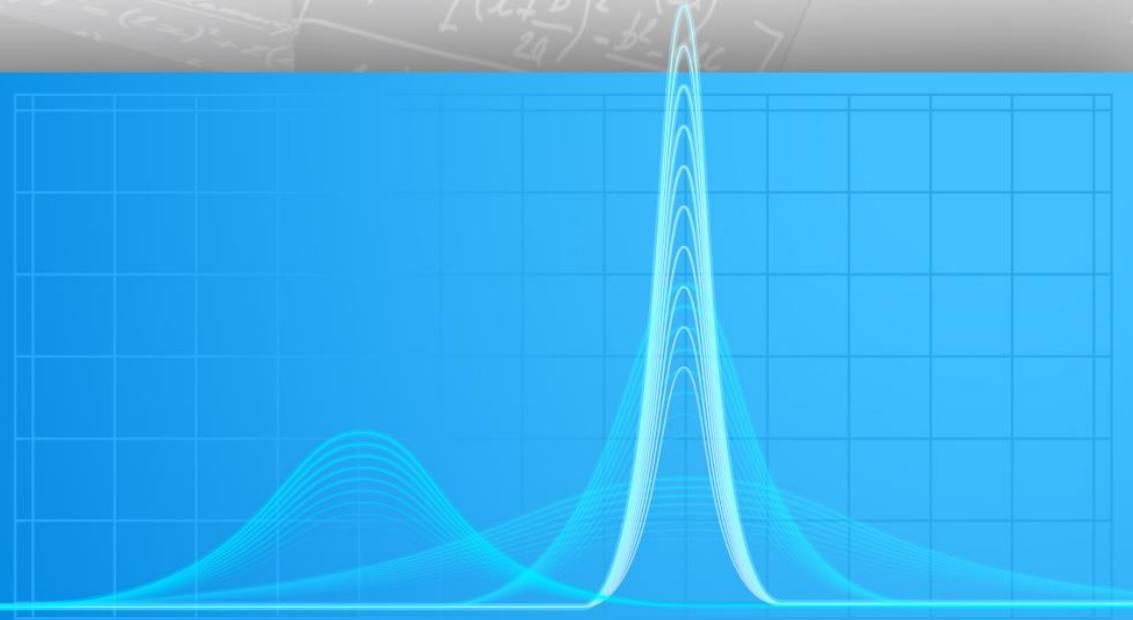
Es más confiable o  
robusto al examinar  
los datos en  
diferentes arboles y  
llegar a un consenso

Qué es  
Random Forest?

