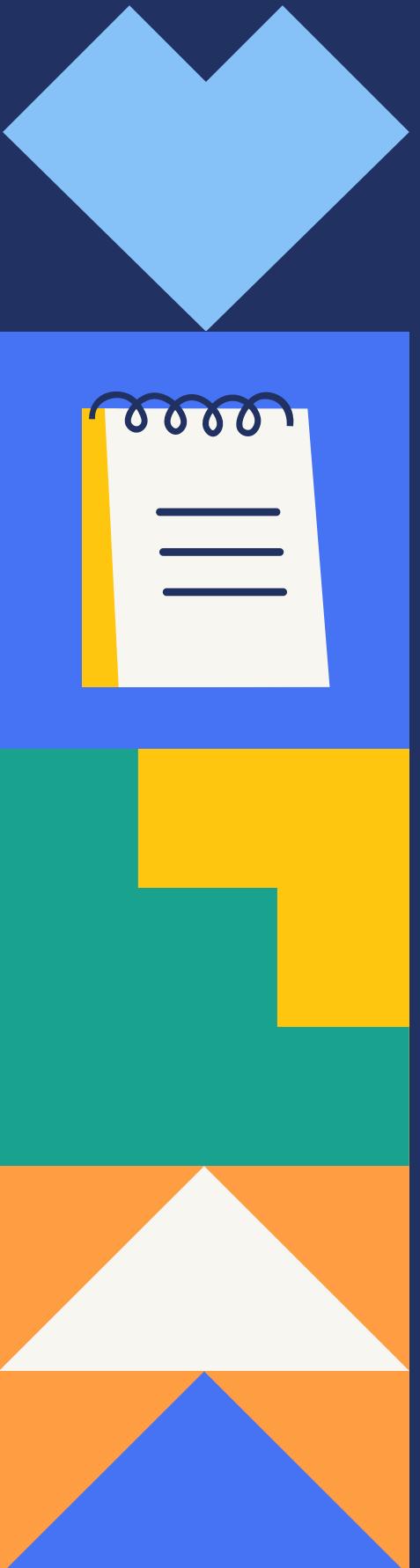


# Fundamentos de Probabilidad y Estadística

para Inteligencia Artificial





# Estructura de la sesión



## Estadística Descriptiva

- Medidas de Tendencia Central
- Medidas de Dispersion
- Correlación
- Ejemplos y aplicaciones para IA



## Teoría de Probabilidad

- Conceptos Clave
- Probabilidad Conjunta, Marginal y Condicional
- Distribuciones de Probabilidad
- Inferencia Probabilística
- Ejemplos

¿Por qué...

necesitamos probabilidad y estadística en IA?



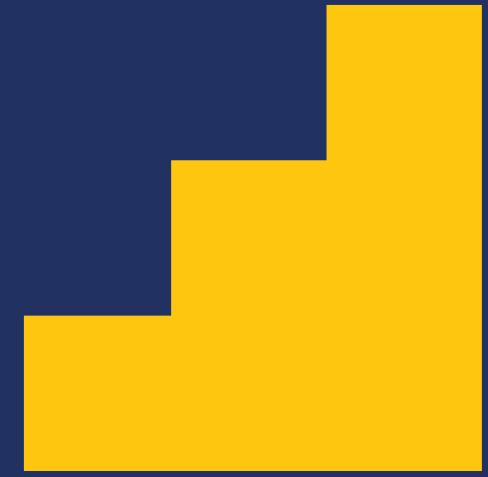
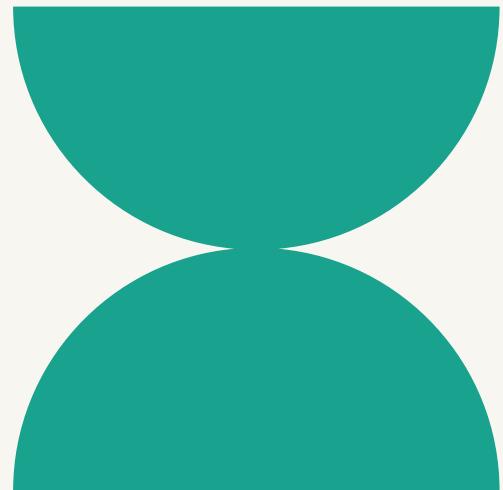
## Incertidumbre

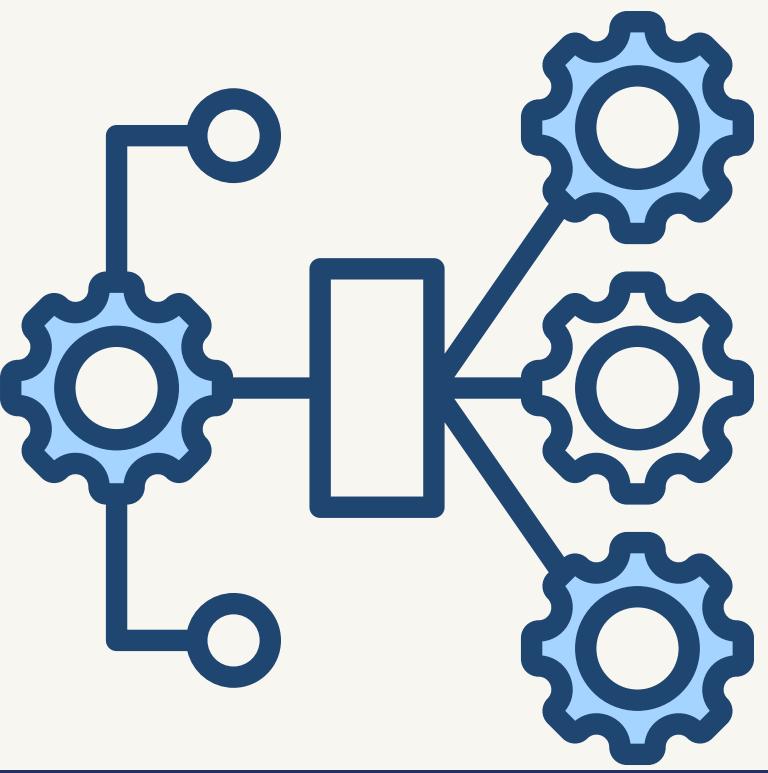
Modelar la  
Incertidumbre y la  
Toma de Decisiones



## Estadística

Analizar Datos y  
Extraer Patrones





¿Por qué...

necesitamos probabilidad y estadística en IA?



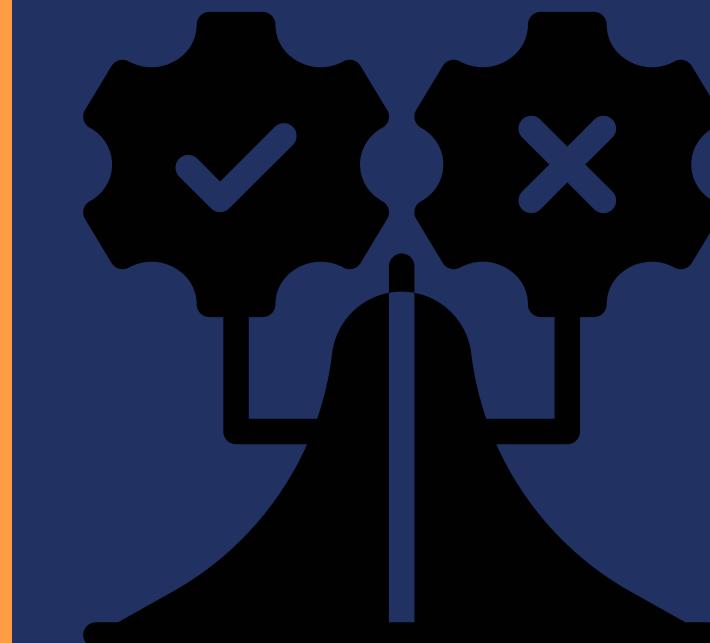
## Probabilidad

Algoritmos como  
clasificadores, Redes  
Neuronales  
Artificiales y Árboles  
de Decisión



## Evaluar Modelos

¿Qué Tan Buenas Son  
Nuestras  
Predicciones?



# ¿Cómo la probabilidad y la estadística nos ayudan a tomar mejores decisiones con datos?

Imagina que sales de casa... No tienes certeza absoluta de si lloverá, pero puedes usar datos y probabilidades para tomar la mejor decisión.



*Diferencia entre probabilidad (incertidumbre en eventos) y estadística (análisis de datos).*



# ¿Debo llevar paraguas hoy?



## Datos históricos y patrones climáticos

- ◆ ¿Cuántos días al año llueve en mi ciudad?
- ◆ ¿Llueve más en ciertas estaciones o meses?
- ◆ Si ayer llovió, ¿qué tan probable es que hoy también llueva?



## Probabilidad condicional

- Supongamos que:
- El pronóstico del tiempo dice que hay un 80% de probabilidad de lluvia.
  - El cielo está nublado y hay alta humedad.

Decisión	¿Llovió?	Consecuencia
Llevaste paraguas	Si	No te mojas
Llevaste paraguas	No	Cargas un paraguas innecesariamente
No llevaste paraguas	Sí	Te mojas y arruinas tu ropa
No llevaste paraguas	No	No Cargas Peso Extra

# Estadística Descriptiva

es una rama de la estadística que se encarga de organizar, resumir y representar datos numéricos

## Medidas de Tendencia Central

- Media, mediana y moda.

## Medidas de Dispersion

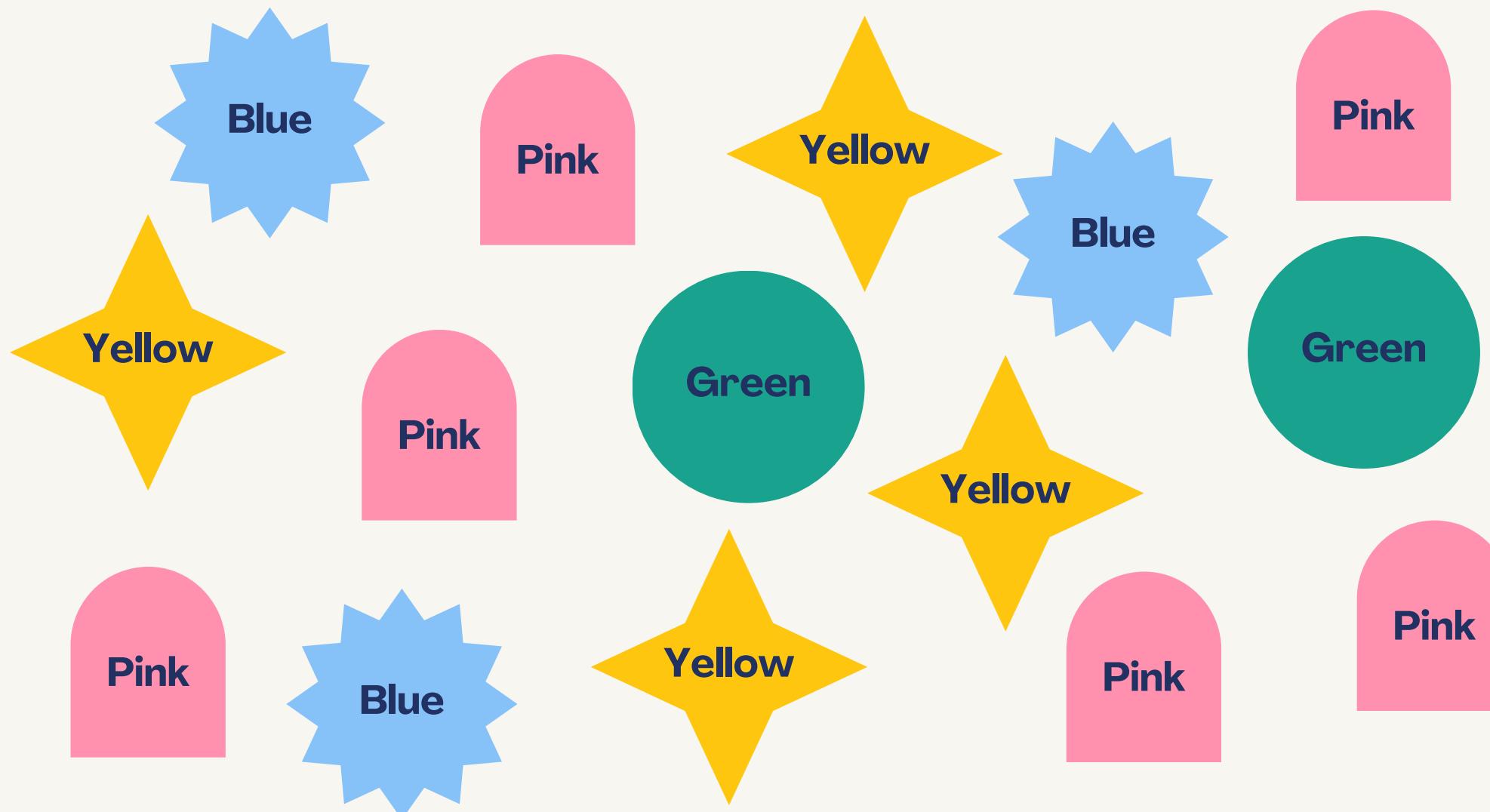
- Varianza y desviación estandar

## Correlación y Covarianza

- Concepto de correlación: Relación entre variables.

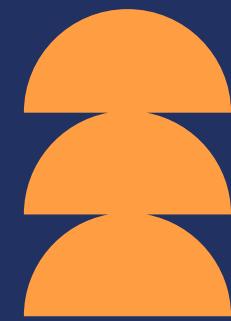
# Votación de Colores Favoritos

Supongamos que realizamos una encuesta a 100 personas sobre su color favorito y obtenemos los siguientes resultados:

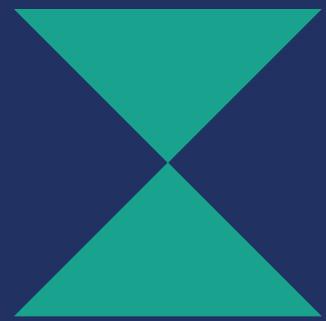


¿Cómo analizar y organizar los resultados?

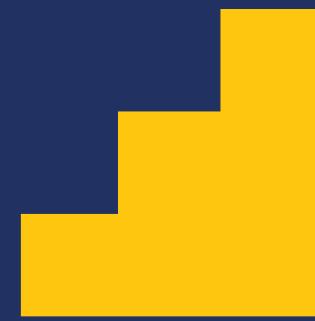
# Preguntas a Resolver



¿Cuál es el color más popular?



¿Cuántos votos tiene un color en promedio?



¿Cuánta variabilidad hay entre los votos?

Estos mismos principios se aplican en Machine Learning para analizar datos antes de entrenar modelos

## Análisis Exploratorio de Datos

Ahora aplicamos la estadística descriptiva para analizar estos datos:

Color	Número de personas
Azul	35
Rojo	20
Verde	15
Amarillo	10
Negro	10
Blanco	10



Which color is the mode?



## Medidas de Tendencia Central:

- Media
- Mediana
- Moda

## Medidas de Dispersión:

- Rango
- Varianza
- Desviación

Color	Número de personas
Azul	35
Rojo	20
Verde	15
Amarillo	10
Negro	10
Blanco	10

# Ejemplo Varianza

## Tiempos de entrega en dos sucursales de una empresa

Imagina que dos sucursales de una empresa entregan paquetes, y queremos analizar la rapidez en la entrega.

### Sucursal 1:

Tiempos en días: 5, 5, 5, 5, 5

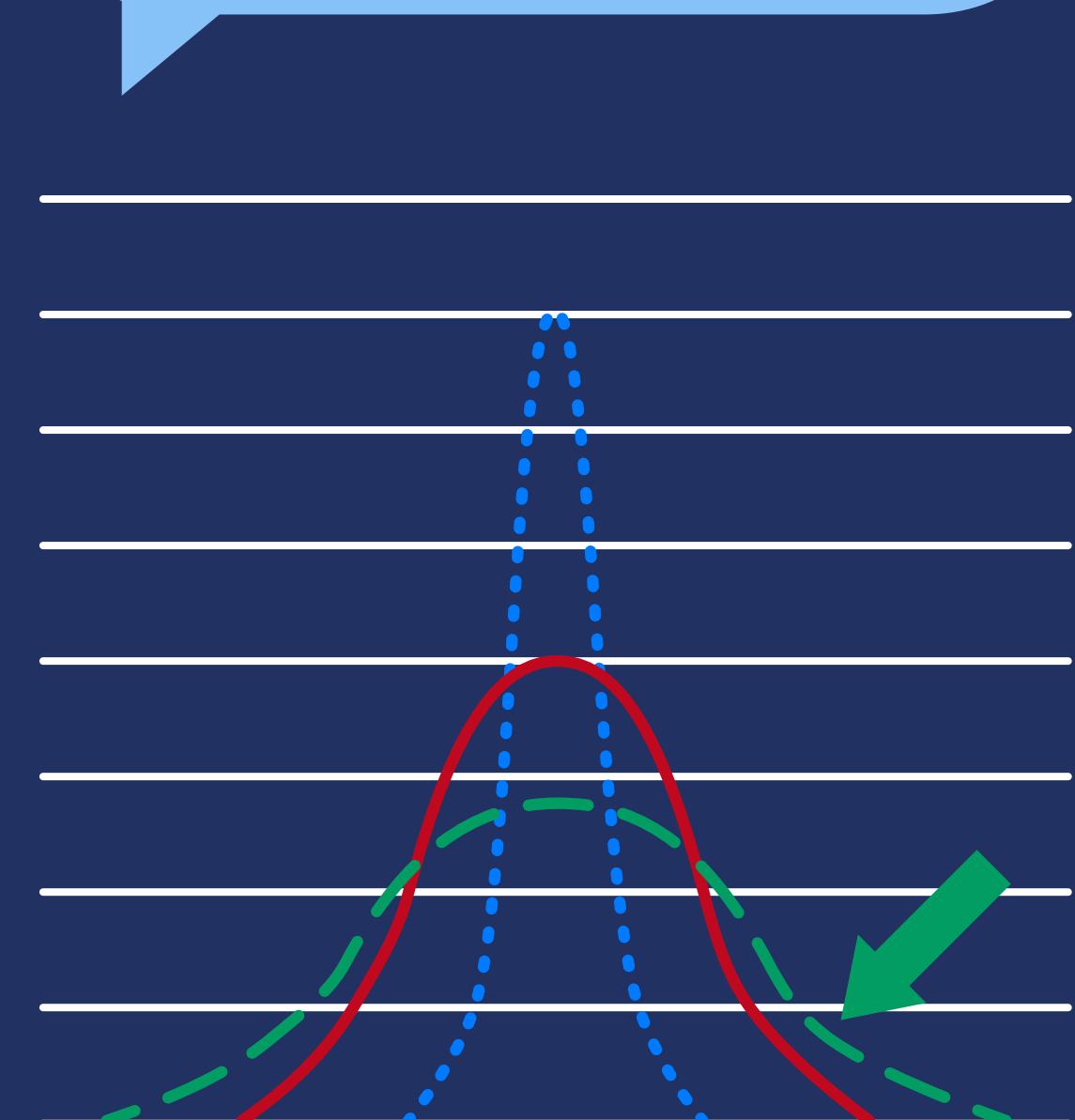
- Media = 5 días
- Varianza = 0 (todas las entregas son en el mismo tiempo)

### Sucursal 2:

Tiempos en días: 1, 3, 5, 7, 9

- Media = 5 días
- Varianza > 0 (algunas entregas son muy rápidas y otras tardan más)

¡Datasets con la misma media pero diferente dispersión!



Supongamos que, además de preguntar el color favorito, también preguntamos la edad de los encuestados y obtenemos la siguiente tabla:

Color Favorito	Número de Votos (X)	Edad Promedio (Y)
Azul	35	25
Rojo	20	30
Verde	15	22
Amarillo	10	35
Negro	10	28
Blanco	10	40



## Relación entre variables

¿Existe una relación entre los votos y la edad?

La covarianza nos indica si existe una relación entre los votos y la edad.

Correlación

¿Qué tan fuerte es la relación?

Estos cálculos ayudan a entender qué variables están relacionadas en un dataset, lo cual es útil para seleccionar características

# Covarianza

Si  $\text{Cov}(X, Y) > 0$ , significa que a mayor número de votos, mayor edad promedio.

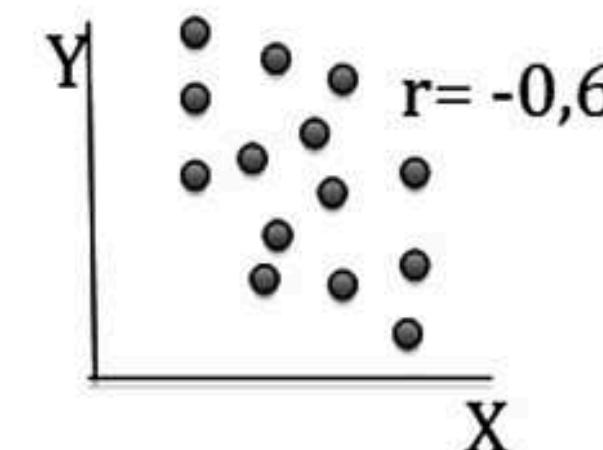
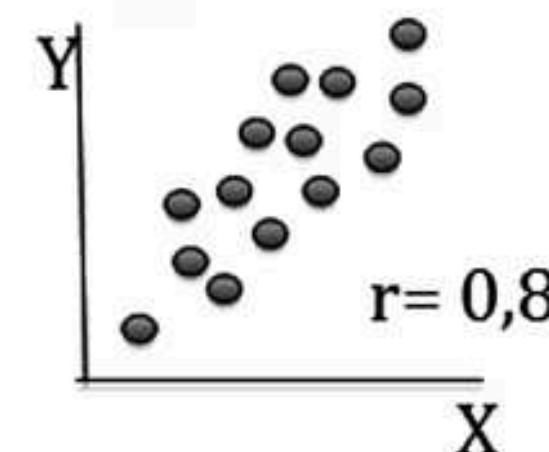
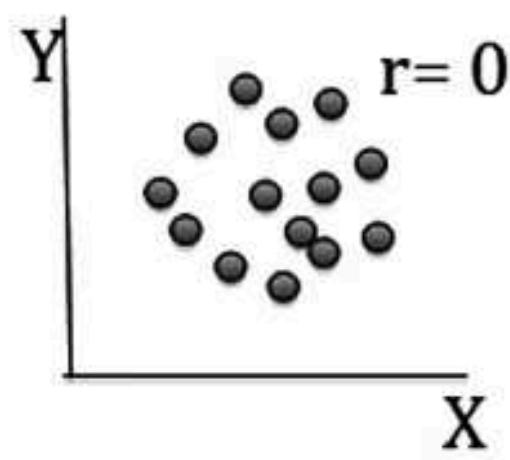
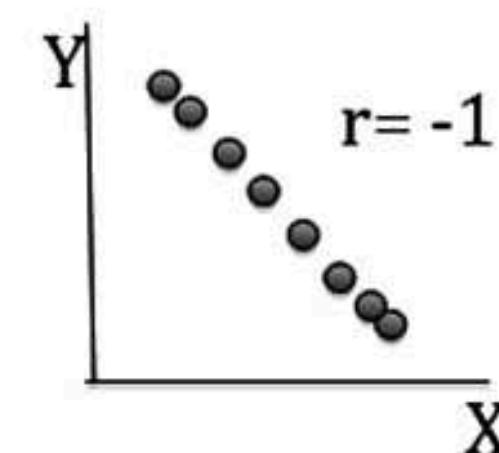
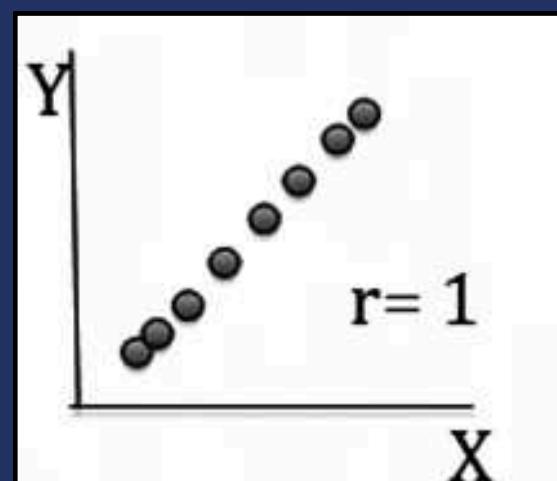
Si  $\text{Cov}(X, Y) < 0$ , significa que a mayor número de votos, menor edad promedio.

Si  $\text{Cov}(X, Y) \approx 0$ , no hay relación entre los votos y la edad.

# Correlación de Pearson

- ✓  $r > 0$
- ✗  $r < 0$

El resultado estará en el rango [-1, 1]:  
Correlación positiva (cuando sube X, sube Y)  
Correlación negativa (cuando sube X, baja Y)  
 $r \approx 0$  No hay relación

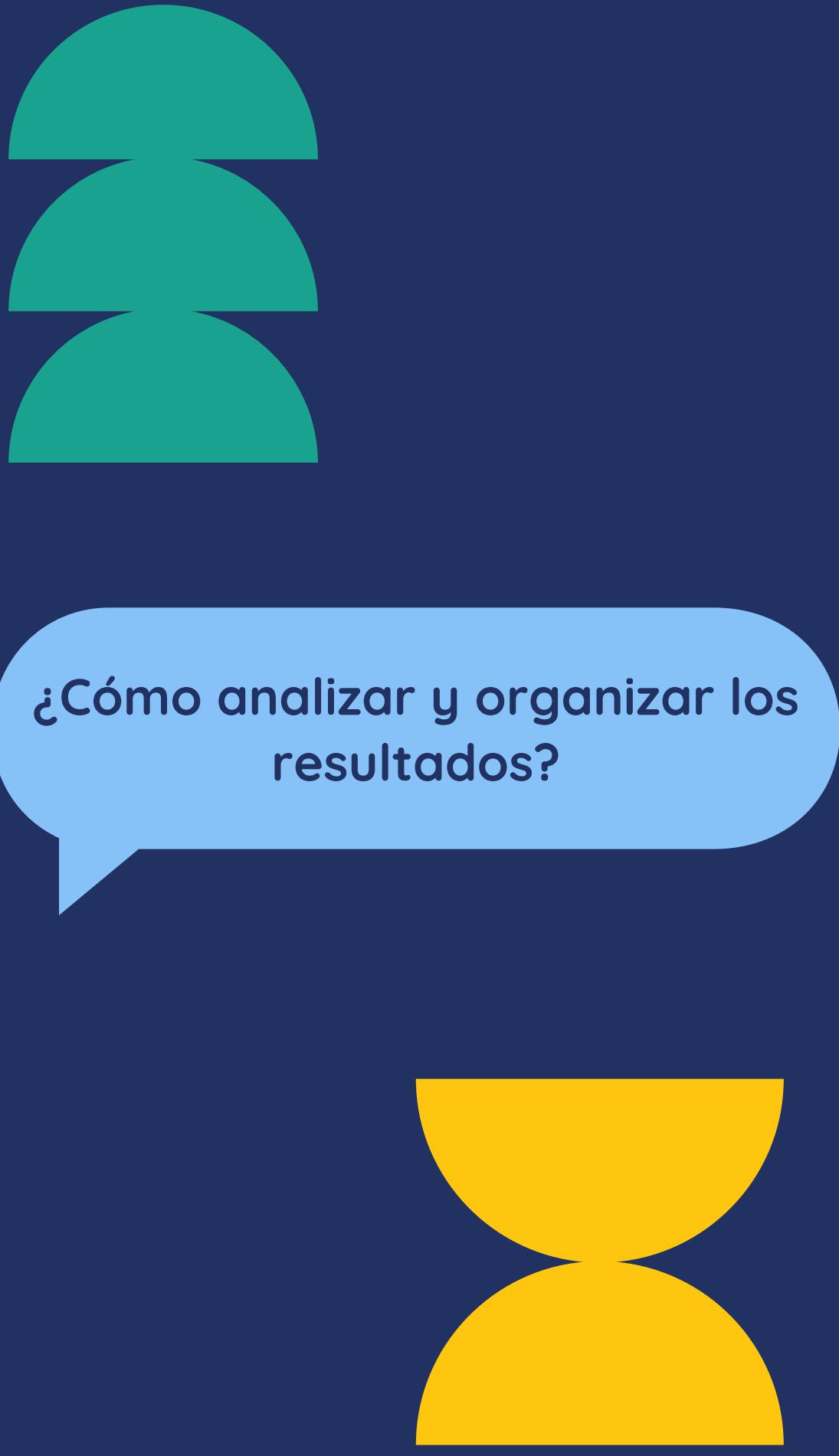


# Con el ejemplo, ¿Cómo interpretar los valores?

- Si encontramos una correlación fuerte entre edad y popularidad de un color, podríamos predecir qué colores son más atractivos para diferentes grupos de edad.
- Si la correlación es débil, significa que los gustos por los colores son independientes de la edad.

Por lo tanto...

- La covarianza indica la dirección de la relación, pero no la intensidad.
- La correlación de Pearson nos da tanto la dirección como la fuerza de la relación de forma estandarizada.



## Presenting Results



# Fundamentos de Probabilidad

La probabilidad compuesta o conjunta es el cálculo de la probabilidad cuando un experimento de probabilidad simple se repite varias veces o se relaciona un experimento con otro.

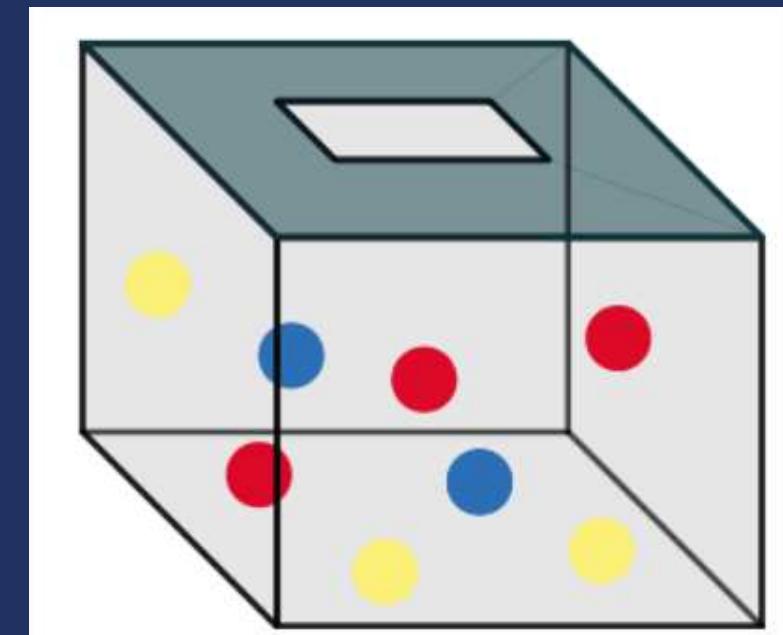
# Sucesos Independientes vs Sucesos Dependientes

## **Sucesos independientes:**

Dos sucesos son independientes si la ocurrencia de uno no afecta la probabilidad de que ocurra el otro. En otras palabras, el hecho de que uno suceda no influye en la probabilidad del otro.

## **Sucesos dependientes:**

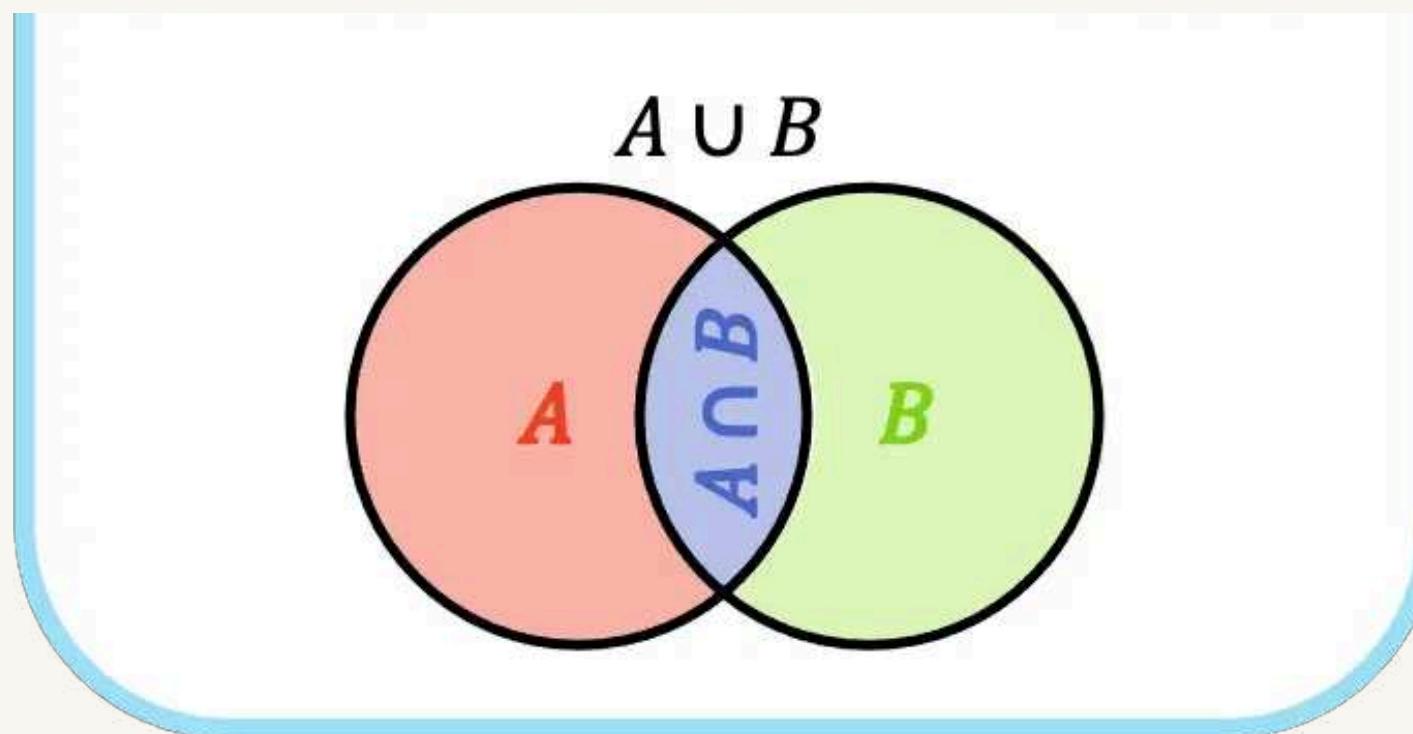
Dos sucesos son dependientes si la ocurrencia de uno afecta la probabilidad de que ocurra el otro. En este caso, la probabilidad de un suceso cambia dependiendo de si el otro suceso ha ocurrido o no.



## Reglas fundamentales de probabilidad

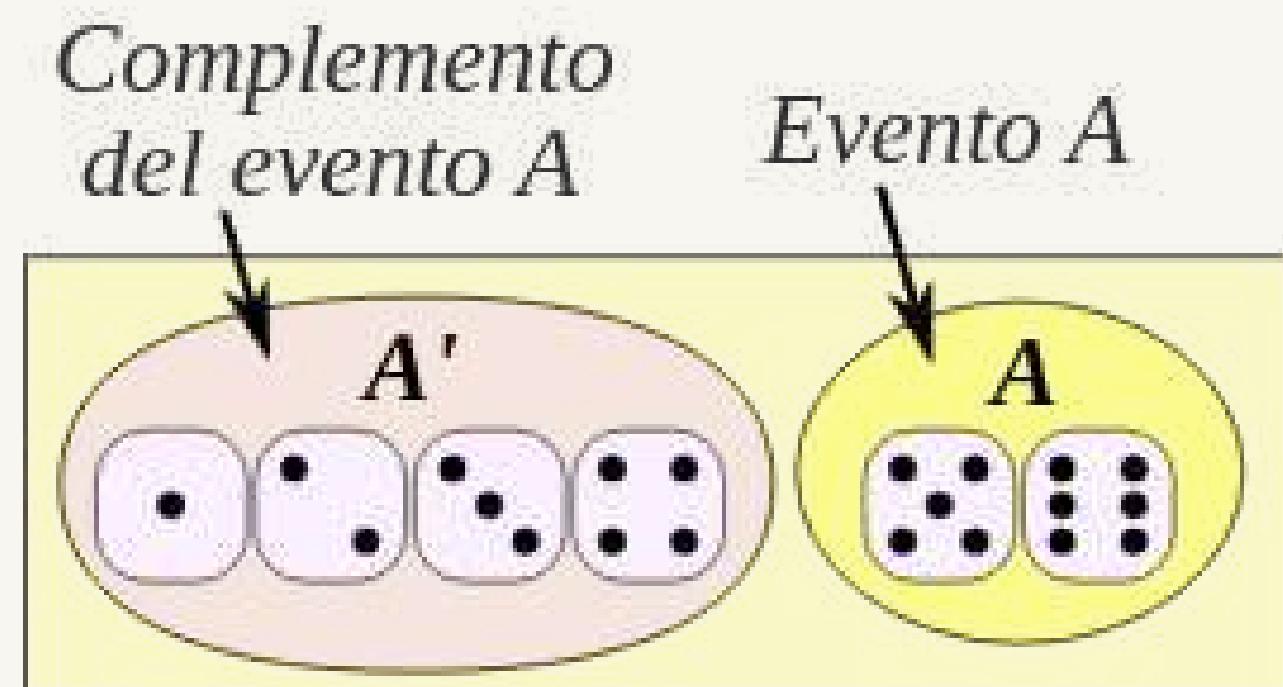
### ◆ Regla de la suma:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



## Reglas fundamentales de probabilidad

- ◆ Regla del complemento:  
 $P(A^c) = 1 - P(A)$

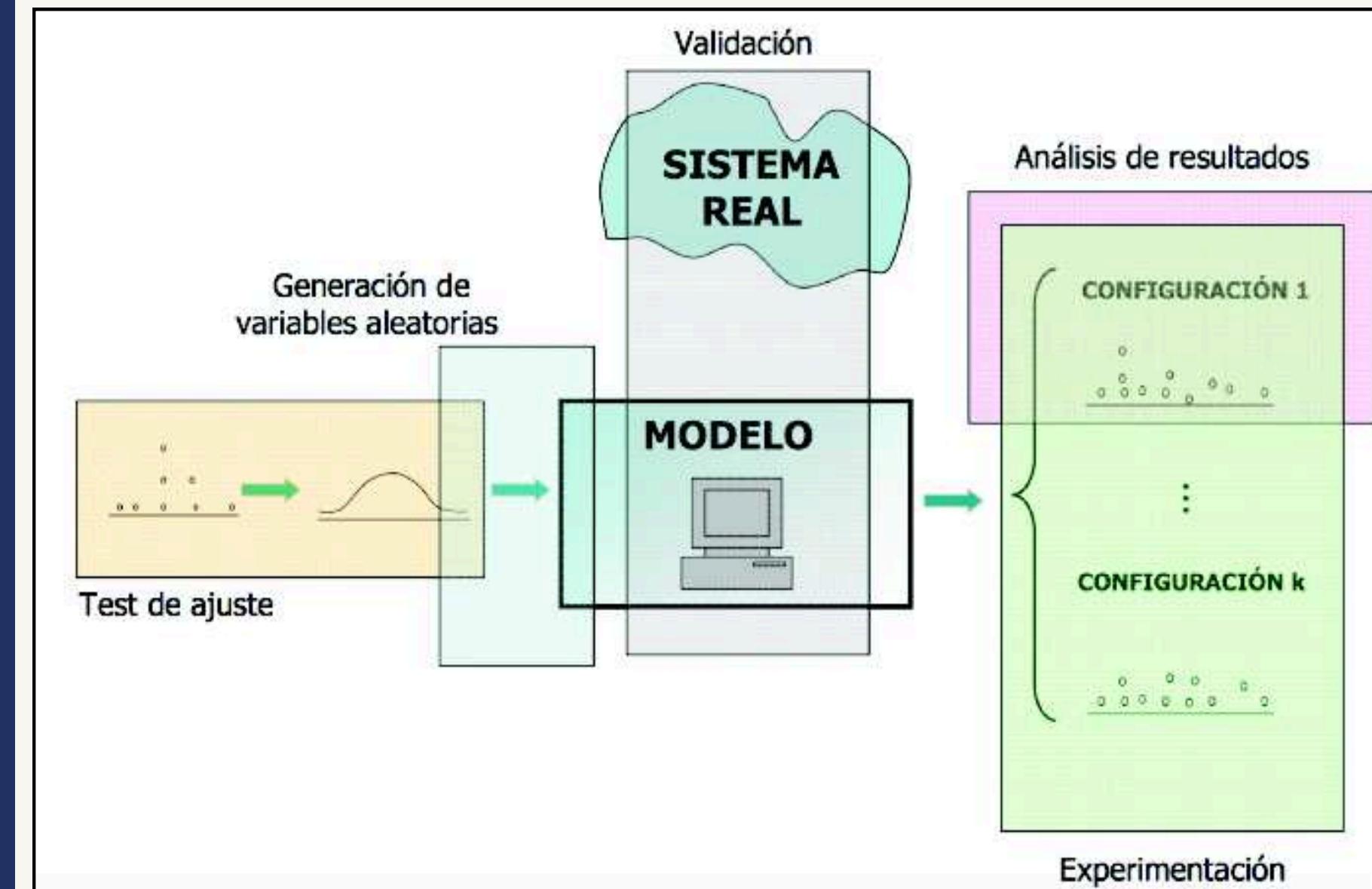


# Distribución de Probabilidad

Describe cómo se reparten las probabilidades de los posibles valores de una variable aleatoria

En un contexto de simulación, cada elemento incierto del sistema, se modela mediante una variable aleatoria con una cierta distribución.

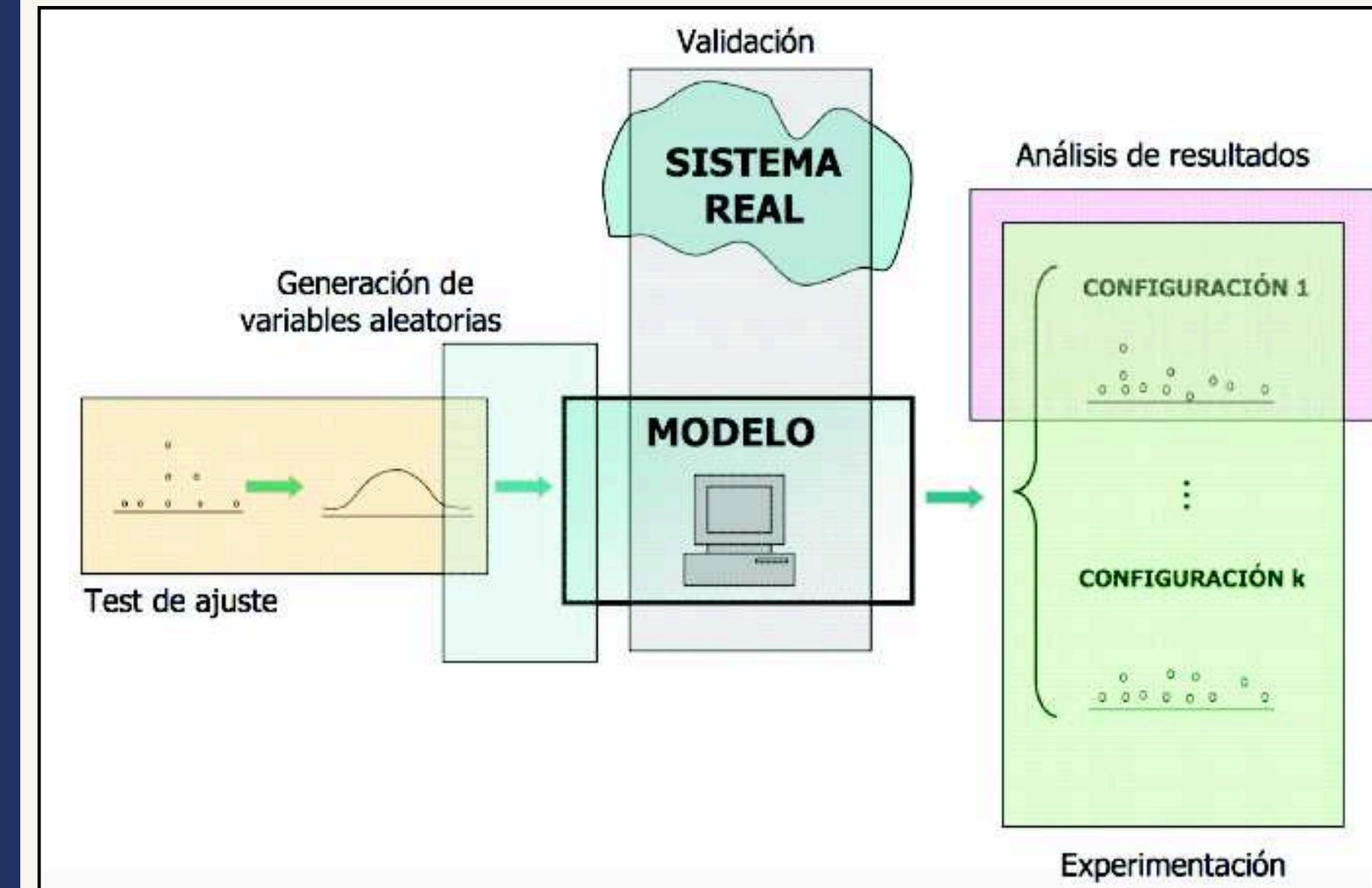
El uso de distribuciones en el modelo permite incorporar el comportamiento estocástico del sistema, imitando su variabilidad natural.



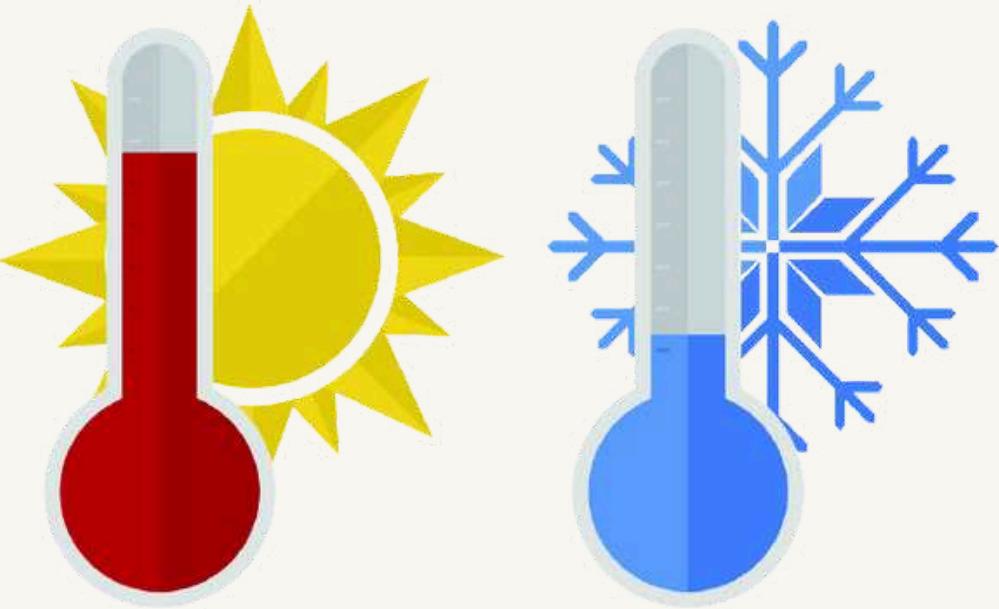
...es un modelo matemático que describe la incertidumbre sobre una variable aleatoria. Se usa para representar y predecir el comportamiento de datos en algoritmos de aprendizaje automático, modelos probabilísticos y redes neuronales.

# Distribución de Probabilidad

1. Se observa el sistema real y se recolectan datos.
2. Se ajustan distribuciones para modelar las variables aleatorias.
3. Se crea un modelo computacional que integra esas distribuciones y la lógica del sistema.
4. Se realizan experimentos cambiando configuraciones.
5. Se analizan los resultados para tomar decisiones o proponer mejoras en el sistema real.



## Distribuciones

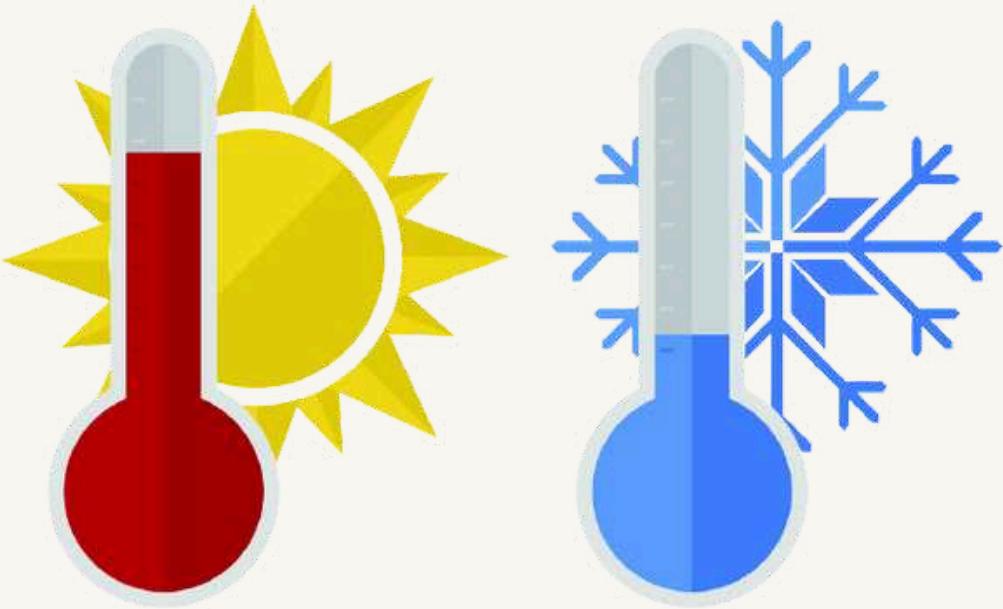


Estado	Eventos
Llueve	9
No Llueve	12

Eventos de lluvia

# Probabilidad conjunta y marginal

## Distribuciones

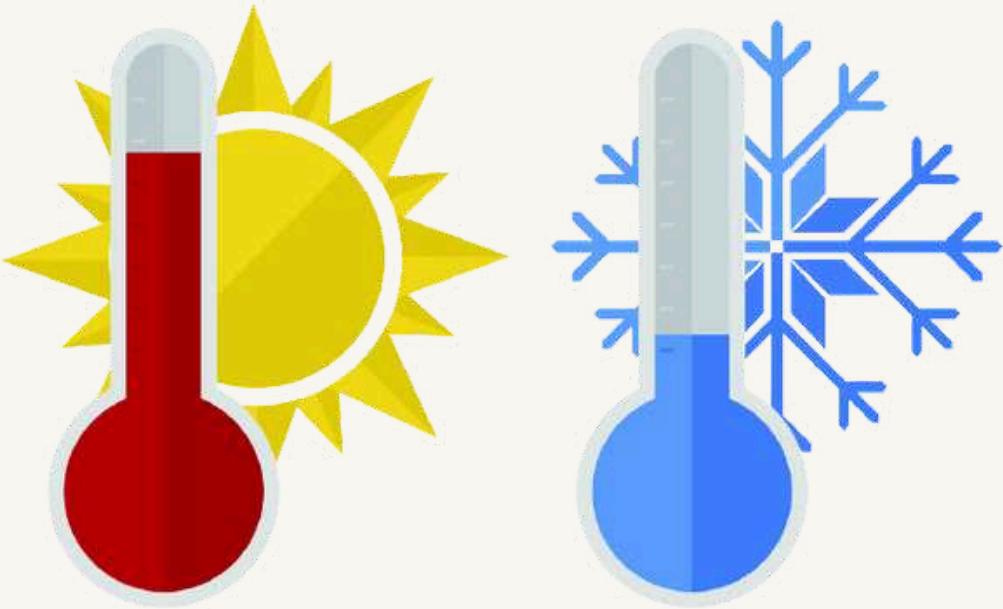


Estado	Eventos
Soleado	10
Nublado	11

Eventos del tiempo

# Probabilidad conjunta y marginal

## Tabla de Eventos



**Soleado      Nublado**

**Llueve**

**No llueve**

	2	7	9
	8	4	12
10		11	

# Probabilidad conjunta y marginal

**Probabilidad Conjunta:** Especifica la probabilidad de que dos eventos sucedan (NO necesariamente dependientes).

**Probabilidad Marginal:** Es la probabilidad de que una sola variable tome un valor específico, ignorando las demás variables.

## Probabilidad Marginal

Var 1	Var 2	Probabilidad
Soleado	Llueve	2/21
Soleado	No Llueve	8/21
Nublado	Llueve	7/21
Nublado	No Llueve	4/21

Marginalizamos variable 2 ignorando la variable 1



Estado 2	Probabilidad
Llueve	9/21
No Llueve	12/21

**Distribucion Marginal:** Es la distribución de una de las variables (o un subconjunto) cuando no nos interesa el resto.

Se obtiene “integrando” o “sumando” la distribución conjunta sobre las variables que se ignoran.

Ejemplo: calcular la **probabilidad que llueva o no llueva**

**NOTA:** Una distribución es el conjunto completo de probabilidades para todos los valores/eventos, mientras que la probabilidad es la medida para un evento en particular.

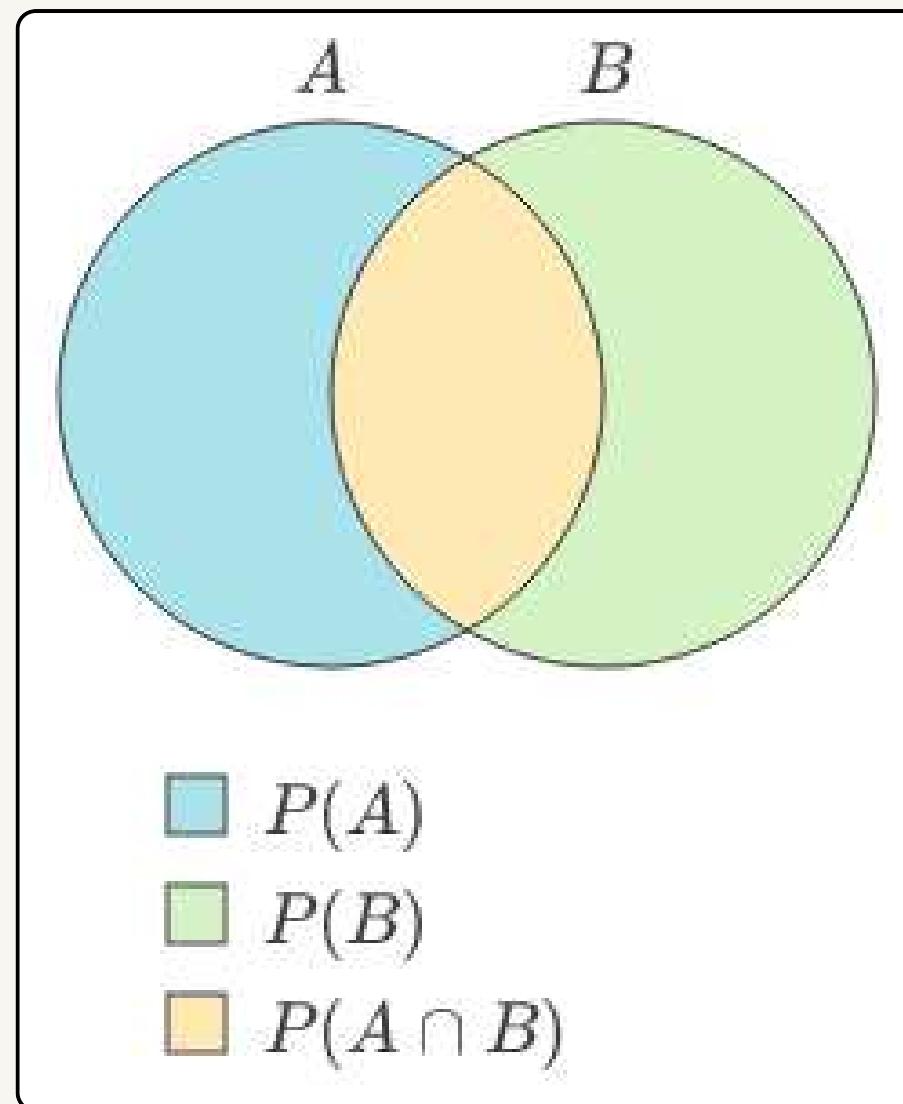
# Probabilidad Condicional

Es la probabilidad de que ocurra un evento dado que otro ya ha ocurrido.