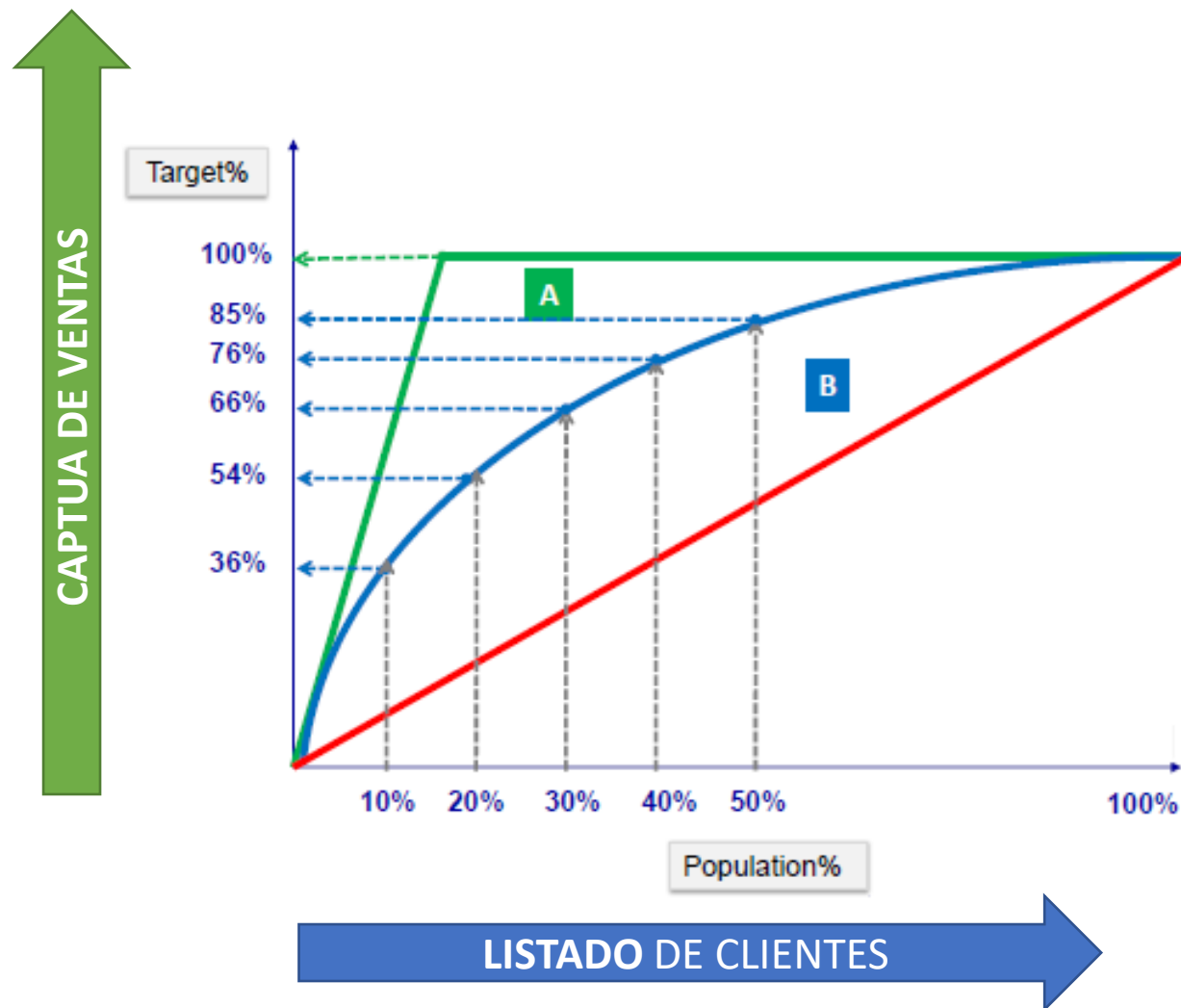
Es el aprendizaje  
consensuado de  
muchos arboles  
variando las muestras  
de aprendizaje