Se denota como:

**P(A | B) <- Cual es la probabilidad que ocurrira A dado que ocurrio B**

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$



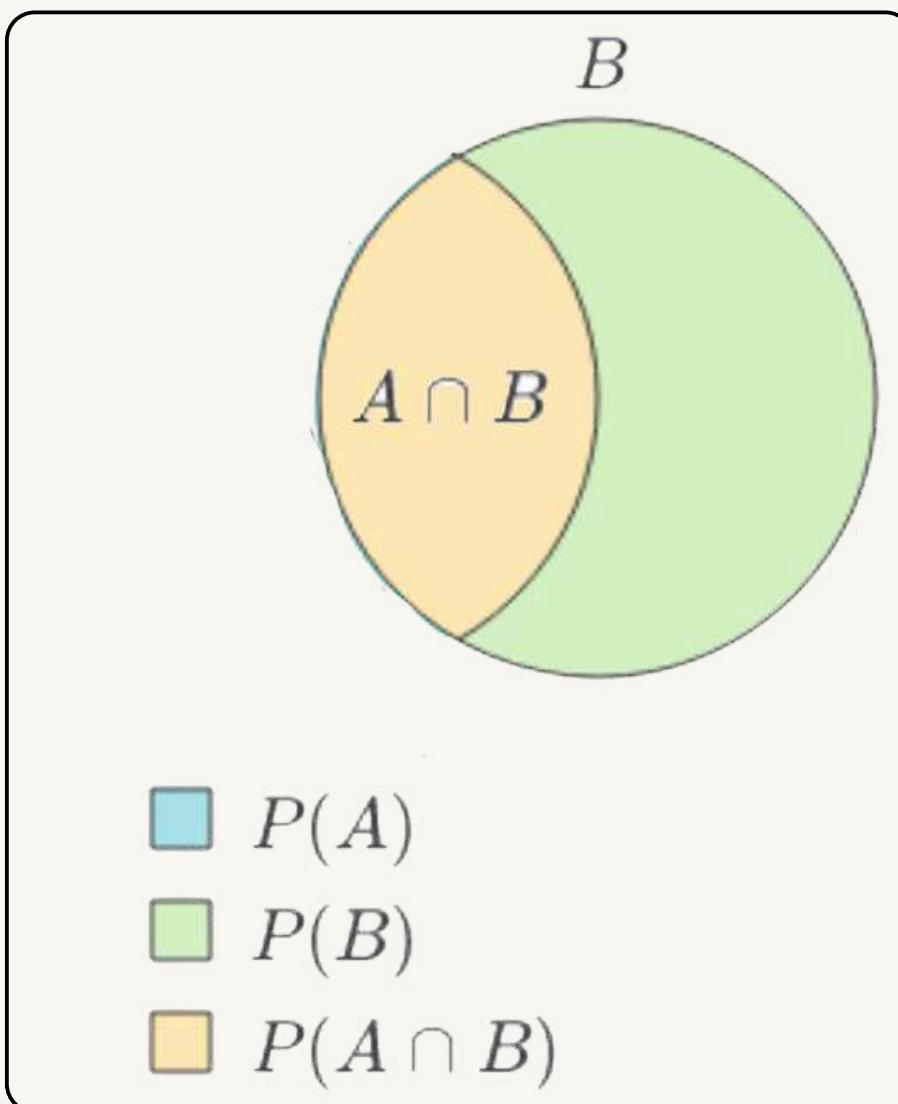
# Probabilidad Condicional

Es la probabilidad de que ocurra un evento dado que otro ya ha ocurrido.

Se denota como:

**P(A | B) <- Cual es la probabilidad que ocurrira A dado que ocurrio B**

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$



# Probabilidad Condicional

Dada la tabla de probabilidad conjunta

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Calcule:

1.  $P(+x|+y)$
2.  $P(-x|+y)$
3.  $P(-y|+x)$

Para:  $P(+x|+y)$

$$P(+x|+y) = \frac{P(+x,+y)}{P(+y)}$$

**NO TENEMOS  $P(+y)$  !!!!!**

# Probabilidad Condicional

Dada la tabla de probabilidad conjunta

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Calcule:

1.  $P(+x|+y)$
2.  $P(-x|+y)$
3.  $P(-y|+x)$

Para:  $P(+x|+y)$

$$P(+x|+y) = \frac{P(+x,+y)}{P(+y)}$$

**NO TENEMOS  $P(+y)$  !!!!!  
entonces:**

**Probabilidad marginal  $P(+y)$**

$$P(+y) = P(+x,+y) + P(-x,+y) = 0.2 + 0.4 = 0.6$$

$$P(+x|+y) = \frac{P(+x,+y) = 0.2}{P(+y) = 0.6}$$

$$P(+x|+y) \approx 0.333$$

# Probabilidad Condicional

Dada la tabla de probabilidad conjunta

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Calcule:

1.  $P(+x|+y)$
2.  $P(-x|+y)$
3.  $P(-y|+x)$

Para:  $P(-x|+y)$

$$P(-x|+y) = \frac{P(-x,+y)}{P(+y)}$$

$$P(-x|+y) = \frac{P(-x,+y) = 0.4}{P(+y) = 0.6}$$

$$P(-x|+y) \approx 0.667$$

# Probabilidad Condicional

Dada la tabla de probabilidad conjunta

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Calcule:

1.  $P(+x|+y)$
2.  $P(-x|+y)$
3.  $P(-y|+x)$

Para:  $P(-y|+x)$

$$P(-y|+x) = \frac{P(-y,+x)}{P(+x)}$$

**NO TENEMOS  $P(+x)$  !!!!!  
entonces:**

**Probabilidad marginal  $P(+x)$**

$$P(+x) = P(+x,-y)+P(+x,+y)=0.2+0.3=0.5$$

$$P(-y|+x) = \frac{P(-y,+x) = 0.3}{P(+x) = 0.5}$$

$$P(-y|+x) = 0.6$$

y si nos piden calcular  $P(X|Y = -y)$ ?

Supongamos la siguiente tabla de probabilidades

Var 1	Var 2	Probabilidad
Soleado	Llueve	0.3
Soleado	No Llueve	0.5
Nublado	Llueve	0.4
Nublado	No Llueve	0.6

**Suma de probabilidades = 1.8**

Normalizar dividiendo entre la suma total

# Normalización

Es el proceso de asegurar que la suma de todas las probabilidades en un espacio dado sea igual a 1.

Supongamos la siguiente tabla de probabilidades

Var 1	Var 2	Probabilidad	Normalizacion	Nueva Probabilidad
Soleado	LLueve	0.3	0.3/1.8	0.1667
Soleado	No Llueve	0.5	0.5/1.8	0.2778
Nublado	Llueve	0.4	0.4/1.8	0.2222
Nublado	No Llueve	0.6	0.6/1.8	0.3333

la suma de las nuevas probabilidades es 1, por lo que la distribución está normalizada

# Normalización

Esto se logra ajustando las probabilidades para que puedan interpretarse como proporciones relativas.

Se parte de una función que no está garantizado que sume 1, y se fuerza a que lo haga dividiéndola entre la suma total de sus valores, obteniendo así una distribución válida.

## INFERENCIA



# Inferencia Probabilística

Es un enfoque utilizado en inteligencia artificial, estadística y ciencias de la computación para modelar y razonar bajo incertidumbre.

Consiste en utilizar probabilidades para representar creencias sobre eventos desconocidos y actualizar esas creencias a medida que se obtiene nueva información (evidencia).

Su objetivo es calcular la probabilidad de ciertas hipótesis o eventos, dados los datos observados y un modelo probabilístico.

## INFERENCIA



# Inferencia Probabilística

### **Incertidumbre:**

- Modela situaciones donde no hay certeza absoluta, como: diagnóstico médico, predicción del clima, detección de spam.
- Ejemplo: ¿Cuál es la probabilidad de que un paciente tenga una enfermedad, dado un resultado positivo en una prueba?

## INFERENCIA



# Inferencia Probabilística

### Modelos Probabilísticos:

- Representan relaciones entre variables mediante distribuciones de probabilidad.
- Ejemplos: Redes Bayesianas, Cadenas de Markov, Modelos de Mezclas Gaussianas.

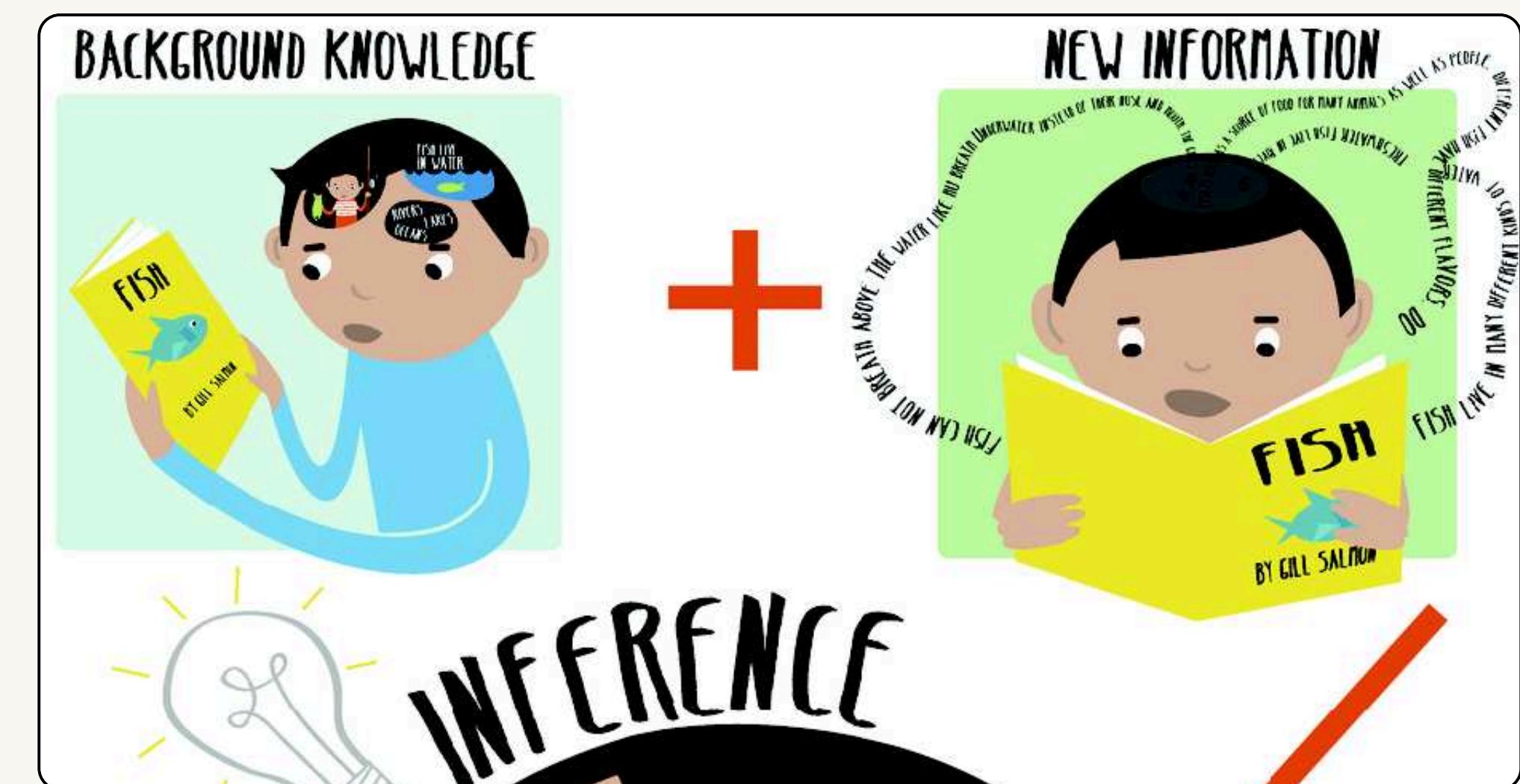
# Inferencia Probabilistica



## INFERENCIA

### Actualización de Creencias:

- Usa reglas como el Teorema de Bayes para ajustar probabilidades ante nueva evidencia.



## INFERENCIA



# Inferencia por Enumeracion

Es un método exacto utilizado en probabilidad y redes bayesianas para calcular probabilidades marginales o condicionales enumerando y sumando todas las posibles combinaciones de variables ocultas (no observadas).

Consiste en:

- Identificar variables:
  - Separar variables en: consulta ( $X$ ), evidencia ( $E$ ), y ocultas ( $Y$ ).
- 1. Calcular la distribución conjunta
- 2. Marginalizar variables ocultas
- 3. Normalizar

### Ejemplo:

Supongamos que estamos modelando si una persona tiene gripe  $X$ , dado que tiene fiebre  $E$ , pero no sabemos si estuvo expuesta a un virus  $H$

- $X = \text{¿El paciente tiene gripe?}$
- $E ?$
- $Y ?$

## INFERENCIA



# Inferencia por Enumeracion

Supongamos los siguientes datos probabilisticos:

Estacion	Temp.	Clima	Prob.
Verano	calido	Soleado	0.30
Verano	calido	Lluvioso	0.05
Verano	frio	Soleado	0.10
Verano	frio	Lluvioso	0.05
Invierno	calido	Soleado	0.10
Invierno	calido	Lluvioso	0.05
Invierno	frio	Soleado	0.15
Invierno	frio	Lluvioso	0.20

Cual es la probabilidad de:

- $P(\text{Clima})$
- $P(\text{Clima} \mid \text{Invierno})$
- $P(\text{Clima} \mid \text{Invierno}, \text{Calido})$

Estacion	Temp.	Clima	Prob.
Verano	calido	Soleado	0.30
Verano	calido	Lluvioso	0.05
Verano	frio	Soleado	0.10
Verano	frio	Lluvioso	0.05
Invierno	calido	Soleado	0.10
Invierno	calido	Lluvioso	0.05
Invierno	frio	Soleado	0.15
Invierno	frio	Lluvioso	0.20

Para  $P(C)$ :

Identificacion de variables:

- Variables de consulta: Clima - C
- Variables de evidencia: Ninguna
- Variables de ocultas: Temperatura - T, Estacion E

$$P(C) = P(C=\text{Soleado}), P(C=\text{Lluvioso})$$

Marginalizar  $P(C)$

$$P(C=\text{Soleado}) = 0.30 + 0.10 + 0.10 + 0.15 = 0.65$$

$$P(C=\text{Lluvioso}) = 0.05 + 0.05 + 0.05 + 0.20 = 0.35$$

Entonces  $P(C)$  es:

Clima	Prob.
Soleado	0.65
Lluvioso	0.35

Estacion	Temp.	Clima	Prob.
Verano	calido	Soleado	0.30
Verano	calido	Lluvioso	0.05
Verano	frio	Soleado	0.10
Verano	frio	Lluvioso	0.05
Invierno	calido	Soleado	0.10
Invierno	calido	Lluvioso	0.05
Invierno	frio	Soleado	0.15
Invierno	frio	Lluvioso	0.20

Para  $P(\text{Clima} | \text{Invierno})$ :

Identificacion de variables:

- Variables de consulta: Clima - C
- Variables de evidencia: Estacion - E = Invierno
- Variables de ocultas: Temperatura - T

$$P(C | E = \text{Invierno}) = P(\text{Soleado} | E = \text{Invierno}), P(\text{Lluvioso} | E = \text{Invierno})$$

$$P(C | E = \text{Invierno}) = \frac{P(C = \text{Soleado}, E = \text{Invierno})}{P(E = \text{Invierno})}, \frac{P(C = \text{Lluvioso}, E = \text{Invierno})}{P(E = \text{Invierno})}$$

Probabilidad total de Invierno (Marginalizacion):

$$P(\text{Invierno}) = 0.10 + 0.05 + 0.15 + 0.20 = 0.50$$

Probabilidades de cada clima en Invierno:

$$P(\text{Soleado, Invierno}) = 0.10 + 0.15 = 0.25$$

$$P(\text{Lluvioso, Invierno}) = 0.05 + 0.20 = 0.25$$

Estacion	Temp.	Clima	Prob.
Verano	calido	Soleado	0.30
Verano	calido	Lluvioso	0.05
Verano	frio	Soleado	0.10
Verano	frio	Lluvioso	0.05
Invierno	calido	Soleado	0.10
Invierno	calido	Lluvioso	0.05
Invierno	frio	Soleado	0.15
Invierno	frio	Lluvioso	0.20

Para  $P(\text{Clima} | \text{Invierno})$ :

$$P(C | E = \text{Invierno}) = \frac{P(C = \text{Soleado}, E = \text{Invierno})}{P(E = \text{Invierno})}, \quad \frac{P(C = \text{Lluvioso}, E = \text{Invierno})}{P(E = \text{Invierno})}$$

Probabilidad total de Invierno (Marginalizacion):

$$P(\text{Invierno}) = 0.50$$

Probabilidades de cada clima en Invierno (Marginalizacion):

$$P(\text{Soleado, Invierno}) = 0.25$$

$$P(\text{Lluvioso, Invierno}) = 0.25$$

$$P(C | E = \text{Invierno}) = \frac{0.25}{0.5}, \frac{0.25}{0.5}$$

Estacion	Clima	Prob.
Invierno	Soleado	0.50
Invierno	Lluvioso	0.50

Estacion	Temp.	Clima	Prob.
Verano	calido	Soleado	0.30
Verano	calido	Lluvioso	0.05
Verano	frio	Soleado	0.10
Verano	frio	Lluvioso	0.05
Invierno	calido	Soleado	0.10
Invierno	calido	Lluvioso	0.05
Invierno	frio	Soleado	0.15
Invierno	frio	Lluvioso	0.20

Para  $P(\text{Clima} | \text{Invierno}, \text{Calido})$ :

Identificacion de variables:

- Variables de consulta: Clima - C
- Variables de evidencia: Estacion - E = Invierno, T = Calido
- Variables de ocultas: Ninguna

$$P(C | \text{Calido}, \text{Invierno}) = P(\text{Soleado} | \text{Calido}, \text{Invierno}), P(\text{Lluvioso} | \text{Calido}, \text{Invierno})$$

$$P(C | \text{Calido}, \text{Invierno}) = \frac{P(C = \text{Soleado}, \text{Cal-Invierno})}{P(\text{Calido-Invierno})}, \frac{P(C = \text{Lluvioso}, \text{Cal-Invierno})}{P(\text{Calido-Invierno})}$$

Probabilidad total de Invierno (Marginalizacion):

$$\mathbf{P(\text{Calido-Invierno}) = 0.10 + 0.05 = 0.15}$$

Probabilidades de cada clima en Invierno:

$$\mathbf{P(\text{Soleado,Calido-Invierno}) = 0.10 = 0.10}$$

$$\mathbf{P(\text{Lluvioso,Calido-Invierno}) = 0.05 = 0.05}$$

Estacion	Temp.	Clima	Prob.
Verano	calido	Soleado	0.30
Verano	calido	Lluvioso	0.05
Verano	frio	Soleado	0.10
Verano	frio	Lluvioso	0.05
Invierno	calido	Soleado	0.10
Invierno	calido	Lluvioso	0.05
Invierno	frio	Soleado	0.15
Invierno	frio	Lluvioso	0.20

Para  $P(\text{Clima} | E=\text{Invierno}, T=\text{Calido})$ :

$$P(C | E = \text{Invierno}) = \frac{P(C = \text{Soleado, Cal-Invierno}), P(C = \text{Lluvioso, Cal-Invierno})}{P(\text{Calido-Invierno})} \quad P(\text{Calido-Invierno})$$

Probabilidad total de Invierno (Marginalizacion):

$$P(\text{Calido-Invierno}) = 0.15$$

Probabilidades de cada clima en Invierno (Marginalizacion):

$$P(\text{Soleado, Calido-Invierno}) = 0.10$$

$$P(\text{Lluvioso, Calido-Invierno}) = 0.05$$

$$P(C | E = \text{Invierno}) = \frac{0.10}{0.15}, \frac{0.05}{0.15}$$

Estacion	Temp.	Clima	Prob.
Invierno	calido	Soleado	0.6667
Invierno	calido	Lluvioso	0.3333

## Regla del Producto

Si sabemos que:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

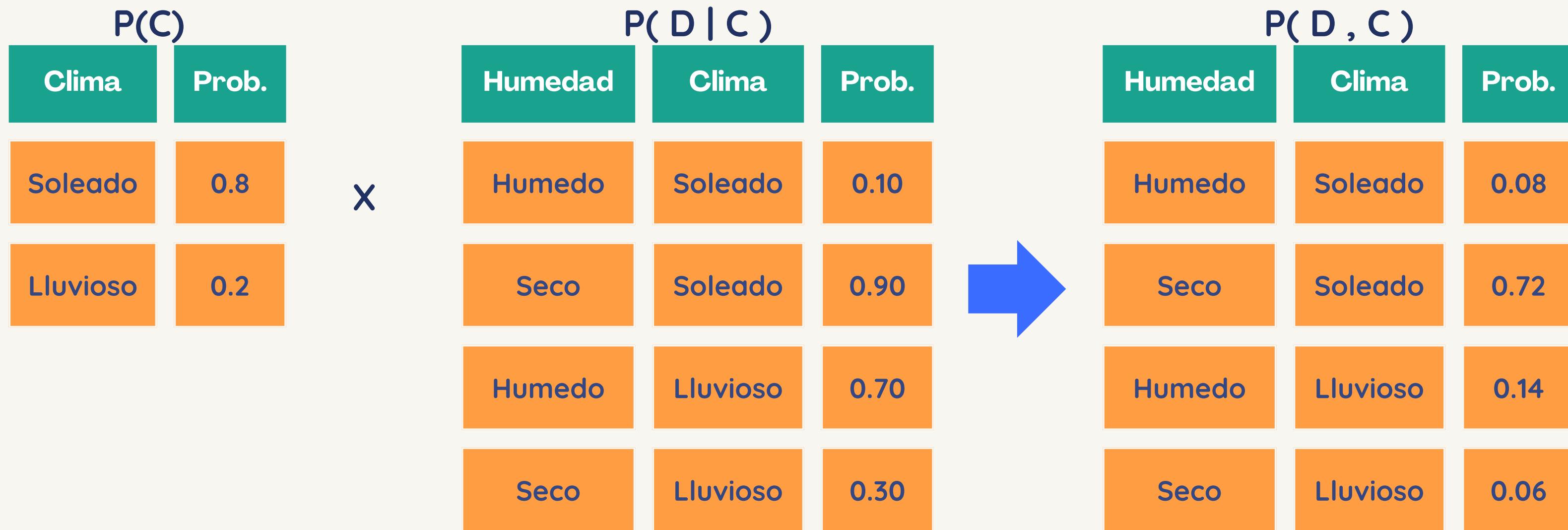


$$P(A,B) = P(A|B) P(B)$$



## Regla del Producto

$$P(A,B) = P(A|B) P(B)$$



## INFERENCIA



# Regla de la Cadena

Es una herramienta en probabilidad y cálculo que permite descomponer probabilidades conjuntas complejas en multiplicaciones de probabilidades condicionales más simples.

$$P(A,B,C) = P(A) \cdot P(B | A) \cdot P(C | A, B)$$

Es decir, la probabilidad de que ocurran **A**, **B** y **C** juntos es igual a la probabilidad de **A**, multiplicada por la probabilidad de **B** dado **A**, multiplicada por la probabilidad de **C** dado **A** y **B**.

## INFERENCIA



# Regla de la Cadena

Ejemplo: Calcular la probabilidad de que usando la regla de la cadena:

1. Hoy esté nublado (N),
2. Luego llueva (L), y
3. Tu paraguas se moje (M).

$$P(N, L, M) = P(N) \cdot P(L | N) \cdot P(M | N, L)$$

- $P(N)$ : Probabilidad de que esté nublado.
- $P(L|N)$ : Probabilidad de lluvia si está nublado.
- $P(M|N, L)$ : Probabilidad de que el paraguas se moje si está nublado y llueve.



# INFERENCIA Y REDES DE BAYES

PARA INTELIGENCIA ARTIFICIAL



# AGENDA

- Teorema de Bayes
- Independencia de variables
- Redes Bayesianas
- D-Separation
- Inferencia por Enumeracion en una red bayesiana
- Eliminacion de variables



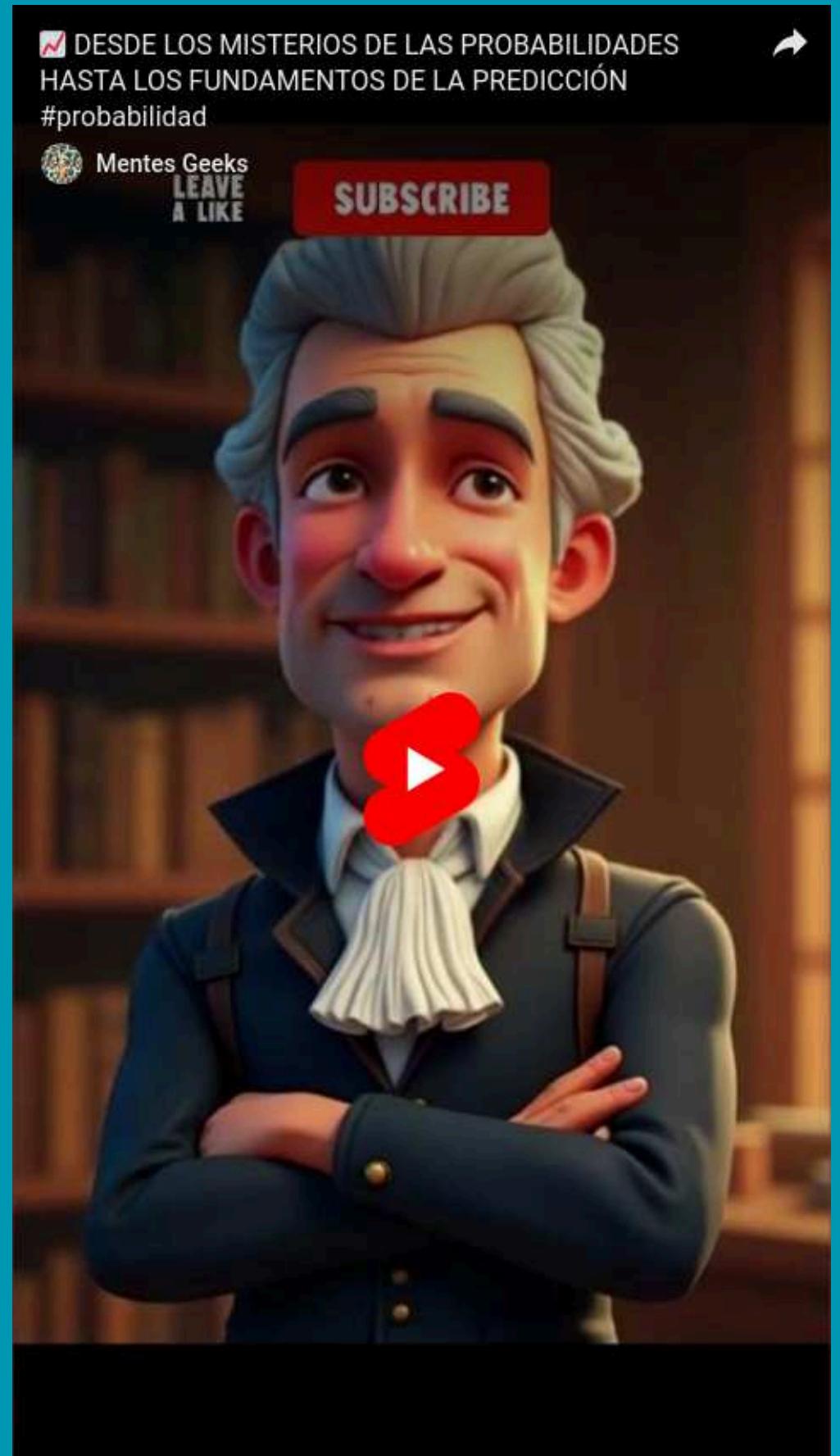
# TEOREMA DE BAYES



# Bayes

Thomas Bayes, matemático británico, estudió el problema de la determinación de la probabilidad de las causas a través de los efectos observados, su teorema se resuelve el problema conocido como **de la probabilidad inversa**

El teorema de Bayes tiene muchas aplicaciones, incluyendo **Aprendizaje Automático**, se usa en modelos de clasificación como el Naïve Bayes.



# Teorema de Bayes

Es una regla matemática que nos dice cómo actualizar nuestras creencias cuando obtenemos nueva información.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(\text{Causa}|\text{Efecto}) = \frac{P(\text{Efecto}|\text{Causa})P(\text{Causa})}{P(\text{Efecto})}$$



# Teorema de Bayes

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

**P(A|B)** = **Probabilidad**  
Si nuestra hipótesis fuera verdadera  
¿qué tan posible es que ocurra la evidencia?

**Estimación posterior**  
¿Qué tan probable es que ocurra nuestra hipótesis  
cuando observamos la evidencia?

**P(B|A) · P(A)**  
**Estimación previa**  
Lo que ya sabíamos.

**P(B)**  
**Marginal**  
¿Qué tan probable es que ocurra la  
nueva evidencia bajo todas las hipótesis posibles?

# Inferencia con Teorema de Bayes

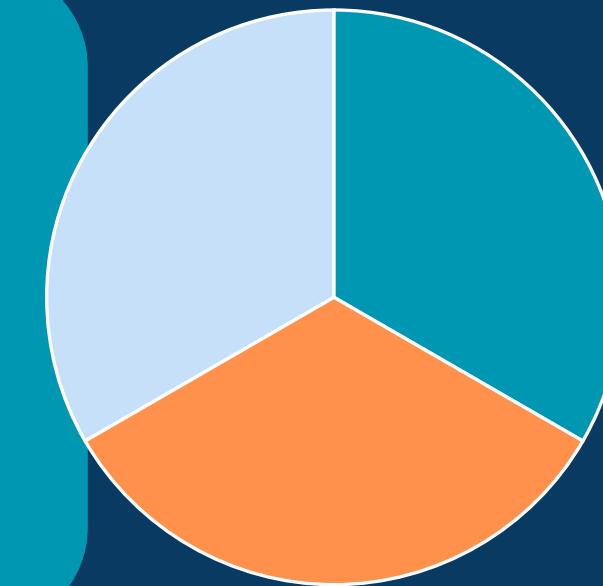
Segun datos del MSPAS durante 2016, 8 de cada 10 pacientes con Chinkunguya presentaban dolor de cuerpo entre sus sintomas.

En una muestra aleatoria de 10,000 Guatemaltecos se determinó que 1 de cada 10 presentaba síntomas de dolor de cuerpo, y existió únicamente un caso confirmado de Chinkunguya.

¿Cual es la probabilidad de tener Chinkunguya si tengo dolor de cuerpo?

## Variables

- Dolor de Cuerpo = D (+d, -d)
- Chinkunguya = C (+c, -c)



### Probabilidad de tener dolor de cuerpo ( $P(+d)$ )

que 1 de cada 10 presentaba síntomas de dolor de cuerpo

$$P(+d) = \frac{1}{10} = 0.10$$

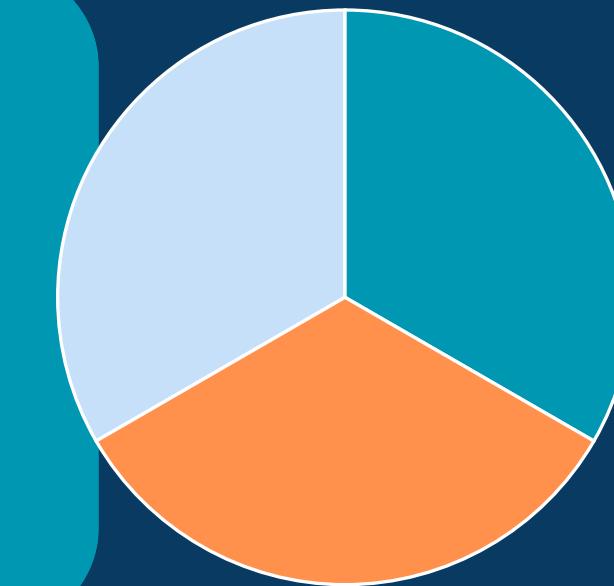
### Probabilidad de tener Chinkunguya ( $P(+c)$ )

En una muestra aleatoria de 10,000 Guatemaltecos y existió únicamente un caso confirmado de Chinkunguya

$$P(+c) = \frac{1}{10,000} = 0.0001$$

## Variables

- Dolor de Cuerpo = D (+d, -d)
- Chinkunguya = C (+c, -c)



Probabilidad de tener dolor de cuerpo dado que se tiene Chinkunguya ( $P(+d | +c)$ )

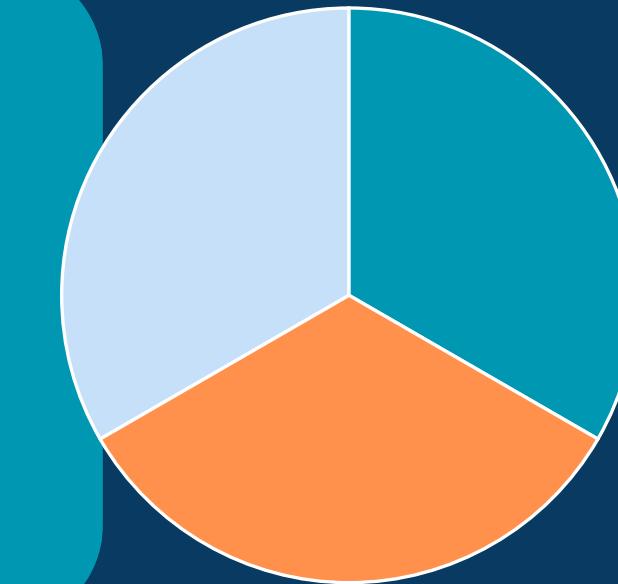
8 de cada 10 pacientes con Chinkunguya presentaban dolor de cuerpo entre sus síntomas

$$P(+d | +c) = \frac{8}{10} = 0.8$$

Entonces si tengo dolor de cuerpo, ¿Cuál es la probabilidad de tener Chinkunguya ( $P( +c | +d)$ ) ?

$$P(+c | +d) = \frac{P(+d | +c)P(+c)}{P(+d)} = \frac{0.8 * 0.0001}{0.1} = 0.0008$$

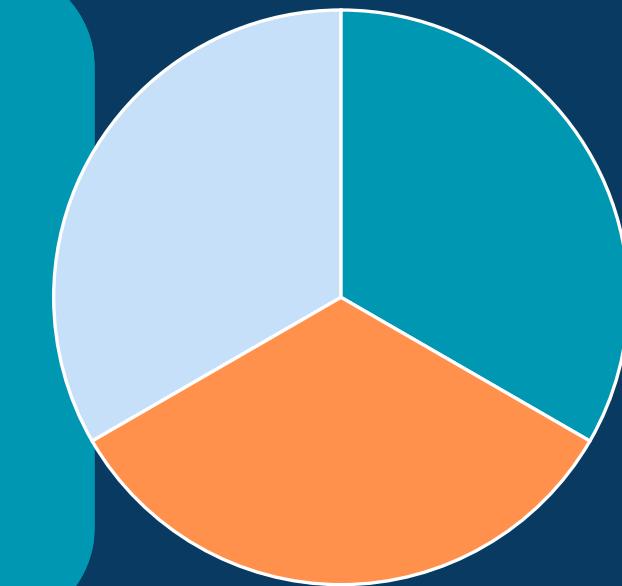
# ¿Que puede hacer el agente con los modelos?



Modelos probabilisticos describen como es el mundo, por tanto el agente puede:

- Razonar acerca de variables dada cierta evidencia (**inferir**)
- Explicación (**razonamiento diagnostico**)
- Predicción (**razonamiento causal**)

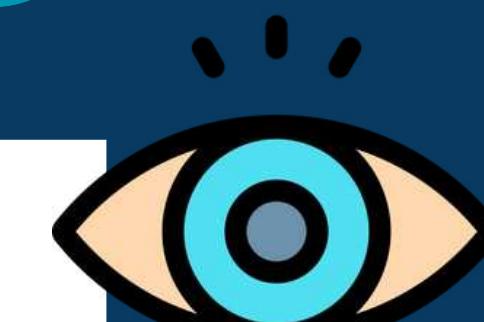
# Ejercicio de inferencia con teorema de Bayes



Clima	Prob.
Soleado	0.8
Lluvioso	0.2

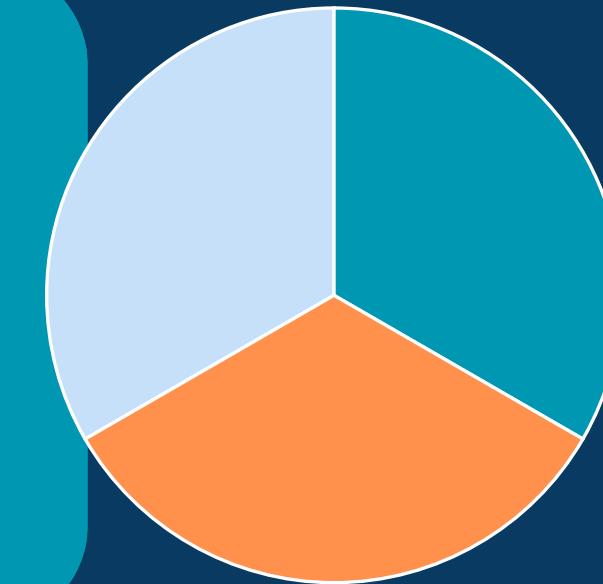
D	C	Prob.
Mojado	Soleado	0.1
Seco	Soleado	0.9
Mojado	Lluvioso	0.7
Seco	Lluvioso	0.3

¿ $P(C|D=\text{seco})?$



Distribucion conjunta  
o distribucion condicional?

# Si el suelo está seco, ¿qué tan probable es que el clima haya sido soleado o lluvioso?



Tenemos:

2 tipos de clima: Soleado y Lluvioso.

2 estados del suelo: Seco y Mojado.

## Probabilidad total de suelo seco

Combinamos ambas opciones (soleado y lluvioso):  $P(D = \text{seco}) = (0.9 \cdot 0.8) + (0.3 \cdot 0.2) = 0.72 + 0.06 = 0.78$

## Aplicamos el Teorema de Bayes

Clima soleado dado que el suelo está seco

$$P(C = \text{soleado}|D = \text{seco}) = \frac{0.9 \cdot 0.8}{0.78} = \frac{0.72}{0.78} \approx 0.9231$$

Clima lluvioso dado que el suelo está seco

$$P(C = \text{lluvioso}|D = \text{seco}) = \frac{0.3 \cdot 0.2}{0.78} = \frac{0.06}{0.78} \approx 0.0769$$

## Conclusión

Si el suelo está seco, hay:

- 92.3% de probabilidad de que el clima sea soleado.
- 7.7% de probabilidad de que el clima sea lluvioso.

# Teorema de probabilidad total

Es una herramienta que nos permite calcular la probabilidad de un evento considerando todas las maneras en que ese evento puede ocurrir, a través de una partición del espacio muestral.

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)(A_i)}$$

8 Ejemplo Teorema de Probabilidad Total y Teorema de Bayes  
Probabilidad Teorema de probabilidad total. Teorema de Bayes

El total de piezas producidas en una fábrica lo hacen tres máquinas A, B y C, que producen, respectivamente el 40%, 35% y 25% de las piezas. Las piezas defectuosas que producen las máquinas A, B y C son, respectivamente, el 1%, 2% y el 3%.

a) Elegida una pieza al azar, calcular la probabilidad de que sea defectuosa.

b) Sabiendo que la pieza elegida es defectuosa, ¿cuál es la probabilidad de que la haya fabricado la máquina C?

The tree diagram illustrates the sample space partitioning:

- Máquina A: 0.40 (total), 0.01 (defectiva D), 0.99 (no defectiva  $\bar{D}$ )
- Máquina B: 0.35 (total), 0.02 (defectiva D), 0.98 (no defectiva  $\bar{D}$ )
- Máquina C: 0.25 (total), 0.03 (defectiva D), 0.97 (no defectiva  $\bar{D}$ )

Calculus:

$P(D) = 0.01 + 0.35 \cdot 0.02 + 0.25 \cdot 0.03 = 0.0185$

$P(C|D) = \frac{P(C \cap D)}{P(D)} = \frac{0.25 \cdot 0.03}{0.0185} = 0.415$

Watch on YouTube

Matematrix



Probabilidad condicional explicada de manera visual (Teorema de Bayes) | Khan Academ...



Share



Watch on YouTube

# Resuelve...

En un sorteo recibe un premio si sacas una canica roja de un saco de 100 canicas, donde 20 canicas son rojas y el resto azules. De las canicas rojas 15 son chicas y 5 grandes, mientras que de las azules 70 son chicas y 10 grandes. Si puedes sentir el tamaño, ¿Qué tamaño te conviene sacar?

## Datos del problema

Total de canicas: 100

Canicas rojas: 20

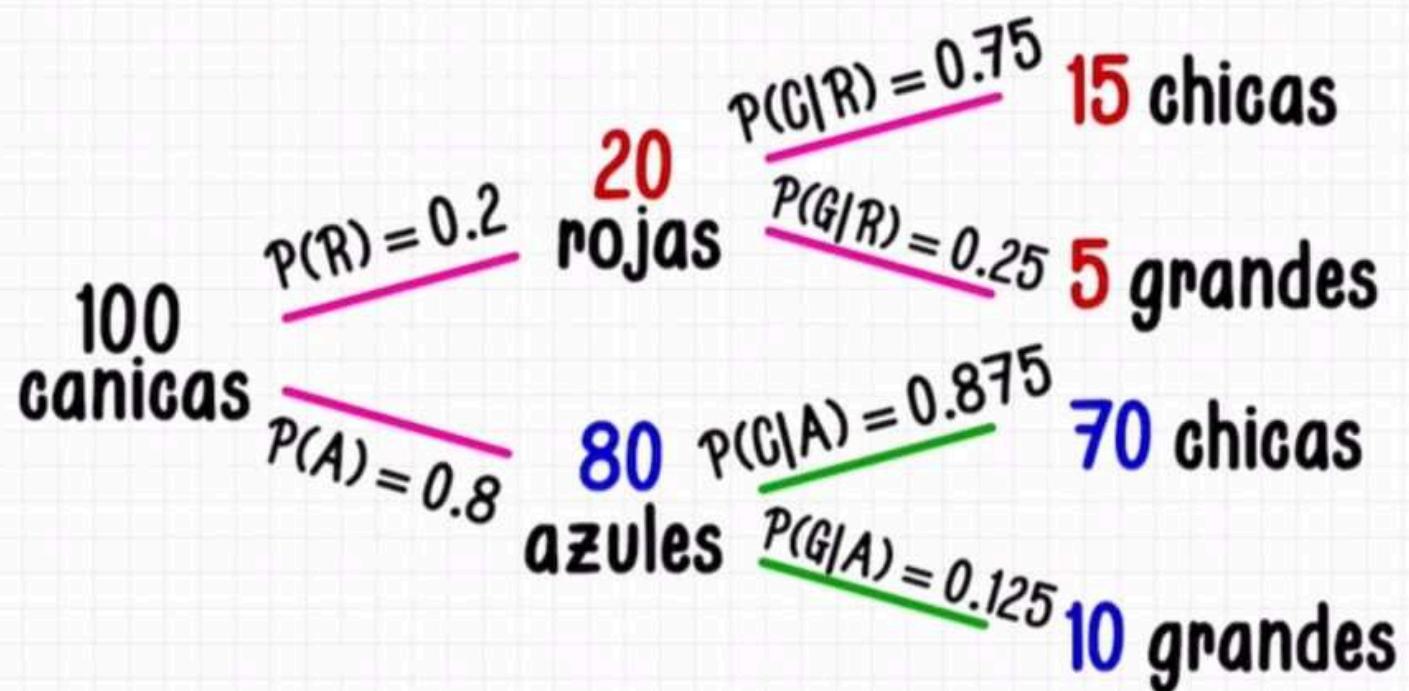
- Chicas: 15
- Grandes: 5

Canicas azules: 80

- Chicas: 70
- Grandes: 10

# Resuelve...

## Teorema de Bayes



$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$\begin{aligned}P(C) &= (0.2)(0.75) + (0.8)(0.875) \\&= 0.15 + 0.7 = 0.85\end{aligned}$$

$$\begin{aligned}P(G) &= (0.2)(0.25) + (0.8)(0.125) \\&= 0.05 + 0.1 = 0.15\end{aligned}$$

$$P(R|C) = \frac{P(C|R) \cdot P(R)}{P(C)} = \frac{(0.75)(0.2)}{0.85} = \frac{0.15}{0.85} = 0.1765 \quad } \quad 17.65\%$$

$$P(R|G) = \frac{P(G|R) \cdot P(R)}{P(G)} = \frac{(0.25)(0.2)}{0.15} = \frac{0.05}{0.15} = 0.3333 \quad } \quad 33.33\%$$



# INDEPENDENCIA DE VARIABLES



# Independencia de Variables

Se refiere a que el comportamiento de una variable no afecta en absoluto al comportamiento de la otra.

Es decir, conocer el valor de una variable no proporciona ninguna información sobre el valor de la otra.

Decimos que son independientes si, para cualquier par de valores  $x$  e  $y$ , se cumple:

$$\forall x, y \ P(x, y) = P(x)P(y) \dashrightarrow X \perp\!\!\!\perp Y$$

La independencia es una suposición fundamental en muchos modelos, ya que simplifica el análisis y el cálculo de probabilidades.

# Independencia de Variables

## Ejemplo: Lanzamiento de dos dados

Variables

- $X$  = Resultado del primer dado (valores posibles: 1, 2, 3, 4, 5, 6)
- $Y$  = Resultado del segundo dado (valores posibles: 1, 2, 3, 4, 5, 6)

**Como es un dado justo:**

$$P(X=x) = 1/6 \text{ para } x=1,2,3,4,5,6$$

$$P(Y=y) = 1/6 \text{ para } y=1,2,3,4,5,6$$

Es decir, cada número tiene la misma probabilidad de salir en cada dado.

# Independencia de Variables

Como son dos dados independientes, la probabilidad de que el primer dado sea un 3 y el segundo dado sea un 5 es:

$$P(X=3 \text{ y } Y=5) = P(X=3) \cdot P(Y=5)$$

$$P(X=3 \text{ y } Y=5) = 1/6 \cdot 1/6 = 1/36$$

Esta es la fórmula de independencia: el producto de las probabilidades individuales (marginales) nos da la probabilidad conjunta.

El resultado del primer dado no afecta al resultado del segundo dado.

Por eso, saber el valor de X (primer dado) no cambia las probabilidades de Y (segundo dado).

**Esto es independencia.**

En este caso, tirar un dado no afecta al otro, por lo que:

$$P(X,Y) = P(X)P(Y) \Rightarrow X \perp Y$$

# Independencia Condicional

Ocurre cuando dos variables son independientes entre sí al condicionar (fijar) una tercera variable

Si conocemos el valor de esa tercera variable, el conocimiento de una de las dos primeras no aporta información adicional sobre la otra.

$$\forall x, y, z \quad P(x, y|z) = P(x|z)P(y|z) \dashrightarrow X \perp\!\!\!\perp Y | Z$$

# Independencia Condicional

## Ejemplo: Clima, paraguas la calle mojada

Variables

- X = La calle está mojada. (sí/no)
- Y = La gente lleva paraguas (sí/no)
- Z = El clima (lluvioso o soleado)

Si vemos a alguien con paraguas, es razonable pensar que probablemente la calle esté mojada (porque quizás llovió), así que:

$$P(X \mid Y) \neq P(X)$$

X y Y no son independientes, ya que ver un paraguas nos da información sobre la calle mojada.

# Independencia Condicional

Si sabemos que el clima es lluvioso, entonces:

- Saber que alguien lleva paraguas ya no añade mucha información extra sobre si hay charcos, porque el clima (Z) ya explica eso.

Entonces:

- $P(X \cap Y, Z=\text{lluvioso}) = P(X | Z=\text{lluvioso})$
- X y Y son condicionalmente independientes dado Z.

**Sin saber el clima, X y Y están relacionados.**

**Sabiendo el clima (Z), X y Y ya no dependen uno del otro. Toda la información sobre charcos y paraguas está explicada por el clima.**

# Diferencias Clave

Independencia	Independencia Condicional
$P(X, Y) = P(X)P(Y)$	$P(X, Y   Z) = P(X   Z) \cdot P(Y   Z)$
No depende de ninguna tercera variable	Involucra las variables X, Y y una tercera Z.
Ejemplo: Lanzar dos dados. El resultado de un dado no afecta al otro.	Ejemplo: El uso de paraguas (Y) y que el suelo esté mojado (X) son independientes si sabes que está lloviendo (Z).

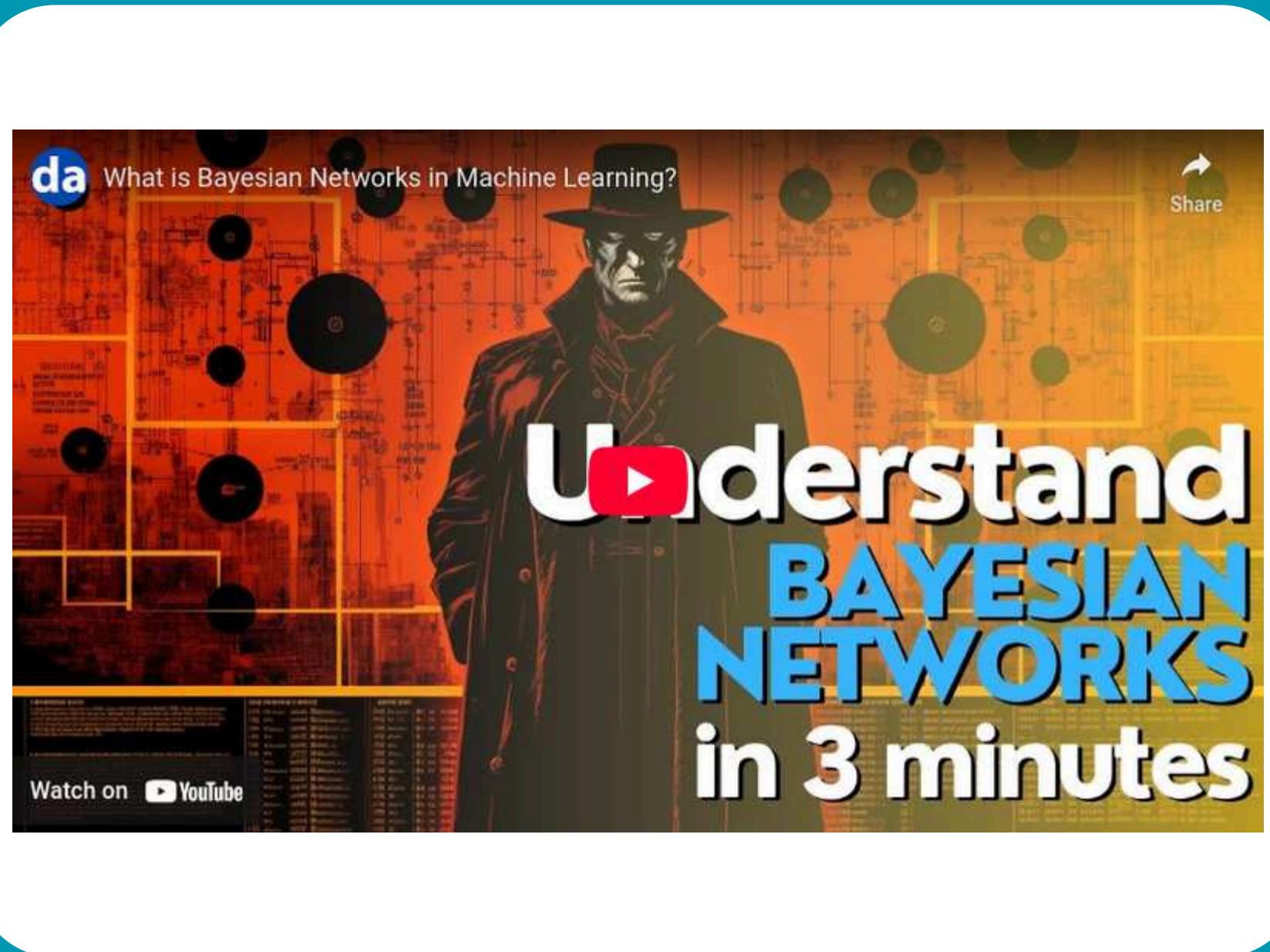
- **Independencia:** X y Y no se afectan.
- **Independencia condicional:** X y Y parecen relacionados, pero esa relación desaparece cuando conoces Z.



# REDES BAYESIANAS



# Redes Bayesianas



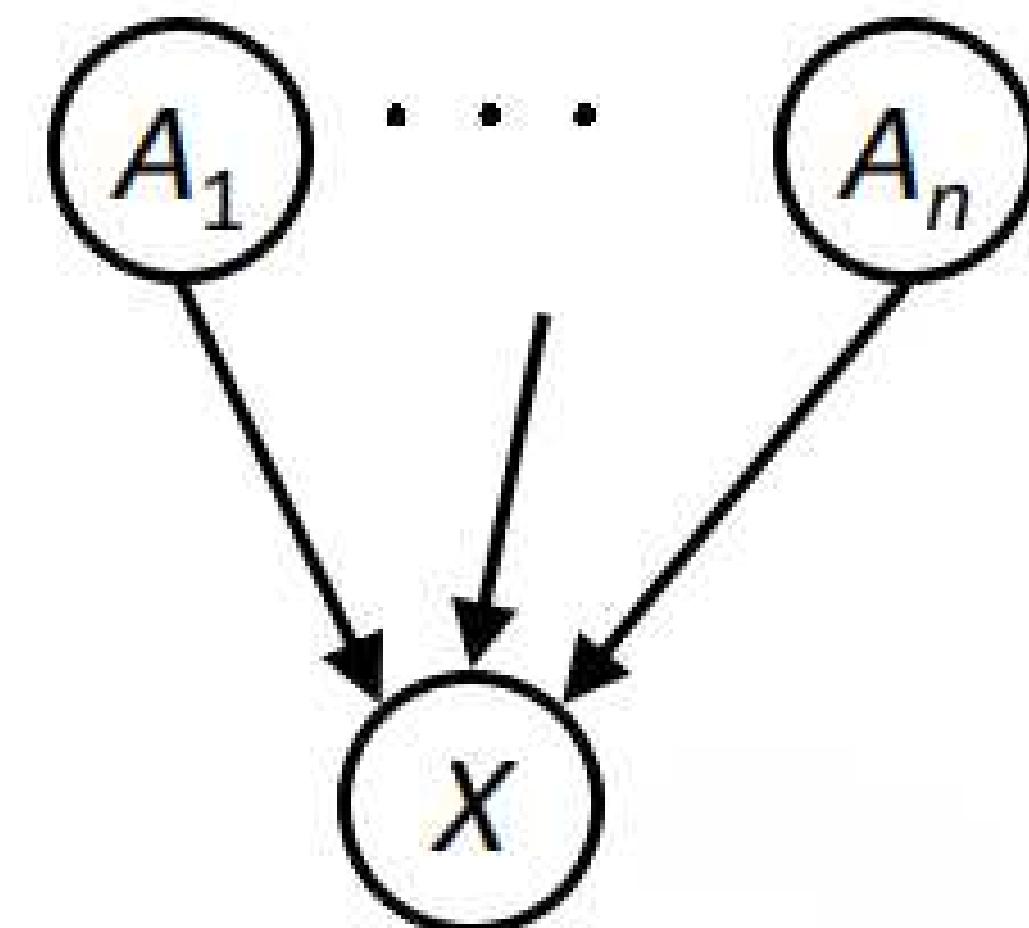
Se conoce también como:

- Red probabilística (Probabilistic Network)
- Red Causal (Causal Network)
- Red de Creencias (Belief Network)
- Mapa de Conocimiento (Knowledge Map)

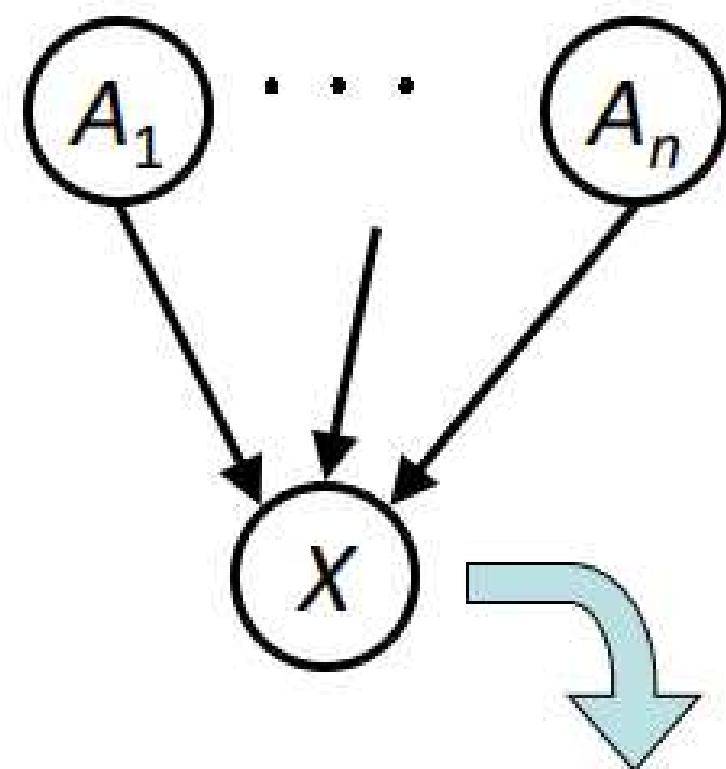
# Redes Bayesianas

Es un modelo probabilístico gráfico que representa un conjunto de variables y las relaciones de dependencia condicional entre ellas, utilizando un grafo dirigido acíclico:

- Cada nodo representa una variable aleatoria
- Las flechas o conexiones entre nodos indican la dirección de la influencia o dependencia, muchas veces interpretada como una relación causal



# Redes Bayesianas



$$P(X|A_1, \dots, A_n)$$

La red bayesiana utiliza el teorema de Bayes para actualizar las probabilidades de los eventos a medida que se incorpora nueva evidencia

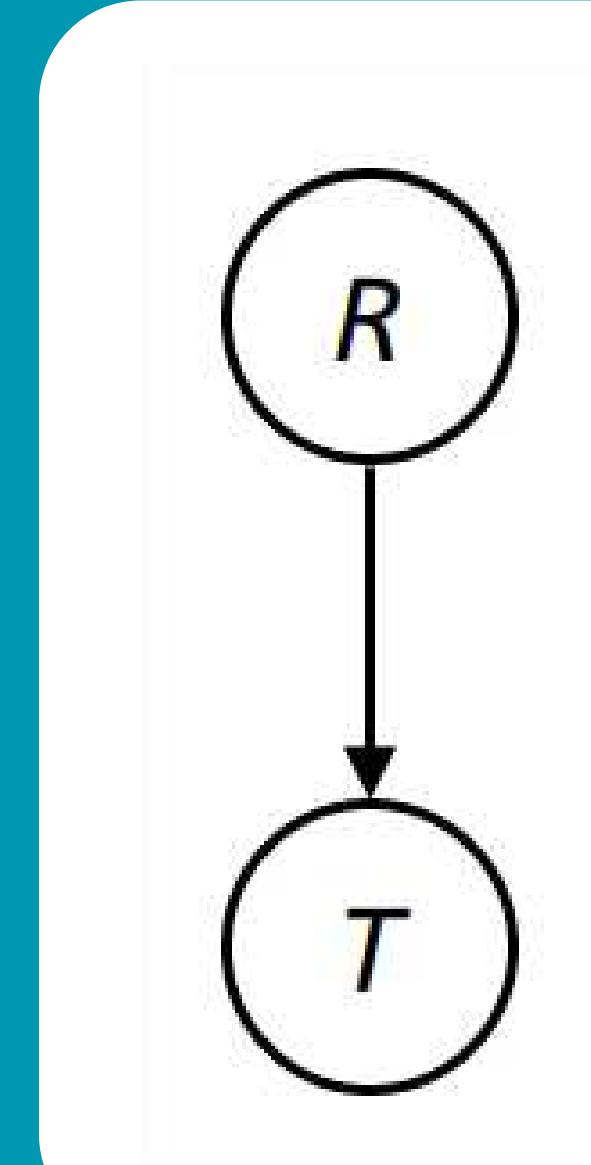
Describe distribuciones conjuntas complejas (modelos), usando distribuciones locales (probabilidades condicionales).