## Entendimiento y comparativa de métodos de modelamiento



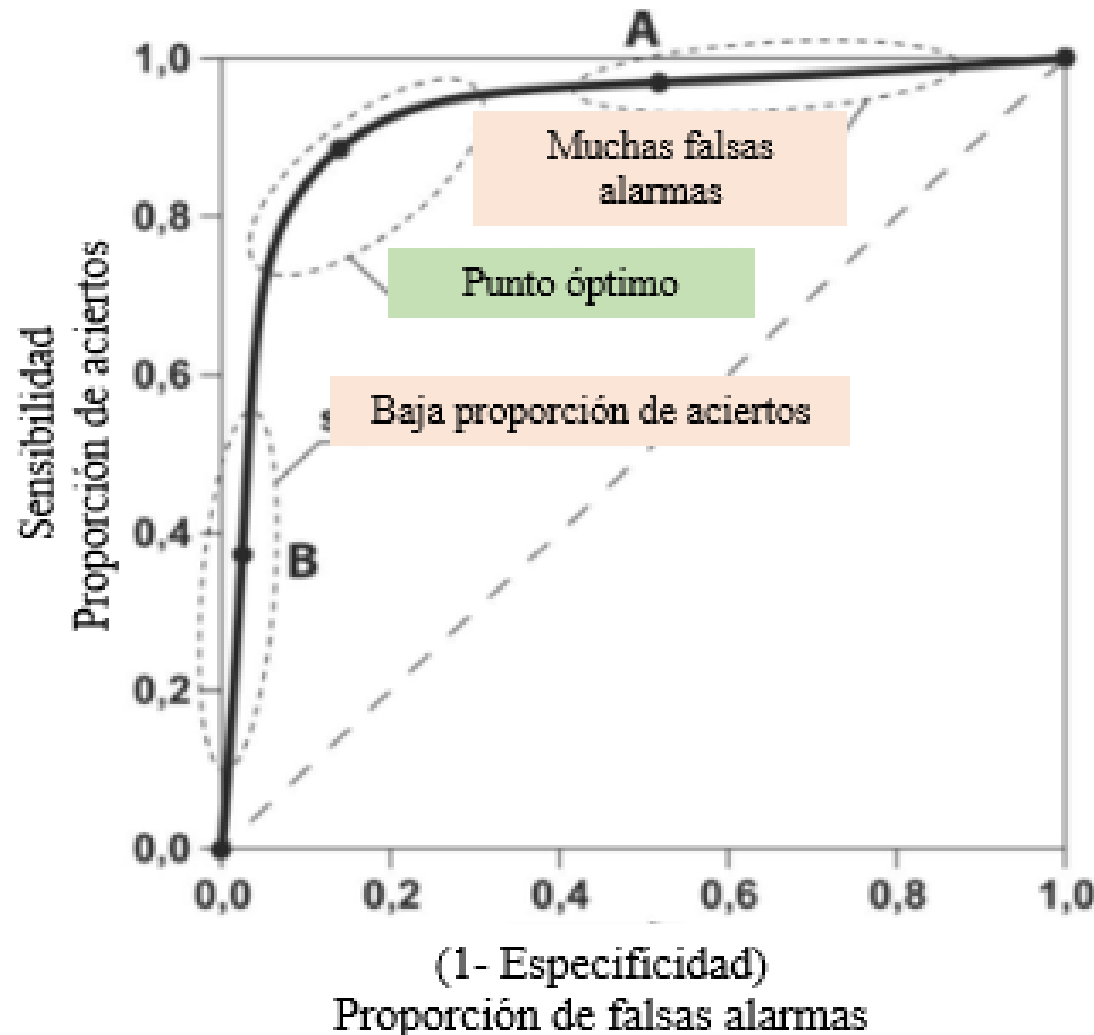
## CURVA DE GANANCIAS



Curva que refleja la contribución de modelo predictivo en la detección de las ventas



## AUC (ÁREA UNDER CURVE)

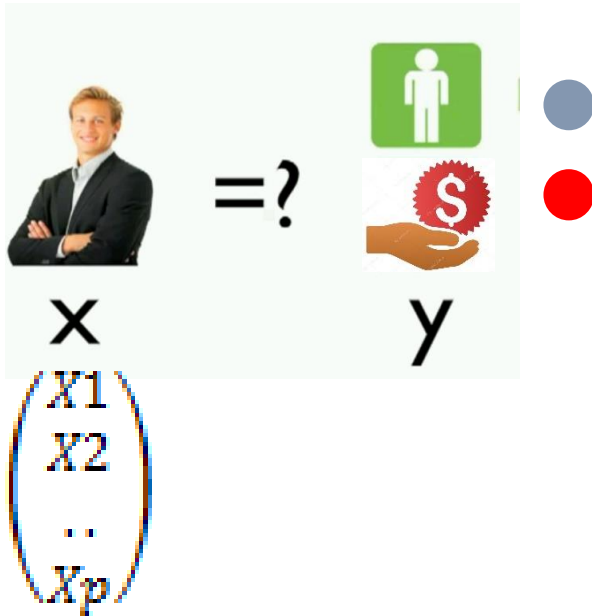


Probabilidad de que el modelo acierte pronosticando las ventas y no cometa falsas alarmas

AUC	Poder predictivo
1.0 o 100%	Predicción perfecta!
$\geq 90\%$	Excelente
$\geq 80\%$ y $< 90\%$	Bueno
$\geq 70\%$ y $< 80\%$	Aceptable
$\geq 60\%$ y $< 70\%$	Regular
50%	Aleatorio