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i|\text{parents}(X_i))$$

# Ejemplo Redes Bayesianas - Tráfico y Lluvia

En una ciudad, se ha observado que la probabilidad de que llueva en un día cualquiera es de  $1/4$ . Además, si llueve, la probabilidad de que haya tráfico es de  $3/4$ , mientras que si no llueve, la probabilidad de tráfico es de  $1/2$ .

Utilizando una red bayesiana, calcula la probabilidad de que llueva y no haya tráfico en un día determinado.



$$P(R)$$

+r	$1/4$
-r	$3/4$

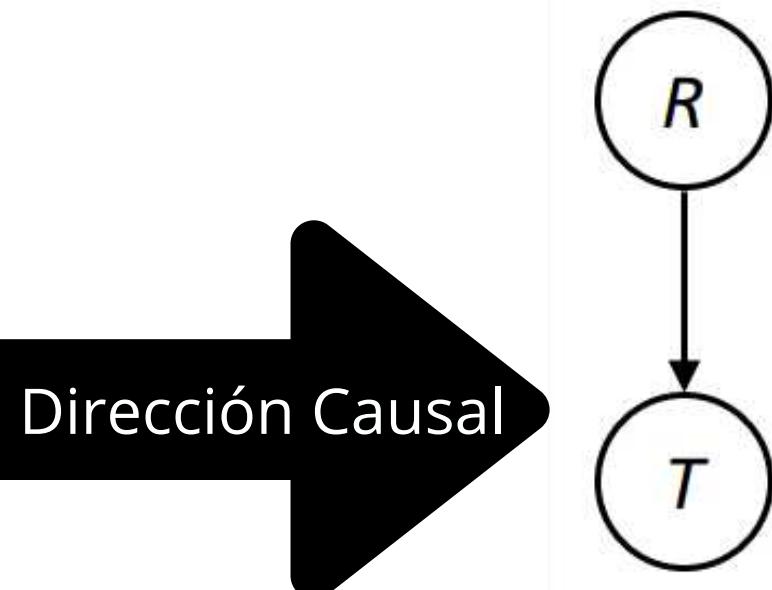
$$P(T|R)$$

+r	+t	$3/4$
	-t	$1/4$
-r	+t	$1/2$
	-t	$1/2$

# Ejemplo Redes Bayesianas - Tráfico y Lluvia

Calcula la probabilidad de que llueva y no haya tráfico en un día determinado.

Utiliza la fórmula de la probabilidad conjunta para calcular  $P(+r, -t)$ .



$P(R)$	
	+r
+r	1/4
-r	3/4

$P(T R)$	
	+t
+r	3/4
-r	1/4
+r	1/2
-r	1/2

$$P(x_1, x_2, \dots x_n) = \prod_{i=1}^n P(x_i | \text{parents}(X_i))$$

$$P(+r, -t) = P(+r)P(-t | +r)$$

$$= \frac{1}{4} * \frac{1}{4} = \frac{1}{16}$$

# Ejemplo Redes Bayesianas – Robo

Estás en el trabajo y recibes una llamada de tu vecino John, que te dice que la alarma de tu casa está sonando. Sin embargo, tu otra vecina Mary no te llama. A veces, la alarma se activa por pequeños terremotos.

**La pregunta es:** *¿Hay un ladrón?*

## Variables

La red bayesiana modela estas variables:

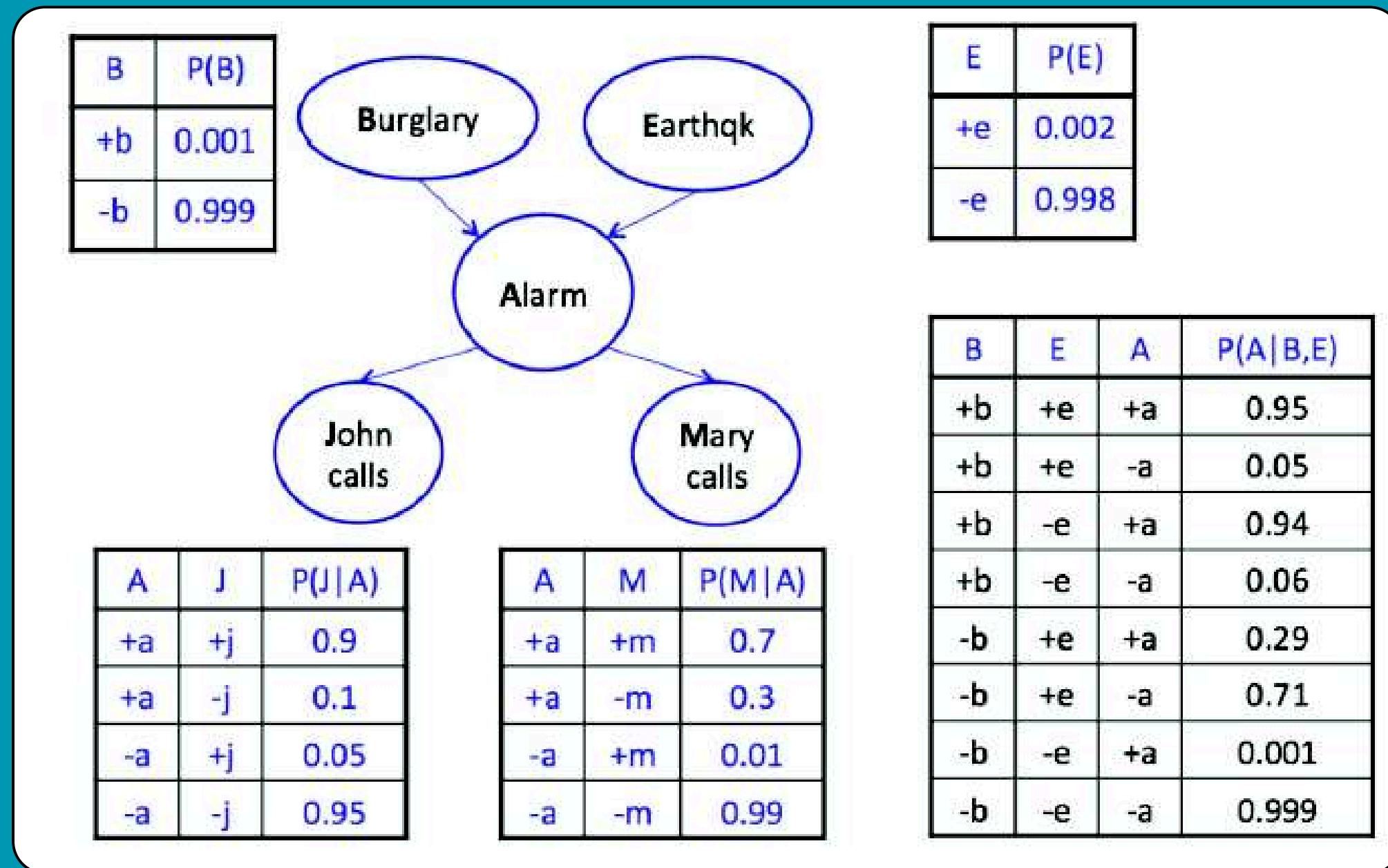
- Burglar: ¿Hay un ladrón?
- Earthquake: ¿Hubo un terremoto?
- Alarm: ¿Está sonando la alarma?
- JohnCalls: ¿John te llama?
- MaryCalls: ¿Mary te llama?

## Relación causal

La red refleja relaciones causales, es decir, cómo unas variables causan otras:

- Si hay un ladrón, puede activar la alarma.
- Si hay un terremoto, también puede activar la alarma.
- Si la alarma suena, eso puede hacer que:
  - John te llame.
  - Mary te llame.

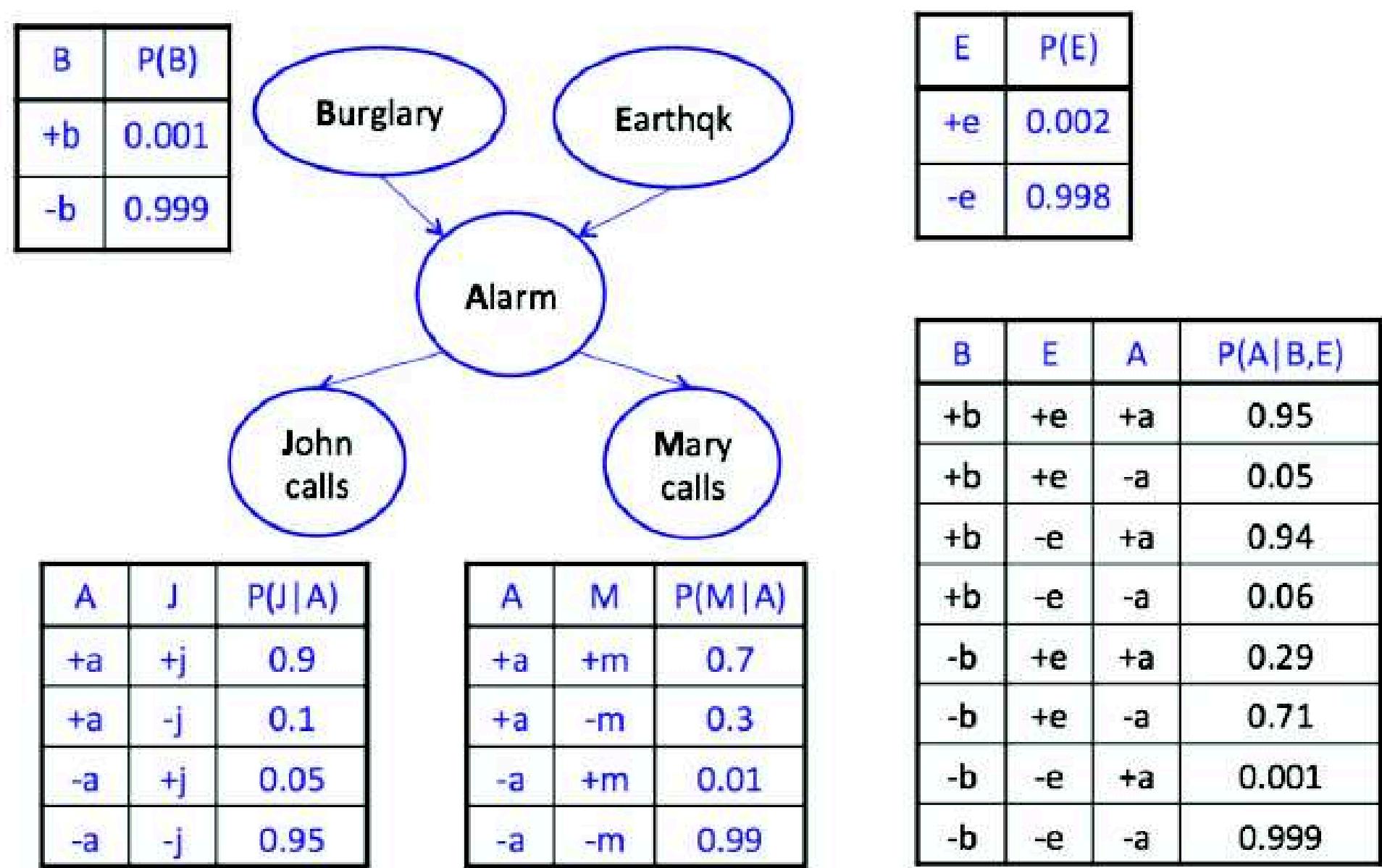
# Ejemplo Redes Bayesianas – Robo



¿Qué te permite esta red?

- Calcular probabilidades conjuntas de todas las variables.
- Estimar, dado que John te llama pero Mary no, cuál es la probabilidad de que haya un ladrón.
- Actualizar tus creencias conforme obtienes evidencia (por ejemplo, si después te enteras que hubo un terremoto, ajustas la probabilidad de que haya un ladrón).

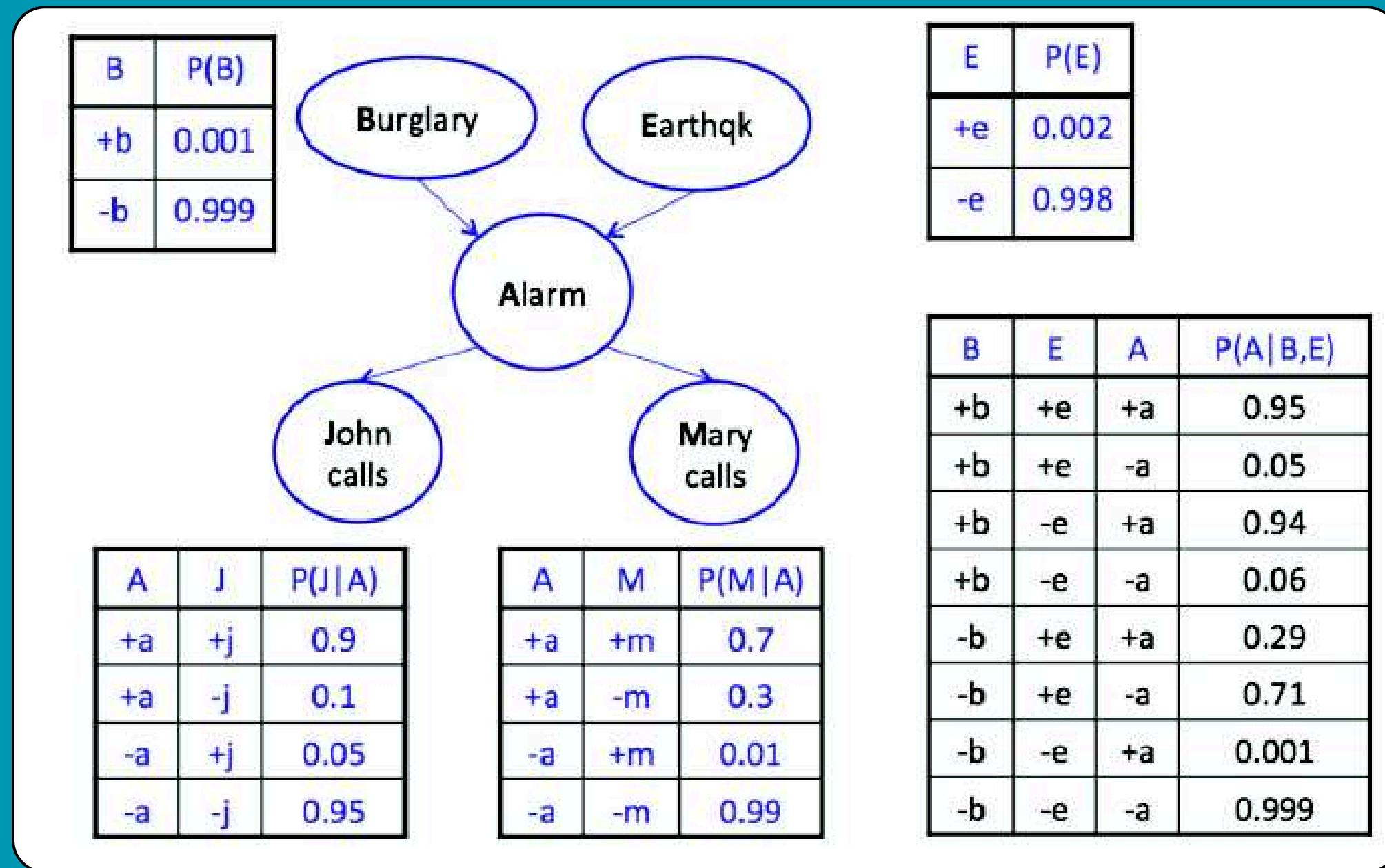
# Ejemplo Redes Bayesianas – Robo



Para este problema en particular, las variables son:

- Burglary (B): Robo.
- Earthquake (E): Terremoto.
- Alarm (A): Alarma.
- John calls (J): Llamada de John.
- Mary calls (M): Llamada de Mary.

# Ejemplo Redes Bayesianas – Robo

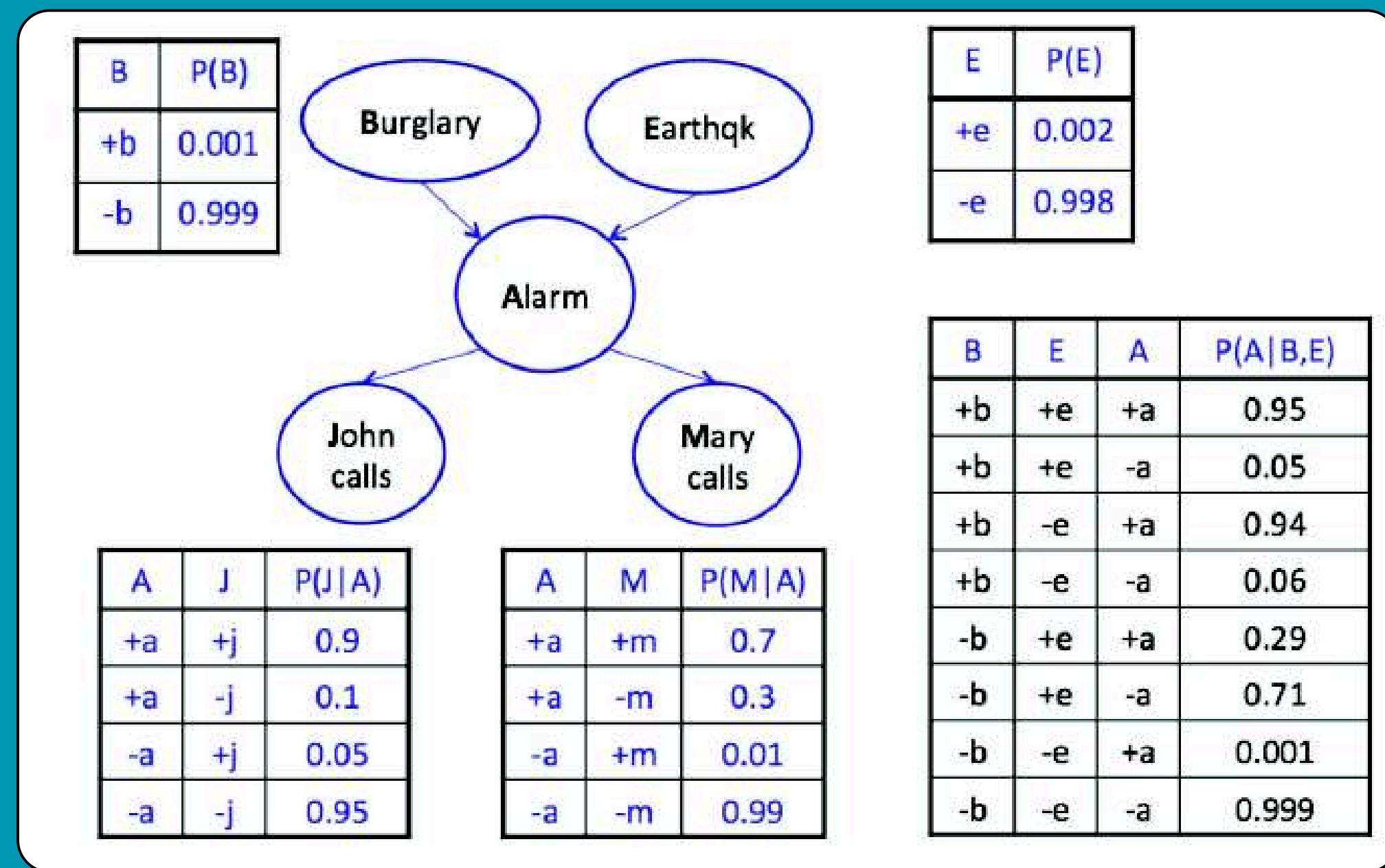


**El orden debe respetar la estructura de la red, es decir, si X es padre de Y, entonces X debe aparecer antes que Y en el orden:**

- Burglary (B) y Earthquake (E) son variables independientes.
- Alarm (A) depende de B y E.
- John calls (J) y Mary calls (M) dependen de A.

Un orden válido es: B,E,A,J,M

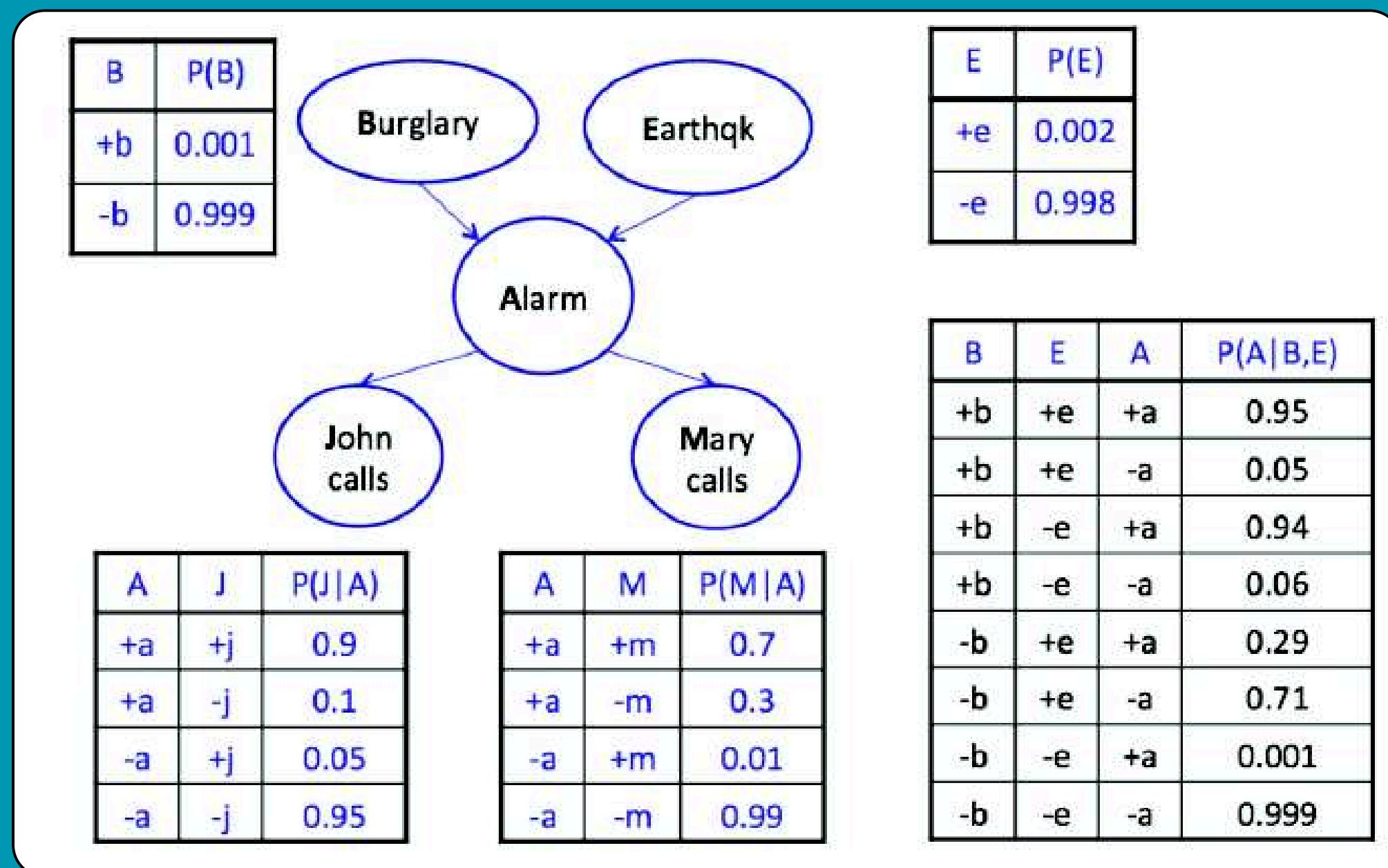
# Ejemplo Redes Bayesianas – Robo



**¿Cuál es la probabilidad que se active la alarma, es debido a un robo y se llama únicamente a Mary?**

- B: Robo (Burglary)
  - B=+b significa que hubo un robo
  - B=-b significa que no hubo un robo
- E: Terremoto (Earthquake)
  - E=+e terremoto
  - E=-e sin terremoto
- ...
- M: Mary llama (Mary calls)
  - M=+m Mary llama
  - M=-m Mary no llama

# Ejemplo Redes Bayesianas – Robo



Queremos encontrar la probabilidad de que la alarma se active debido a un robo y que solo Mary llame.

Esto se puede expresar como:

$$P(+b, -e, +a, -j, +m)$$

Probabilidad de que la alarma suene, haya robo, no haya terremoto, Mary llame, y John no

# Ejemplo Redes Bayesianas – Robo

$$\begin{aligned} P(+b, -e, +a, -j, +m) &= \\ P(+b)P(-e)P(+a|+b, -e)P(-j|+a)P(+m|+a) &= \\ 0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7 \end{aligned}$$

- La regla de la cadena codifica distribuciones conjuntas en secuencia de variables
- Las redes de Bayes asumen que la influencia directa son los padres
- Es comprobable a través del grafo

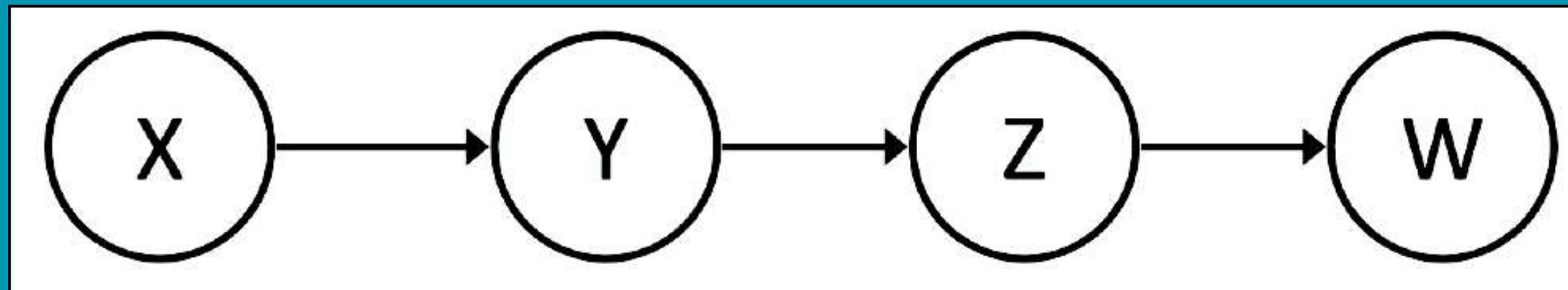
$$P(x_i|x_1 \cdots x_{i-1}) = P(x_i|\text{parents}(X_i))$$



# D-SEPARATION



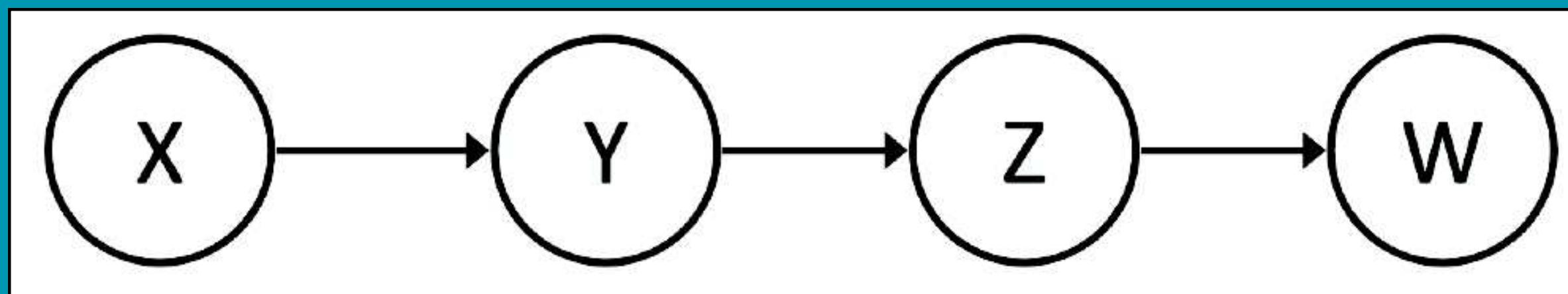
# ¿Son dos nodos independientes dada la evidencia?



Para determinar si dos nodos son independientes dado otro nodo, debemos ver si la influencia de uno sobre el otro se ve interrumpida por la evidencia.

# ¿Son dos nodos independientes dada la evidencia?

Digamos que queremos saber si "Estudiar" (X) y "Pasar el Examen" (W) son independientes, dado el conocimiento de "Confianza" (Z)



**Estudio:** Si el estudiante estudia o no.

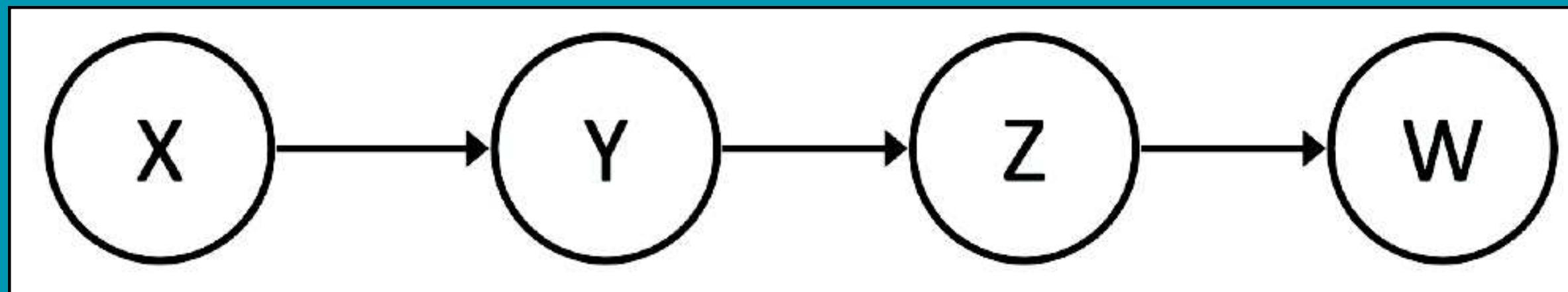
**Dormir Bien:** Si el estudiante durmió bien la noche anterior al examen.

**Confianza:** El nivel de confianza del estudiante al entrar al examen.

**Pasar el Examen:** Si el estudiante pasa o no el examen.

# ¿Son dos nodos independientes dada la evidencia?

Conociendo Z...



Si sabemos el nivel de confianza (Z), entonces la influencia de estudiar (X) en pasar el examen (W) se modula a través de la confianza (Z).

En este caso, estudiar (X) y pasar el examen (W) son independientes dado que conocemos la confianza (Z).

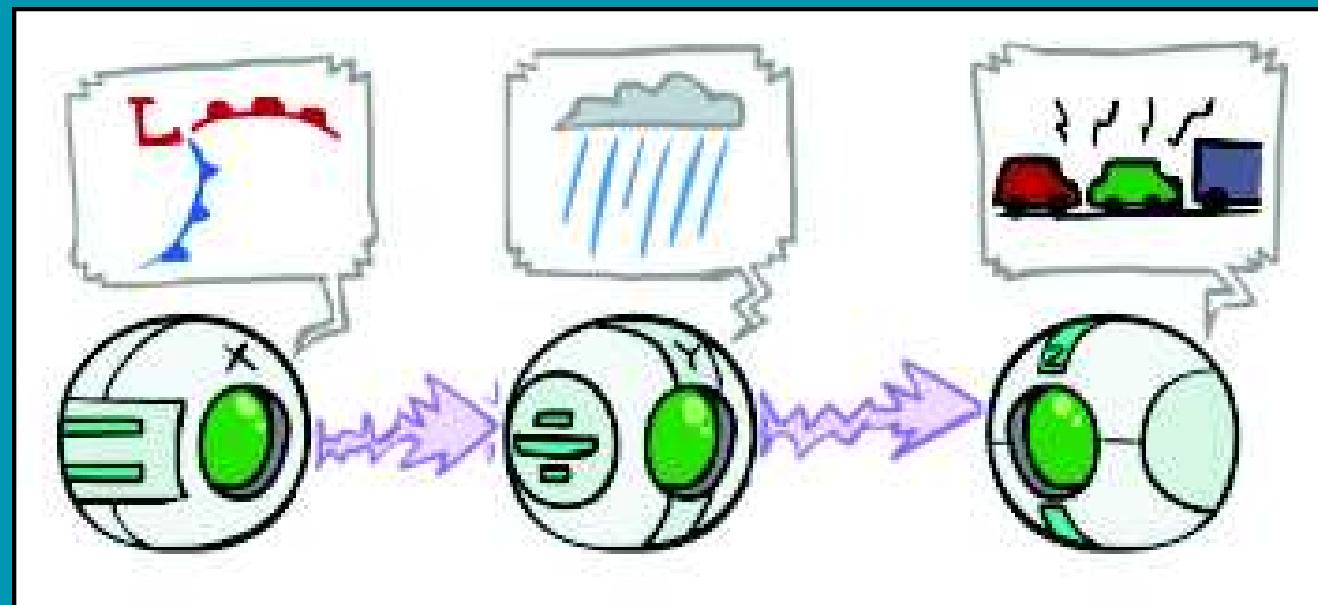
# D-Separation

En una red bayesiana, dos nodos A y B están d-separados (es decir, son condicionalmente independientes) dado un conjunto de evidencia E, si todos los caminos entre A y B quedan “bloqueados” por E.

Existen tres patrones básicos a tener en cuenta:

- **Casual Chain**
- **Common Case**
- **Common Effect**

# D-Separation - Casual Chain

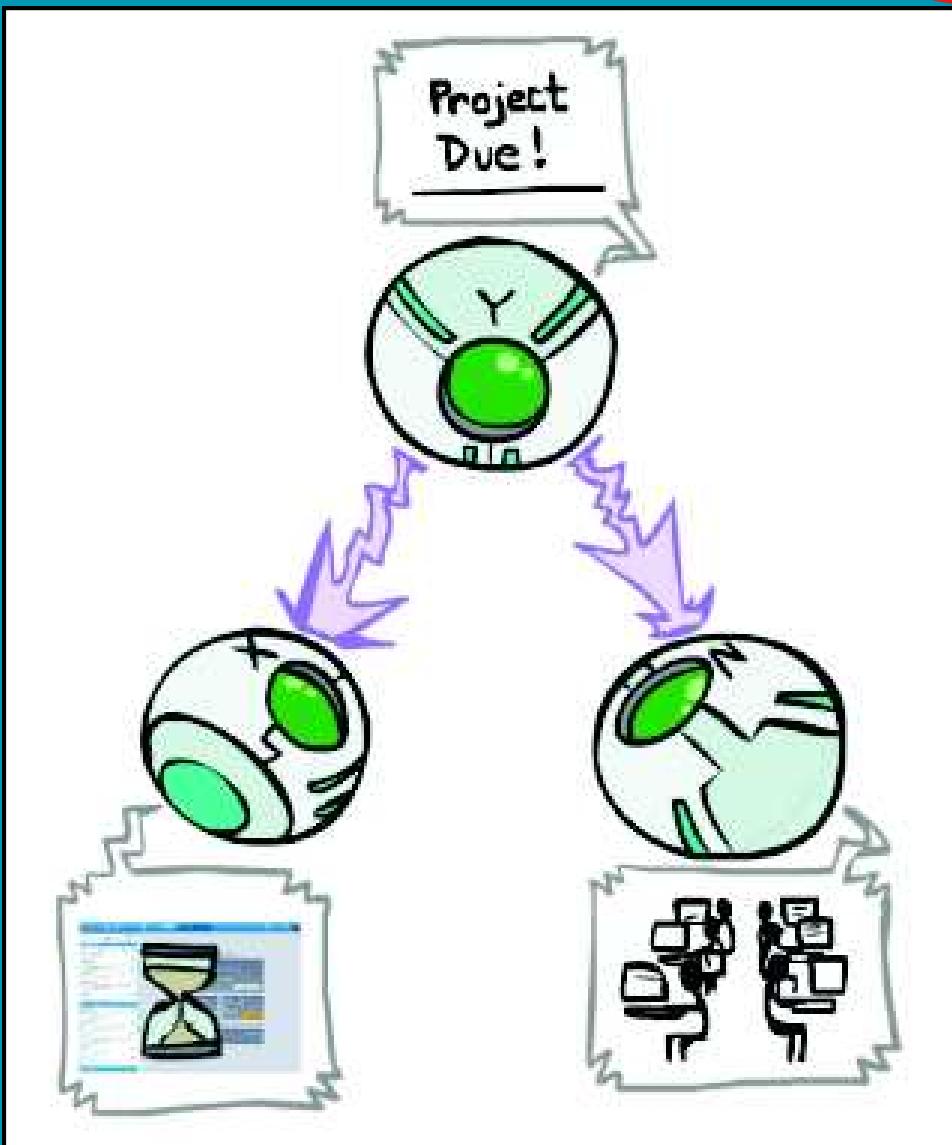


$$\begin{aligned} P(z|x,y) &= \frac{P(x,y,z)}{P(x,y)} \\ &= \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)} \\ &= P(z|y) \end{aligned}$$

$$P(x,y,z) = P(x)P(y|x)P(z|y)$$

**Sin condicionar en Y, el camino está abierto (dependencia).**  
**Si condicionas en Y, bloqueas el camino (independencia).**

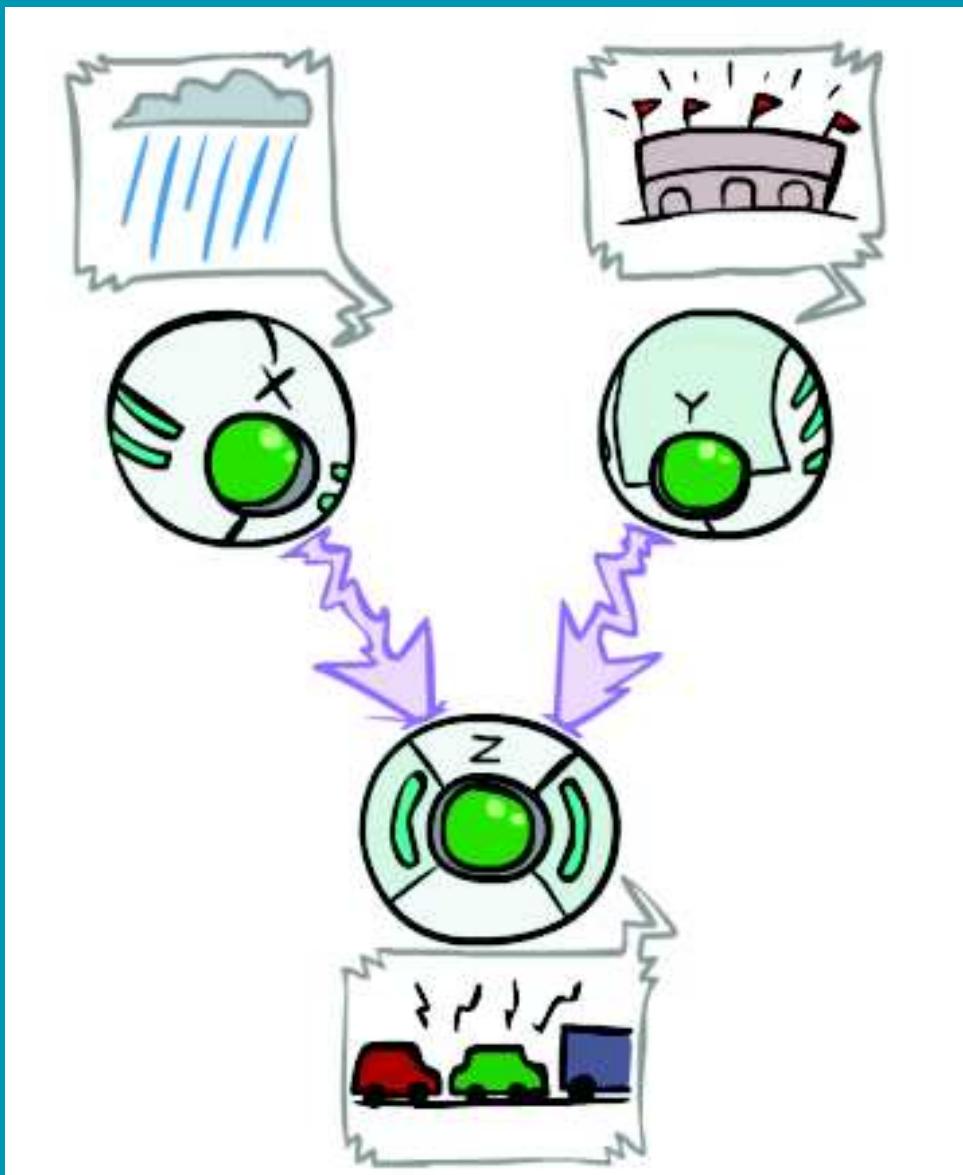
# D-Separation - Common Cause



$$\begin{aligned} P(z|x,y) &= \frac{P(x,y,z)}{P(x,y)} \\ &= \frac{P(y)P(x|y)P(z|y)}{P(y)P(x|y)} \\ &= P(z|y) \end{aligned}$$

Sin condicionar en Y, el camino está abierto (dependencia).  
Si condicionas en Y, bloqueas el camino (independencia).

# D-Separation - Common Effect

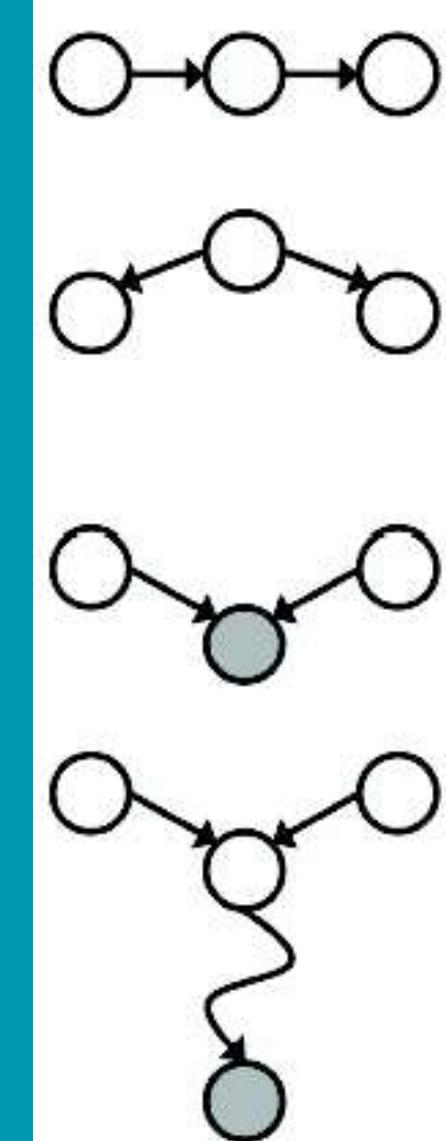


- Sin condicionar en  $Z$ , el camino está bloqueado (independencia).
- Si condicionas en  $Z$  (o en un descendiente de  $Z$ ), se abre el camino (dependencia).

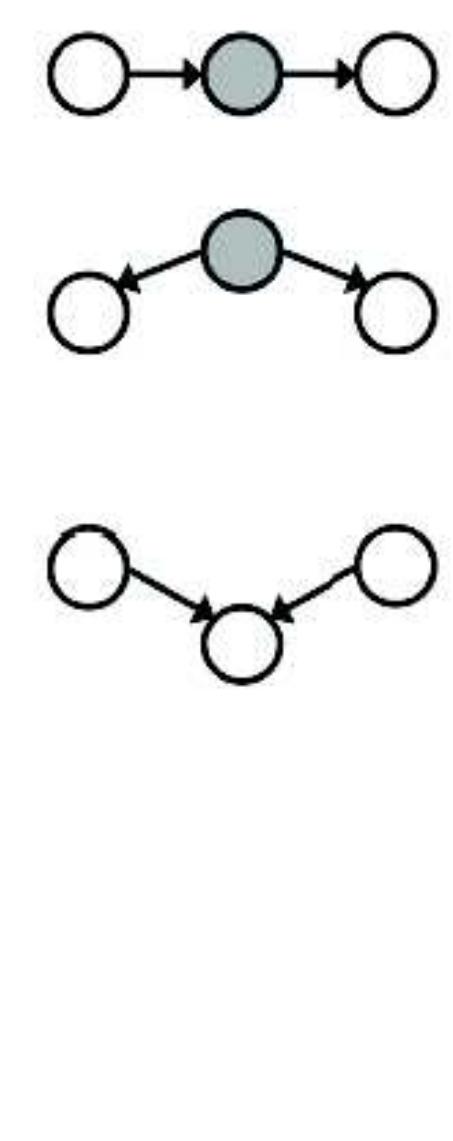
# D-Separation

- Debe existir camino entre las variables
- Una vez se establece el(los) camino(s) se analizara por triadas
- Un camino es activo si todas las triadas son activas
- Si hay un camino activo, no hay independencia

Triada Activa



Triada Inactiva





# INFERENCIA

Dada una Red Bayesiana, ¿Cuál es  $P(X | e)$ ?



# Inferencia

Probabilidad posterior

$$P(Q|E_1 = e_1, \dots, E_k = e_k)$$

Evidence variables:  $E_1 \dots E_k = e_1 \dots e_k$   
Query\* variable:  $Q$   
Hidden variables:  $H_1 \dots H_r$

$X_1, X_2, \dots, X_n$   
*All variables*

# Inferencia por Enumeracion

La inferencia por enumeración en redes bayesianas consiste en calcular la probabilidad conjunta de todas las variables relevantes mediante la suma (o marginalización) de todas las combinaciones posibles de valores para las variables no observadas.

# Inferencia por Enumeracion

Supongamos que queremos resolver:

$$P(+b| +j, +m)$$

Entonces:

$$P(+b| +j, +m) = \frac{P(+b, +j, +m)}{P(+j, +m)}$$

$$P(+b, +j, +m) = \sum_a \sum_e P(+b)P(e)P(a|b, e)P(j|a)P(m|a)$$

# Inferencia por Enumeracion

$$\begin{aligned} P(+b, +j, +m) = & \quad P(+b)P(+e)P(+a|+b, +e)P(+j|+a)P(+m|+a) + \\ & P(+b)P(+e)P(-a|+b, +e)P(+j|-a)P(+m|-a) + \\ & P(+b)P(-e)P(+a|+b, -e)P(+j|+a)P(+m|+a) + \\ & P(+b)P(-e)P(-a|+b, -e)P(+j|-a)P(+m|-a) \end{aligned}$$

Calculamos toda (o casi toda la distribución conjunta) y luego enumeramos variables.

# Inferencia por Enumeración

Este enfoque es conceptualmente sencillo, presenta algunas desventajas importantes:

**Complejidad Exponencial:** La cantidad de combinaciones posibles crece exponencialmente con el número de variables.

**Ineficiencia Computacional:** Evaluar todas las combinaciones resulta impracticable para redes grandes.

**Alto Consumo de Memoria:** Guardar todas las combinaciones posibles consume una gran cantidad de memoria.

**Redundancia de Cálculo:** Muchas combinaciones son redundantes, ya que se vuelven a calcular probabilidades parciales repetidamente.



# ELIMINACION DE VARIABLES



# Eliminación de Variables

La técnica de eliminación de variables aborda estas desventajas al evitar la enumeración completa de todas las combinaciones.

**Agrupación de Sumas:** Se agrupan y simplifican sumas parciales antes de realizar la marginalización.

**Factorización Inteligente:** Se explota la estructura de independencia de la red para calcular solo las combinaciones necesarias.

**Uso de Factores:** Se representan los factores condicionales en lugar de trabajar directamente con la distribución conjunta completa.

# Eliminación de Variables – Factor 1

Podemos identificar los diferentes factores o elementos que estan involucrados dentro de una red bayesiana

Distribución conjunta

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Distribución conjunta parcial  
(selected)  $P(x, Y)$

$P(cold, W)$

T	W	P
cold	sun	0.2
cold	rain	0.3

# Eliminacion de Variables – Factor 2

Condicional simple  
(suma 1)  $P(Y|X)$

$$P(W|cold)$$

T	W	P
cold	sun	0.4
cold	rain	0.6

Familia de condicionales  $P(Y|X)$

$$P(W|T)$$

T	W	P
hot	sun	0.8
hot	rain	0.2
cold	sun	0.4
cold	rain	0.6

$$P(W|hot)$$

$$P(W|cold)$$

# Eliminacion de Variables – Factor 3

Familia especificada  
(consecuencias)  $P(y|X)$

$$P(rain|T)$$

T	W	P
hot	rain	0.2
cold	rain	0.6

$P(rain|hot)$   
 $P(rain|cold)$

Una vez identificados los elementos o factores dentro de la red bayesiana podemos empezar el proceso de eliminacion con el siguiente ejemplo

# Ejemplo - Eliminacion de Variables

Hoy juegan los rojos, puede que haya trafico y llegue tarde a clase.

## Variables

Juegan los rojos (**R**)

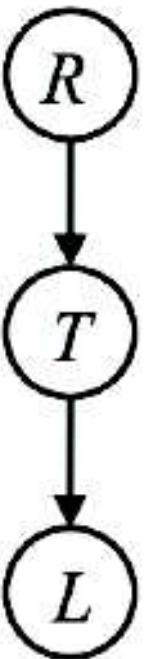
Trafico (**T**)

Llegue tarde a clase (**L**)

¿Llego tarde a clase?

# Ejemplo - Eliminación de Variables

Dada la red bayesiana



$$P(R)$$

+r	0.1
-r	0.9

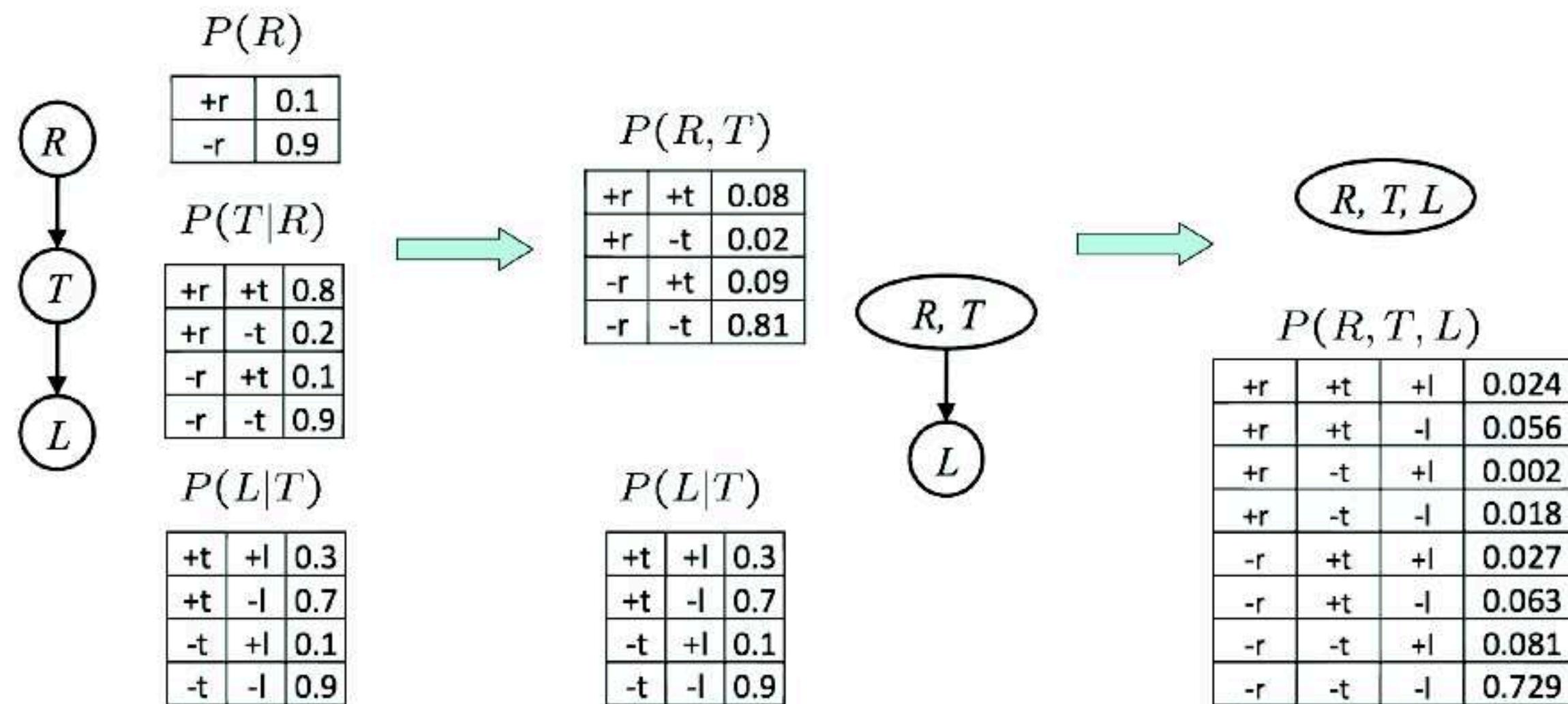
$$P(T|R)$$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

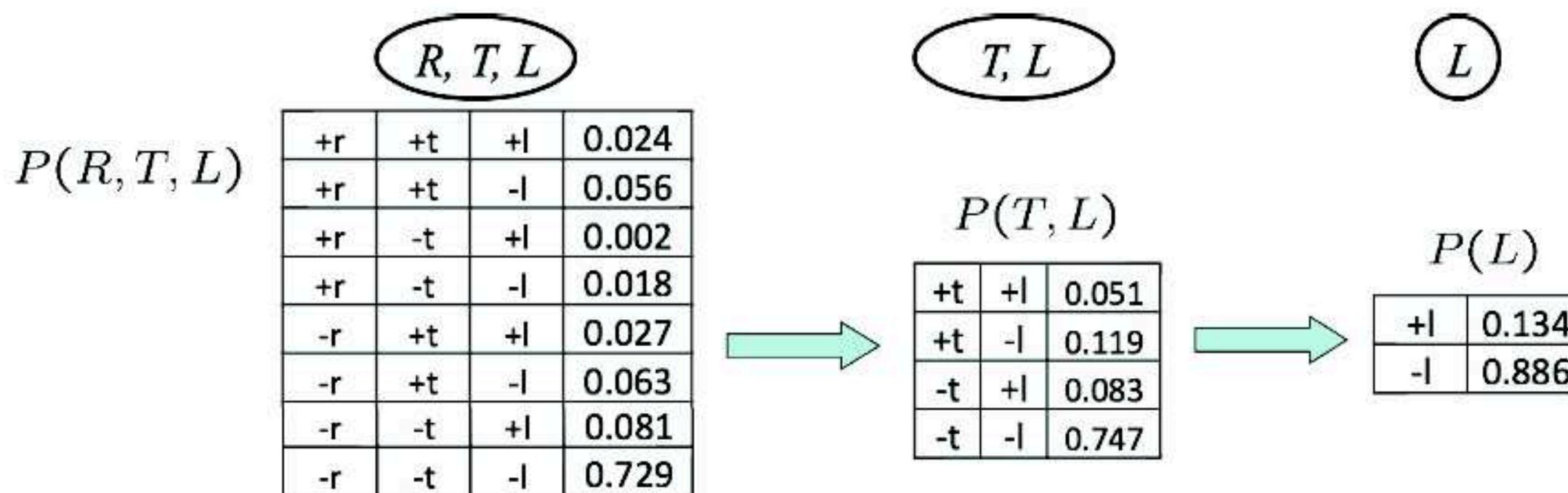
$$P(L|T)$$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