# ALGORITMO REGRESIÓN LOGÍSTICA

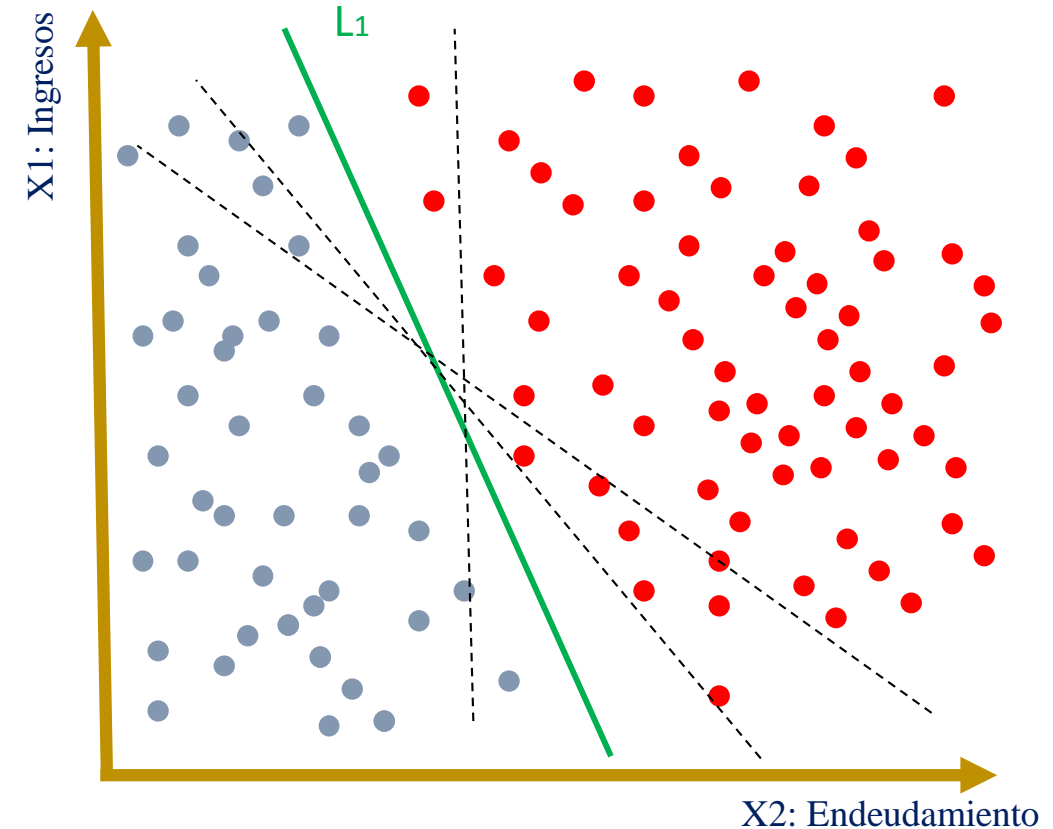
El problema



Data histórica



Representación gráfica



Para el ejemplo visual definamos :

Variable Y

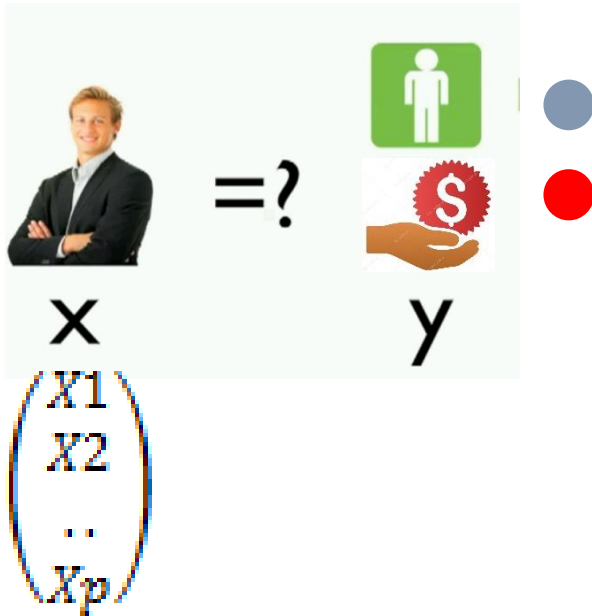


$X_1$  Ingresos mensuales  $\begin{pmatrix} 245 \\ 3450 \end{pmatrix} \begin{pmatrix} 345 \\ 4532 \end{pmatrix} \begin{pmatrix} 234 \\ 3452 \end{pmatrix} \begin{pmatrix} 45 \\ 1234 \end{pmatrix} \cdots \begin{pmatrix} 345 \\ 1232 \end{pmatrix} \begin{pmatrix} 345 \\ 123 \end{pmatrix}$   
 $X_2$  Endeudamient SSFF