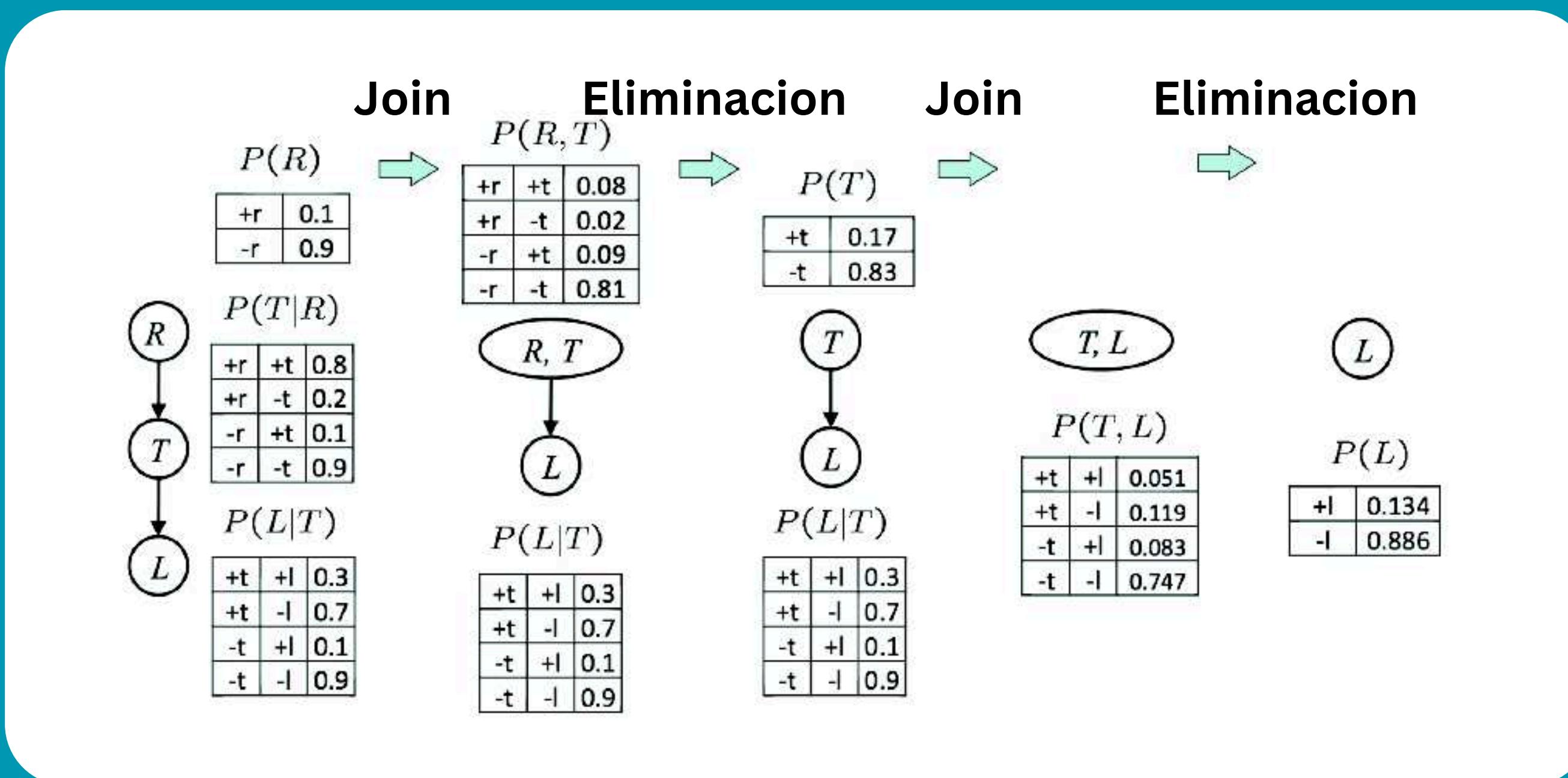
# EV - Join



# EV - Eliminacion - Suma



# EV - Alternativa



# Ejemplo - Eliminación de Variables

Dada la red bayesiana



$P(R)$	
+r	0.1
-r	0.9

$P(T R)$		
+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L T)$		
+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Consultar

$$P(L|r)$$

# Ejemplo - Eliminacion de Variables

Aplicamos condicion simple  $P(L | +r)$ , eliminamos  $-r$

$$P(+r)$$

+r	0.1
----	-----

$$P(T|+r)$$

+r	+t	0.8
+r	-t	0.2

$$P(L|T)$$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

# Ejemplo - Eliminacion de Variables

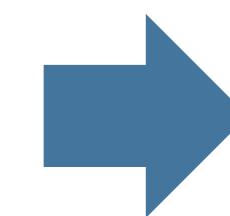
Join

$$P(+r)$$

+r	0.1
----	-----

$$P(T| +r)$$

+r	+t	0.8
+r	-t	0.2



$$P(T,+r)$$

+r	+t	0.08
+r	-t	0.02

# Ejemplo - Eliminacion de Variables

Join

$P(T,+r)$

+r	+t	0.08
+r	-t	0.02

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$P(L,T,+r)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.02
+r	-t	-l	0.018

# Ejemplo - Eliminacion de Variables

## Eliminacion T

$P(+r, L)$

+r	+l	<b>0.026</b>
+r	-l	<b>0.074</b>

# Ejemplo - Eliminación de Variables

## Normalizar

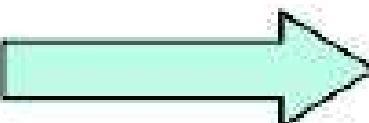
$$P(+r, L)$$

+r	+l	0.026
+r	-l	0.074

$$P(L | +r) = \frac{P(L, +r)}{P(+r)}$$

$$P(L | + r)$$

+l	0.26
-l	0.74



$$P(+r) = P(+r, +l) + P(+r, -l) = 0.026 + 0.074 = 0.1$$

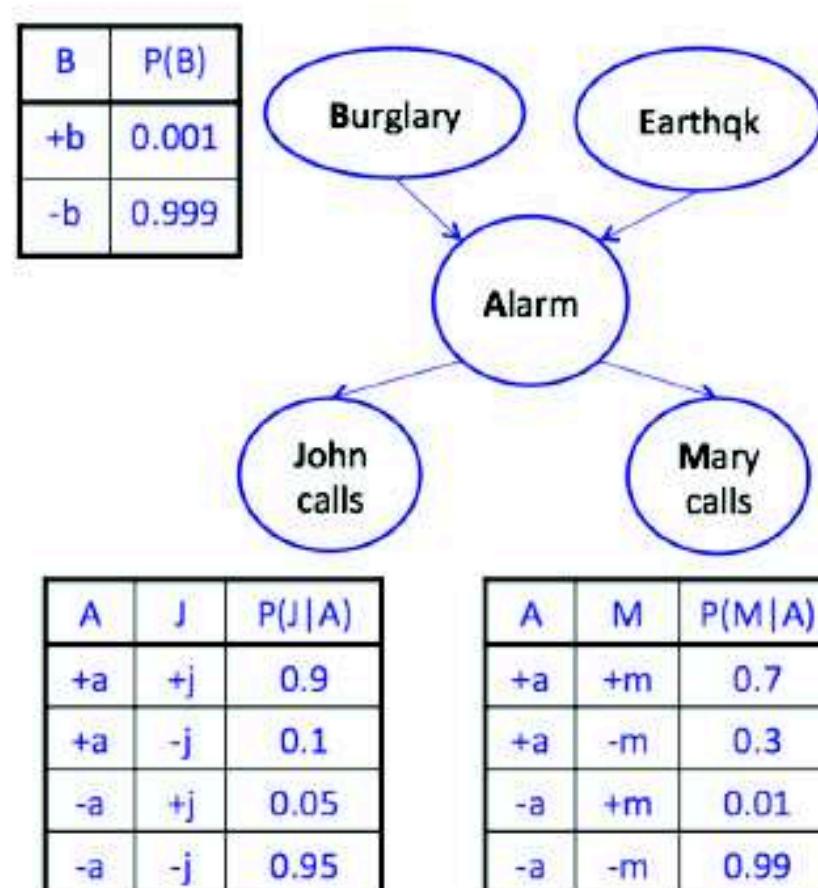
$P(+r)$  es la probabilidad marginal de  $+r$ , que se obtiene sumando  $P(L, +r)$  sobre todos los valores posibles de  $L$ . Así, los valores de la tabla se “reescalan” de modo que sumen 1, convirtiéndose en una dist. condicional

# PROCEDIMIENTO GENERAL

- 1) Obtenemos una consulta
- 2) Iniciamos con los factores (probabilidades conjuntas c/ evidencia)
- 3) Mientras existan variables ocultas se hace join y eliminaciones
- 4) Normalizamos

# Ejemplo - EV en Red Bayesiana - Robo

$P(B|j, m)$   
Consulta



B	E	A	$P(A B,E)$
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

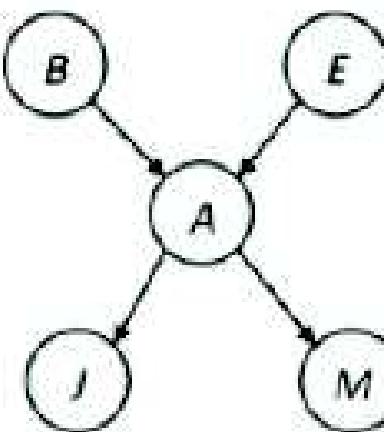
# Ejemplo - EV en Red Bayesiana - Robo

## Identificamos Variables

Consulta: B

Evidencia: j,m

Ocultas: E, A



## Identificamos involucradas en la red segun variables

$$P(B|j, m) \propto P(B, j, m)$$

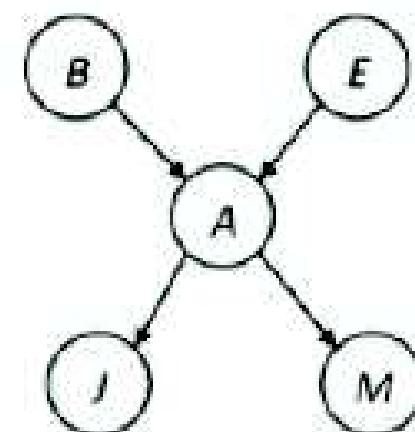
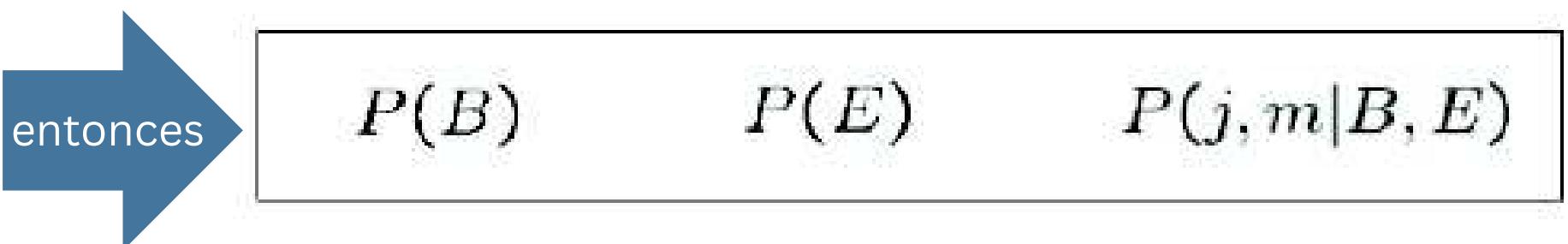
$P(B)$	$P(E)$	$P(A B, E)$	$P(j A)$	$P(m A)$
--------	--------	-------------	----------	----------

# Ejemplo - EV en Red Bayesiana - Robo

**Buscamos eliminar variables ocultas, empezando con A**

**A**

$$\begin{array}{ccc} P(A|B, E) & \text{Join} & \text{Eliminacion} \\ P(j|A) & \xrightarrow{\times} & P(j, m, A|B, E) \xrightarrow{\sum} P(j, m|B, E) \\ P(m|A) & & \end{array}$$



# Ejemplo - EV en Red Bayesiana - Robo

Continuamos eliminando E, que es una variable oculta

despues de  
remover A

$$P(B)$$

$$P(E)$$

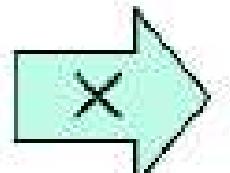
$$P(j, m|B, E)$$

E

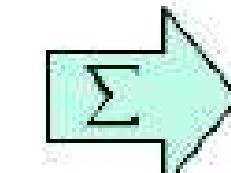
Join

$$P(E)$$

$$P(j, m|B, E)$$



Eliminacion

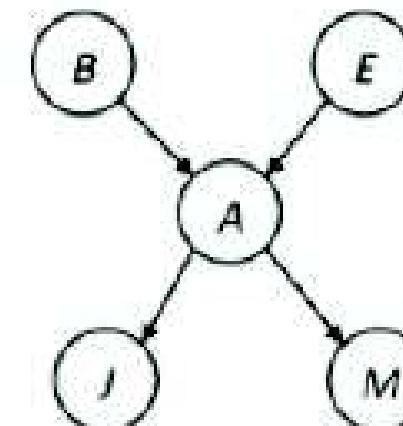


$$P(j, m|B)$$

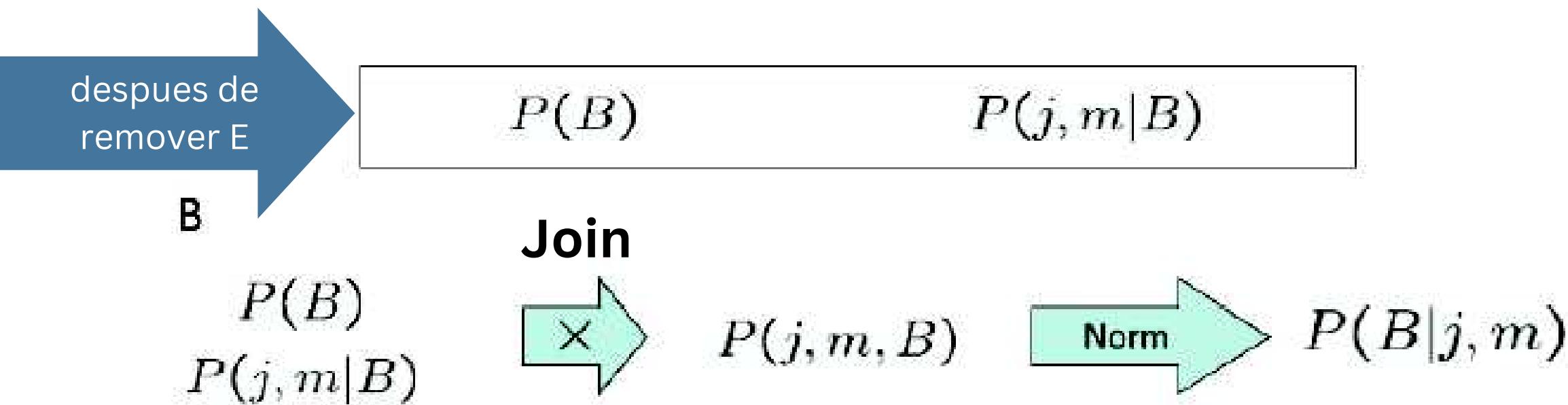
despues de  
remover E

$$P(B)$$

$$P(j, m|B)$$



# Ejemplo - EV en Red Bayesiana - Robo

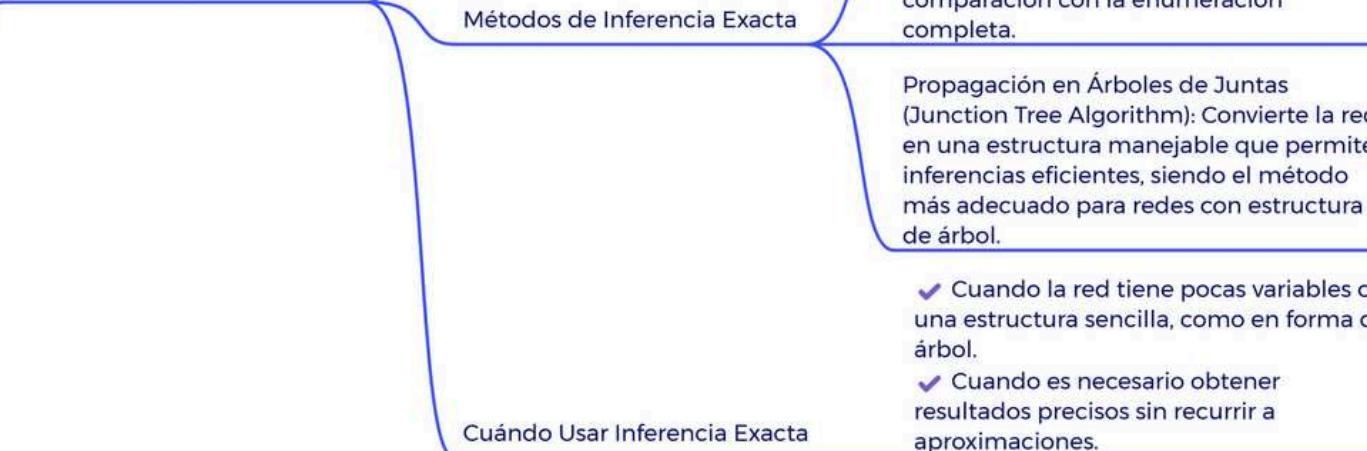


**Ya tenemos las variables, solo queda hacer un ultimo  
Join y normalizar**

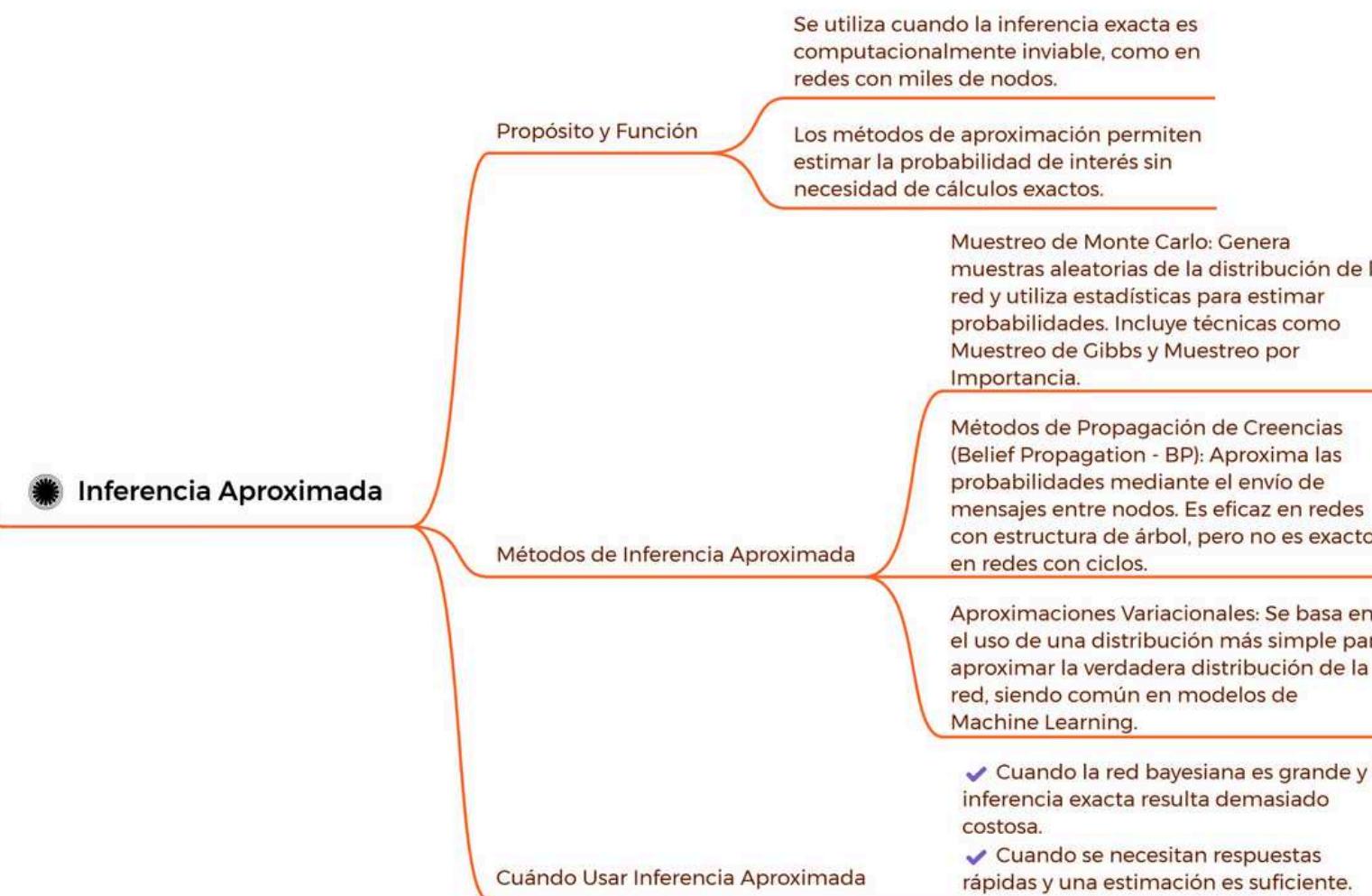
## Tipos de Inferencia en Redes Bayesianas



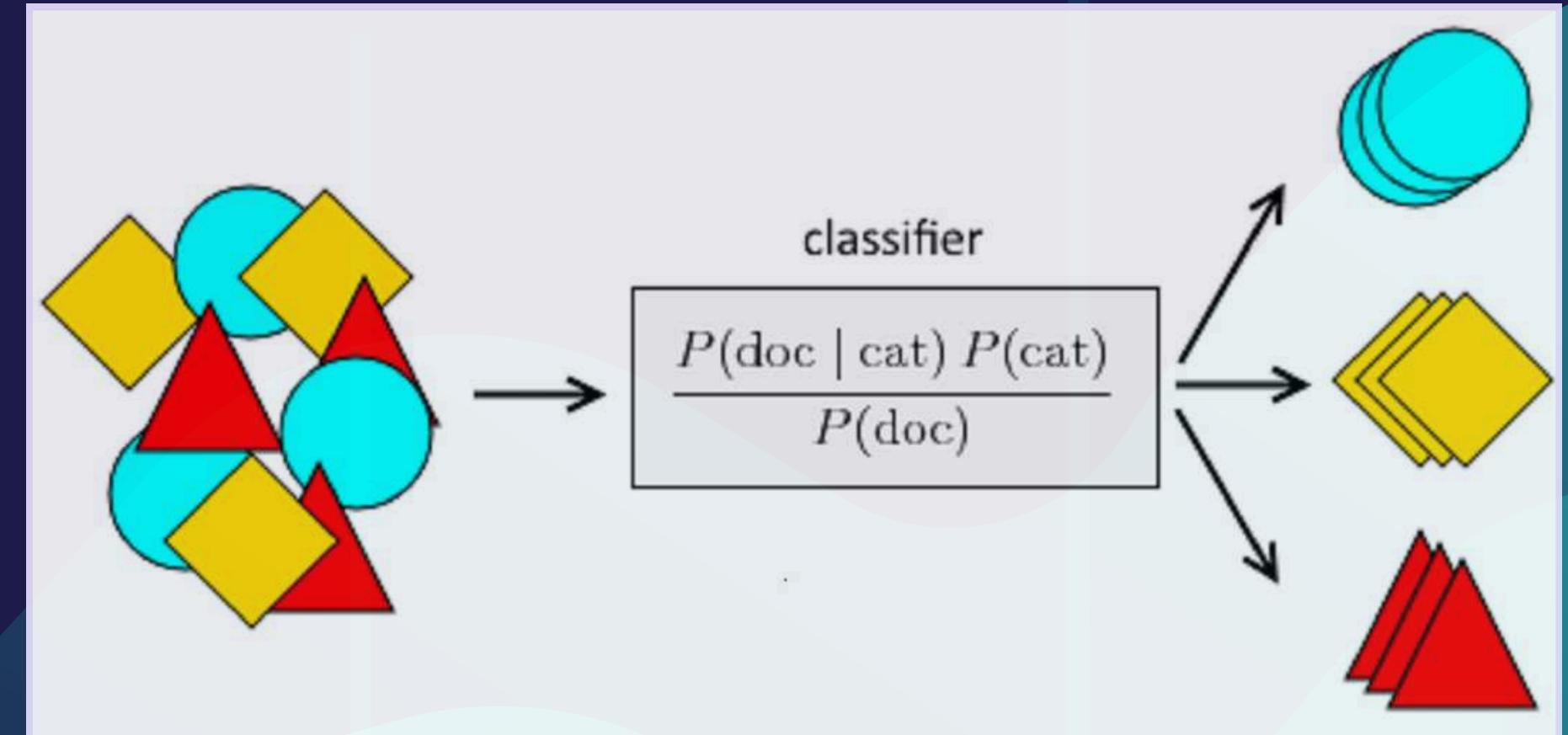
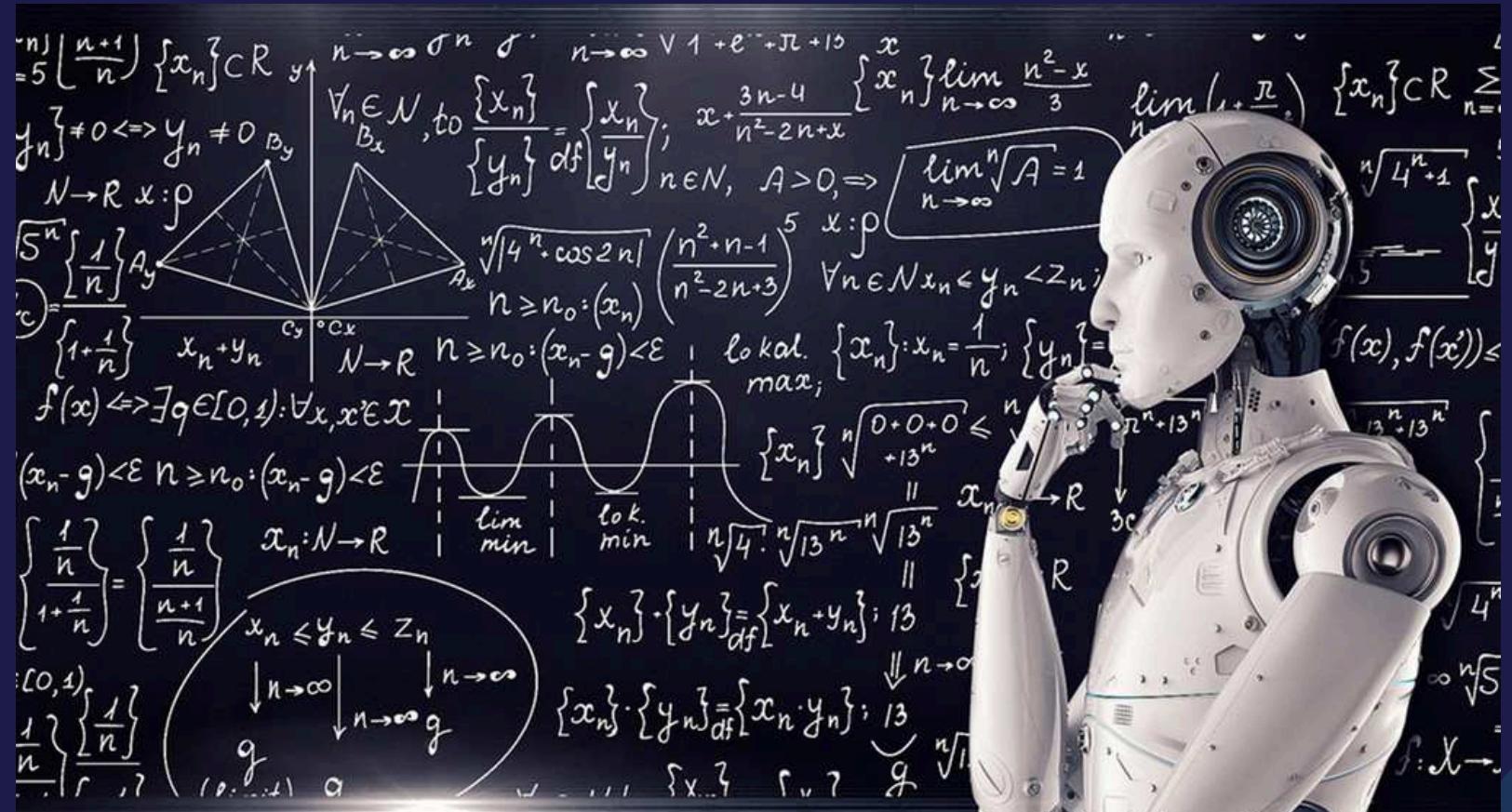
### Inferencia Exacta