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

# ALGORITMO REGRESIÓN LOGÍSTICA

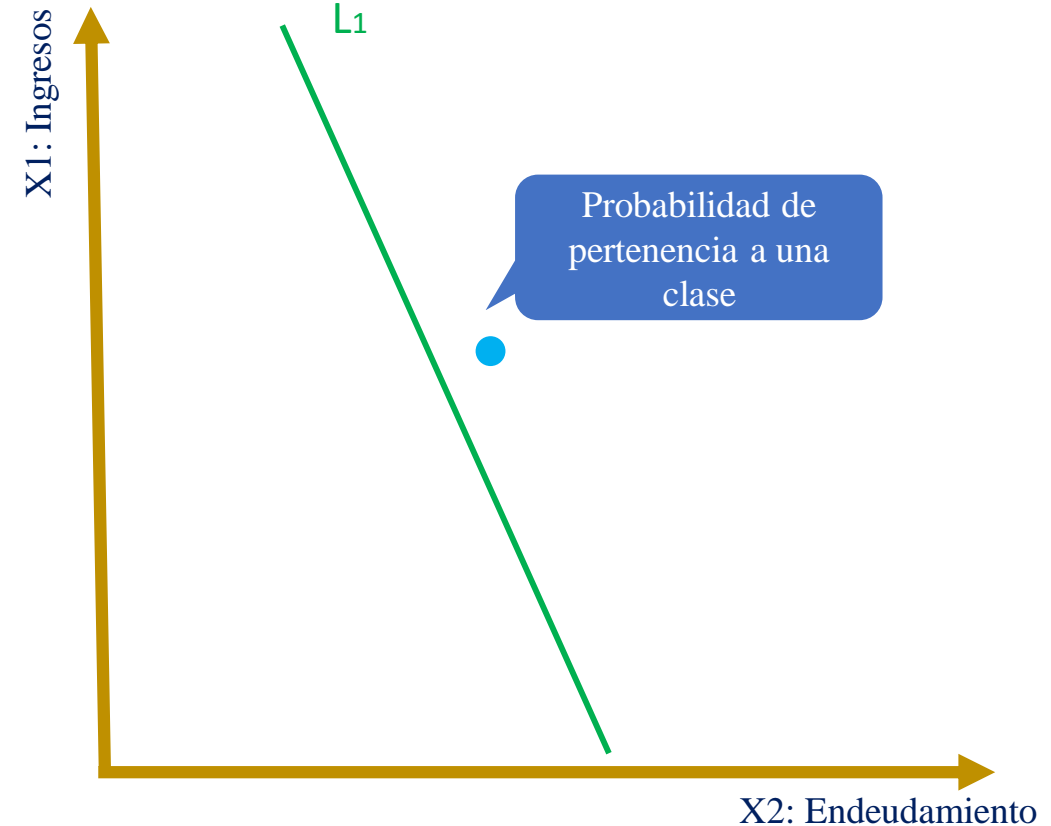
## El problema



## Data histórica



## Representación gráfica



Para el ejemplo visual definamos :

Variable Y



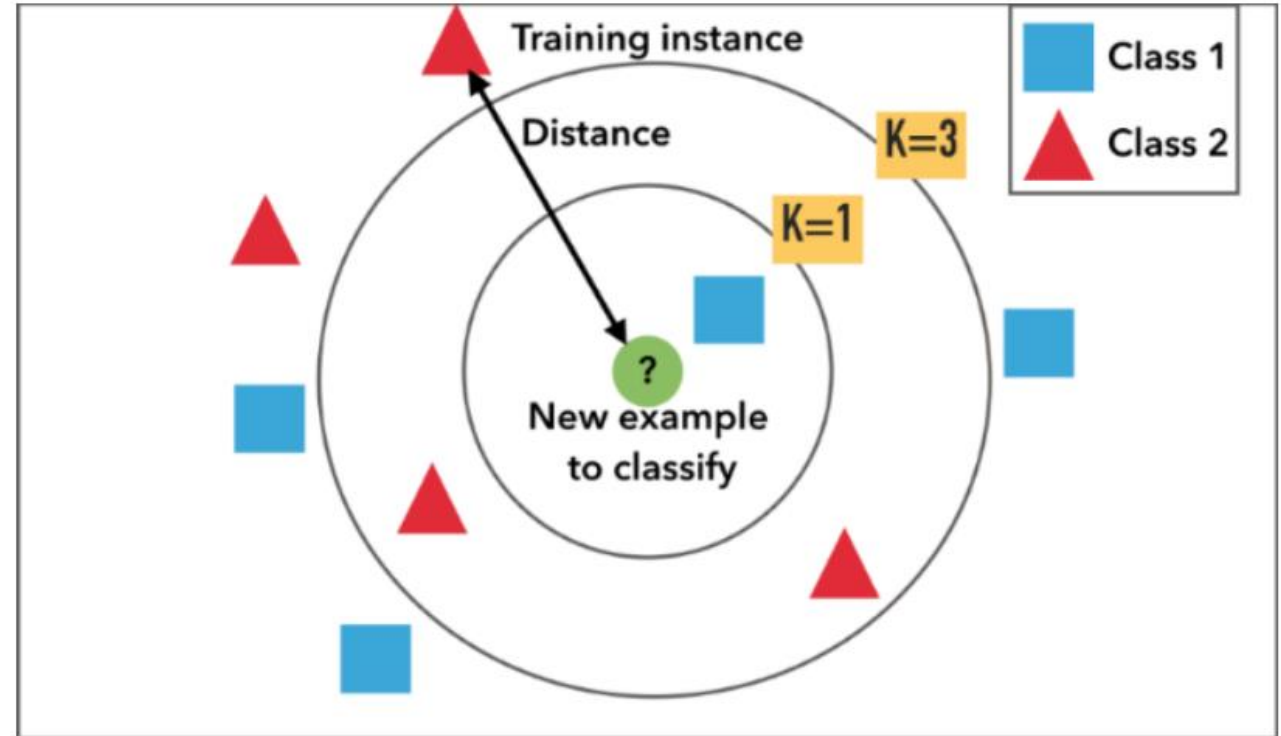
$X_1$	Ingresos mensuales	$\begin{pmatrix} 245 \\ 3450 \end{pmatrix}$	$\begin{pmatrix} 345 \\ 4532 \end{pmatrix}$	$\begin{pmatrix} 234 \\ 3452 \end{pmatrix}$	$\begin{pmatrix} 45 \\ 1234 \end{pmatrix}$	$\cdots$	$\begin{pmatrix} 345 \\ 1232 \end{pmatrix}$	$\begin{pmatrix} 345 \\ 123 \end{pmatrix}$
$X_2$	Endeudamiento SSFF							

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

# ALGORITMO KNN (K NEAREST NEIGHBOR)



No soy Lazy!  
Solo estoy altamente motivado a no hacer nada



$$\text{Euclidean distance : } d(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

$$\text{Squared Euclidean distance : } d(x, y) = \sum (x_i - y_i)^2$$

$$\text{Manhattan distance : } d(x, y) = \sum |x_i - y_i|$$

# ALGORITMO NAIVE BAYES

Algoritmo que supone que todas las características son independientes entre sí



**Cliente  
compra o no?**



**Conozco  
características  
del cliente**



Likelihood

Prior

$$P(Y|X_1, \dots, X_n) = \frac{P(X_1, \dots, X_n|Y)P(Y)}{P(X_1, \dots, X_n)}$$

Normalization  
Constant



# ALGORITMO NAIVE BAYES

Algoritmo que supone que todas las características son independientes entre sí



**Cliente  
compra o no?**



% Clientes con esa  
característica que  
han comprado

% Clientes que  
han comprado



$$P(Y|X_1, \dots, X_n) = \frac{P(X_1, \dots, X_n|Y)P(Y)}{P(X_1, \dots, X_n)}$$


% Clientes con esa  
característica

**Conozco  
características  
del cliente**



# ALGORITMO NAIVE BAYES

	 outlook	 temperature	 humidity	 windy	 play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

**OR**  
**#YAY OR #NAY**  
**CAN USE NAIVE BAYES**  
**ALGORITHM**  
**P(C) OR P(CLASS)**  
**P(YES) = 9 / 14**  
**P(NO) = 5 / 14**