### Inferencia Aproximada



# Naive Bayes



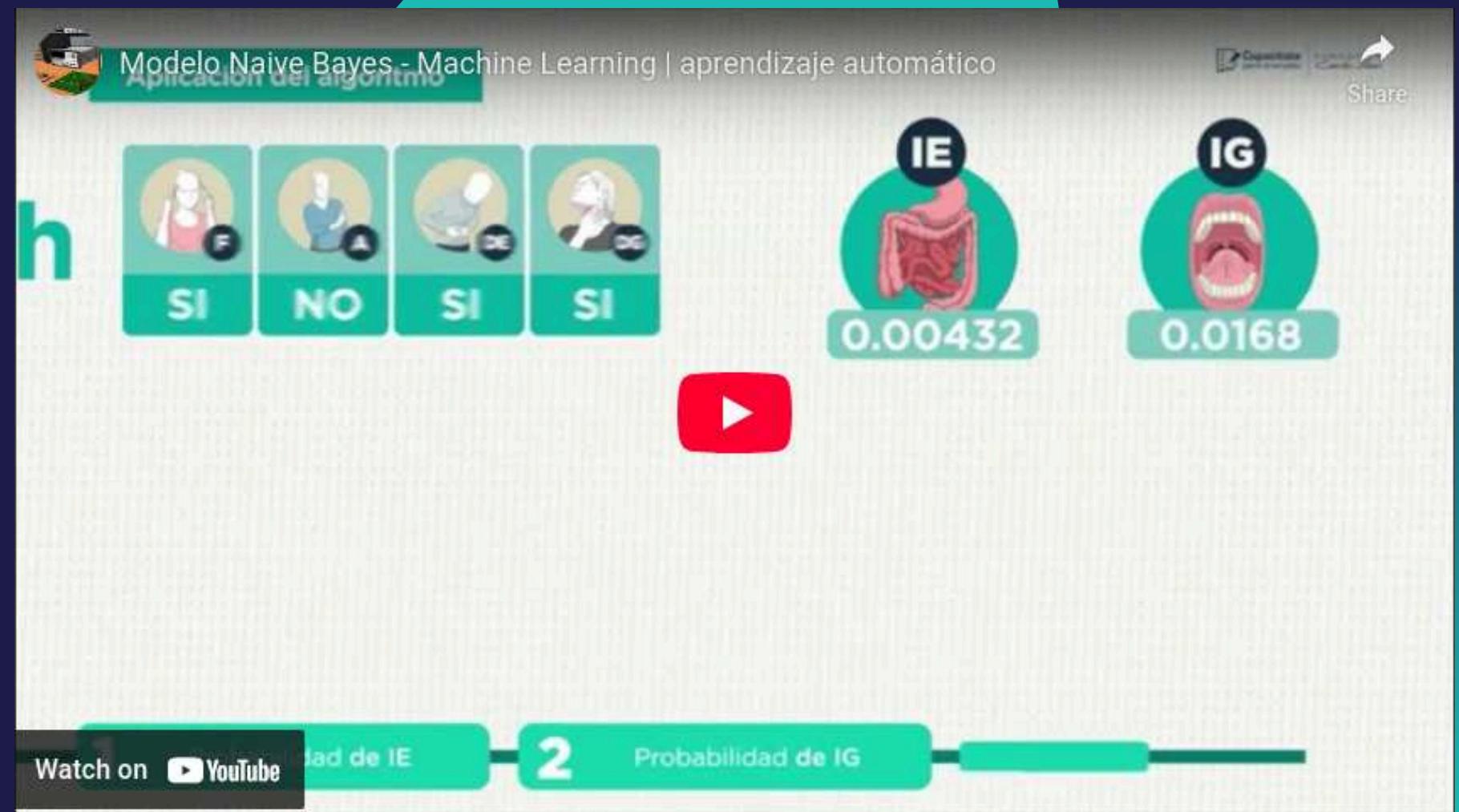
# Agenda

- ✓ Naïve Bayes como Clasificador
- ✓ Clasificacion Basada en Modelos
- ✓ Ejemplos de clasificacion y aplicaciones
- ✓ Algoritmo Naïve Bayes
- ✓ Maximum likelihood
- ✓ Over-fitting
- ✓ K-fold
- ✓ Laplace Smoothing
- ✓ Suma de Logaritmos

# Naïve Bayes

Es una clase especial de algoritmo de clasificación de Aprendizaje Automatico, o Machine Learning

Proporcionan una manera fácil de construir modelos con un comportamiento muy bueno debido a su simplicidad



# Naïve Bayes

El principio de parsimonia o navaja de Ockham en el contexto de Machine Learning y clasificadores como Naive Bayes, se refleja en la preferencia por modelos simples que generalicen bien en lugar de modelos complejos que puedan sobreajustarse a los datos.

En este caso se asume por simplificación que:

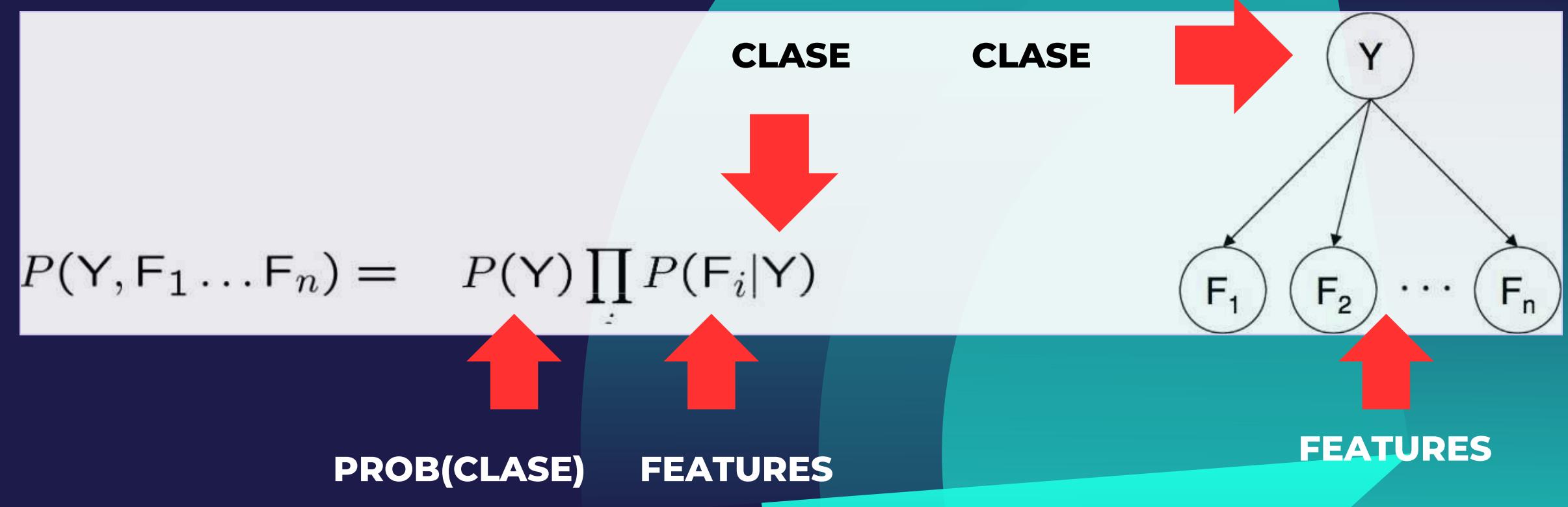
***Las características (features) son independientes entre sí, dado el resultado (clase).***

**Principio de simplicidad o parsimonia:**  
"En igualdad de condiciones, la explicación más sencilla, suele ser la correcta".



# Naïve Bayes

Naïve Bayes, asume una red bayesiana donde los atributos (*features*) son independientes entre si, pero las features dependen de la clase



- La **clase** representa la clasificación a la que puede pertenecer una instancia
- La **feature** representa alguna de las características o parámetros que se tienen de esa instancia.

ES COMO UN CHEF QUE USA UNA RECETA (EL TEOREMA DE BAYES) PARA COMBINAR INGREDIENTES (CARACTERÍSTICAS) Y HACER UN PLATO (LA CLASIFICACIÓN)

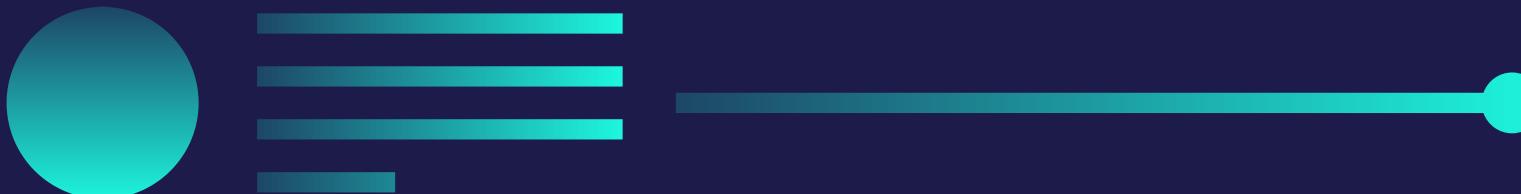
# Clasificacion basada en Modelos





HASTA AHORA

*¿Como hemos modelado  
sistemas con redes  
bayesianas?*



# Modelo



## Variables

Definir parámetros o variables (evidencia, consulta, ocultas)



## Estructura

Definir estructura mediante una Red bayesiana (BN) y Grafos



## Probabilidades

Se usan las reglas de probabilidad condicional y el teorema de Bayes



## Inferencia

Una vez modelada la red, se pueden responder las preguntas consultadas

# Clasificación

Consiste en asignar una etiqueta o clase a una muestra de datos en función de sus características.

Los clasificadores se entrena con datos etiquetados para aprender a predecir la clase de nuevas muestras.

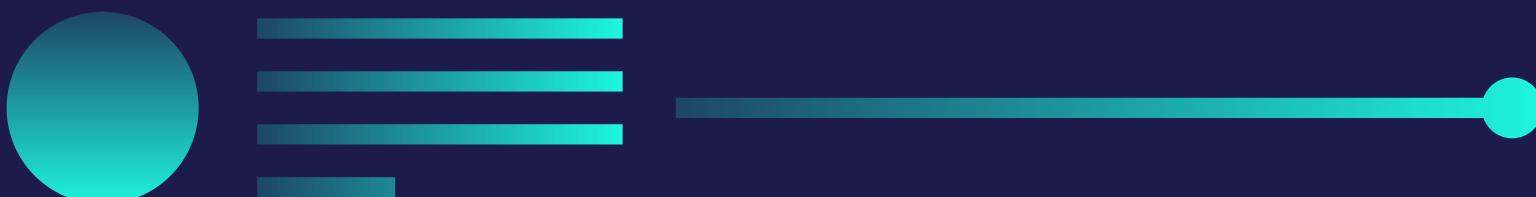
Para clasificar necesitamos:

1. Recopilación de Datos
2. Preprocesamiento de Datos
  - a. Limpiar los datos
  - b. Escalar o normalizar las características
3. División del Conjunto de Datos
  - a. Conjunto de entrenamiento (70%)
  - b. Conjunto de prueba (15%)
  - c. Conjunto de validacion (15%)
4. Selección del Modelo (Naive Bayes)
5. Entrenamiento del Modelo
6. Evaluación del Modelo



PERO...

*¿Cómo evaluarían si un  
modelo es bueno o malo  
para clasificar?*



# ¿Qué son las métricas de evaluación?



## Definición

Herramientas para medir el rendimiento de un modelo.



## Generalización

Un modelo con alta precisión en los datos de entrenamiento no necesariamente generaliza bien.



## Decidir

Permiten comparar diferentes modelos y elegir el mejor.



## Guían

Informan qué aspectos del modelo deben ajustarse para optimizar su rendimiento.

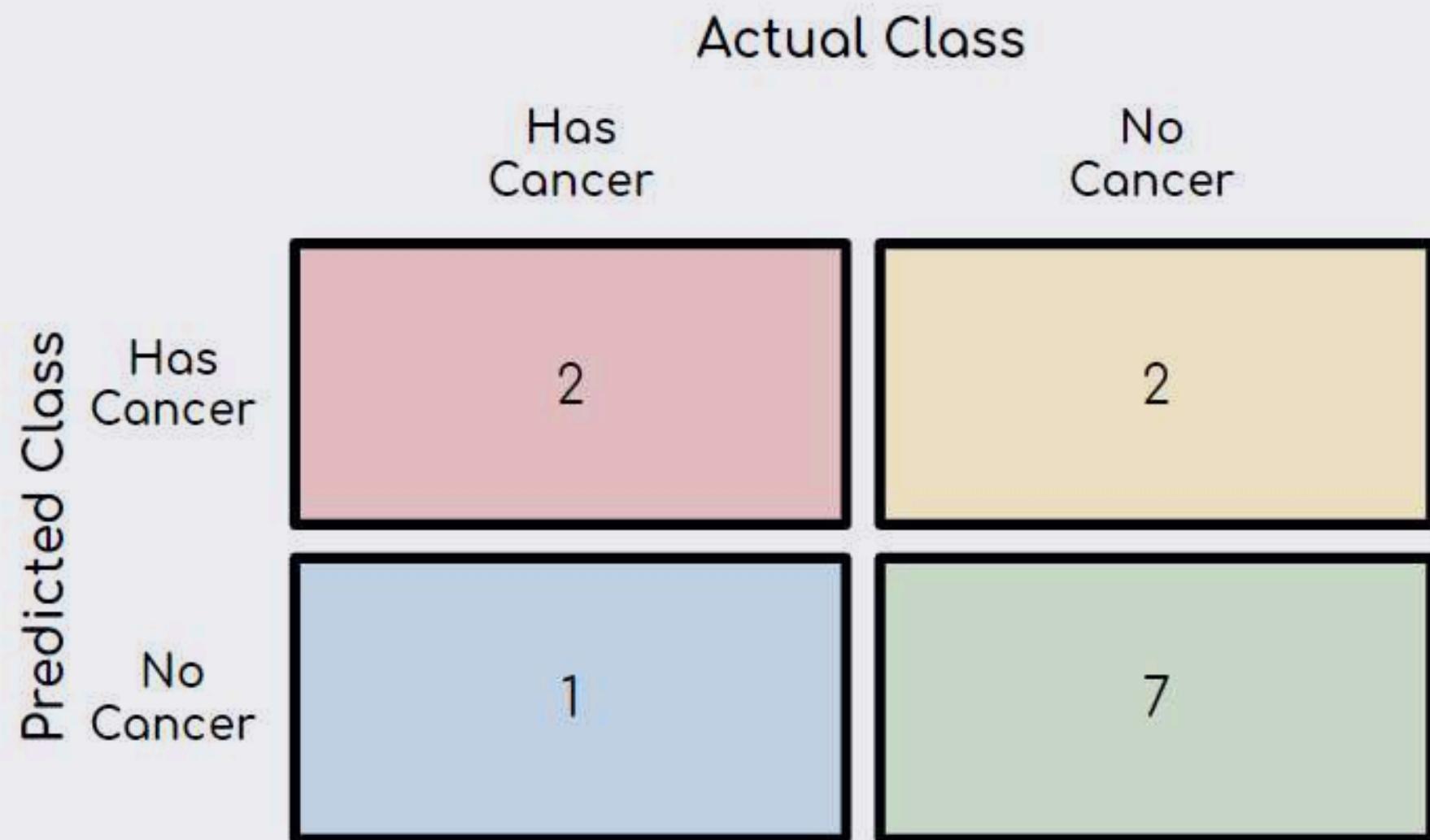
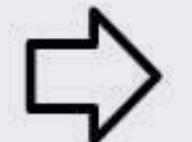
# Matriz de Confusión

Una tabla que muestra las predicciones del modelo versus las clases reales.

		Valor Actual o Real	
		Positivo	Negativo
Valor Predicho	Positivo	TP	FP
	Negativo	FN	TN

# Matriz de Confusión

Predicted Class	Actual Class
No Cancer	No Cancer
Has Cancer	Has Cancer
No Cancer	No Cancer
No Cancer	No Cancer
No Cancer	Has Cancer
Has Cancer	Has Cancer
No Cancer	No Cancer
Has Cancer	No Cancer
No Cancer	No Cancer
No Cancer	No Cancer
Has Cancer	No Cancer
No Cancer	No Cancer



# Evaluación del Modelo

Usar métricas como:

- **Exactitud (Accuracy)**: mide qué tan bien el modelo clasifica correctamente

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- TP (True Positive): Clasificado correctamente como positivo.
- TN (True Negative): Clasificado correctamente como negativo.
- FP (False Positive): Clasificado incorrectamente como positivo (falso positivo).
- FN (False Negative): Clasificado incorrectamente como negativo (falso negativo).

- **Precisión**: mide la proporción de ejemplos positivos predichos que realmente son positivos.

$$\text{Precision} = \frac{TP}{TP + FP}$$

**¿Qué métrica usarías si te importa más evitar falsos positivos?**

# Evaluación del Modelo

## Ejemplo comparativo

Tenemos un modelo de clasificación que predice si un tumor es maligno o benigno.

- Maligno: Tumor peligroso.
- Benigno: Tumor no peligroso.

Supongamos que tenemos 1000 tumores y el modelo realiza las siguientes predicciones:

1000 casos	Predictión: Maligno	Predictión: Benigno
Real: Maligno	30	70
Real: Benigno	20	880

# Clasificación

## Ejemplo comparativo

El **Accuracy** mide qué tan bien el modelo clasifica correctamente en general, tanto malignos como benignos.

$$\text{Accuracy} = \frac{70 + 880}{70 + 880 + 20 + 30} = \frac{950}{1000} = 0.95 \quad (95\%)$$

El modelo acierta en el 95% de los casos totales.

Si el conjunto de datos está muy desbalanceado (por ejemplo, 950 benignos y 50 malignos), el modelo puede lograr un accuracy alto simplemente prediciendo todo como benigno, por tanto,

**puede ser engañoso**

# Clasificación

## Ejemplo comparativo

La **Precision** mide qué porcentaje de los tumores clasificados como malignos realmente lo son.

$$\text{Precision} = \frac{70}{70 + 20} = \frac{70}{90} = 0.78 \quad (78\%)$$

De todos los tumores que el modelo clasificó como malignos, el 78% realmente lo eran.

# Evaluación del Modelo

Usar métricas como:

- **Recall (Sensibilidad o Tasa de Verdaderos Positivos)**: mide la capacidad del modelo para detectar correctamente los casos positivos (malignos).

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Recall} = \frac{70}{70 + 30} = \frac{70}{100} = 0.70 \quad (70\%)$$

El modelo identifica correctamente el 70% de los tumores malignos

- **F1-Score**: combina la **Precision** y el **Recall** en una única métrica para equilibrar ambos aspectos.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{F1-Score} = 2 \times \frac{0.78 \times 0.70}{0.78 + 0.70} = 2 \times \frac{0.546}{1.48} = 0.74 \quad (74\%)$$

Indica que el modelo logra un equilibrio razonable entre precisión y sensibilidad, útil en casos de clases desbalanceadas

# Evaluación del Modelo

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) <b>Type II Error</b>	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) <b>Type I Error</b>	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
	Precision	$\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TN + FP + FP + FN)}$

# Evaluación del Modelo

## ¿Qué métrica usarías si te importa más evitar falsos positivos?

La **precisión** mide qué proporción de las predicciones positivas realizadas por el modelo son correctas. Penaliza los falsos positivos, ya que estos afectan directamente su valor.

- **Ejemplo:** En un sistema de detección de fraudes, es importante minimizar los falsos positivos para no bloquear transacciones legítimas.

## ¿Qué métrica priorizarías en un sistema médico donde es crítico detectar todas las enfermedades?

El **recall** mide qué proporción de los casos positivos reales son correctamente identificados por el modelo. En un sistema médico, detectar todas las enfermedades (minimizar los falsos negativos) es crucial, incluso si eso implica aceptar algunos falsos positivos.

- **Ejemplo:** En la detección de cáncer, es preferible que el modelo clasifique erróneamente algunos casos como positivos (aunque no lo sean) a que pase por alto un caso real de cáncer.

# Resumen

Entonces:

**Accuracy**

Útil cuando las clases están equilibradas

**Precision**

Importante cuando el costo de un **falso positivo** es alto (como falsos diagnósticos positivos).

**Recall**

Importante cuando el costo de un **falso negativo** es alto (como pasar por alto un tumor maligno).

**F1-Score**

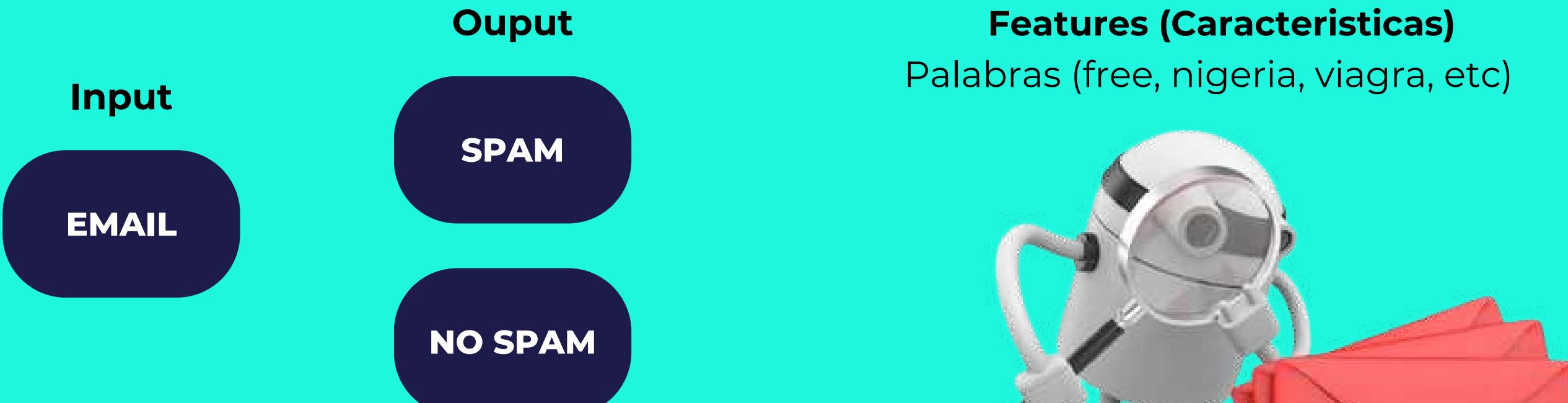
Útil cuando se necesita un equilibrio entre precisión y sensibilidad.

# Ejemplo de clasificación



# Ejemplo SPAM

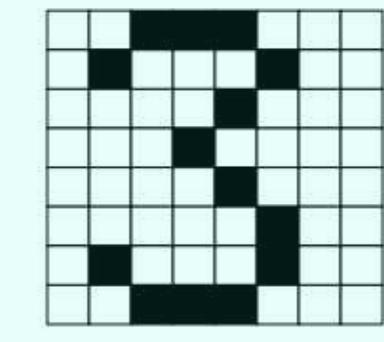
Supongamos que tenemos un correo electronico en el cual queremos identificar correo valido y correo SPAM



**OBTENER UNA GRAN CANTIDAD DE EMAILS, CLASIFICAR CADA UNO DE ELLOS COMO SPAM O HAM, APRENDER DE LOS EMAILS PREVIAMENTE CLASIFICADOS**

# Ejemplo Reconocimiento de Números en Imágenes

Supongamos que queremos identificar números dentro de una imagen



OBTENER UNA GRAN CANTIDAD DE MUESTRAS DE NÚMEROS PREVIAMENTE CLASIFICADOS

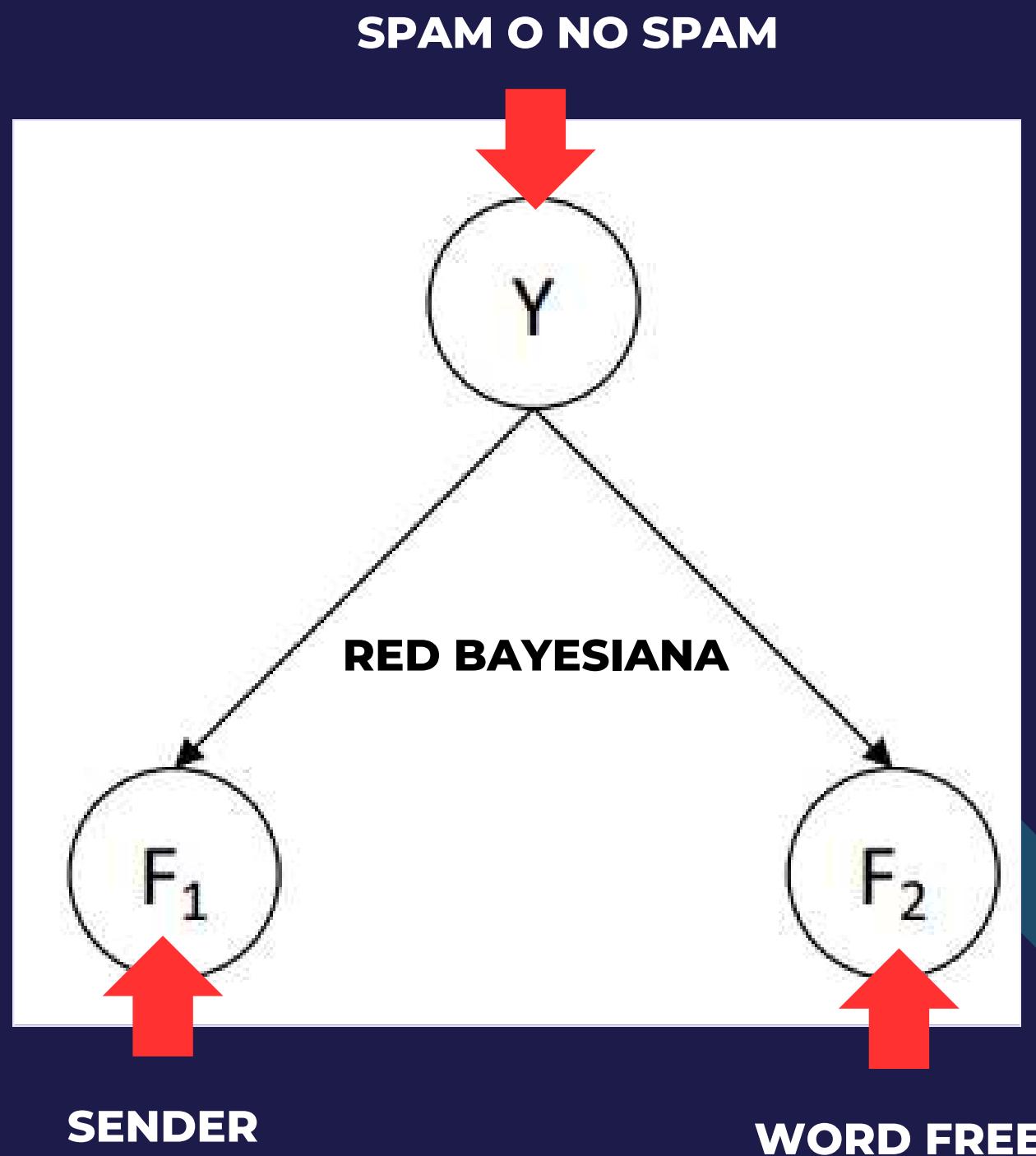
# Aplicaciones de Clasificación

- ✓ Detección spam
- ✓ OCR
- ✓ Diagnósticos médicos
- ✓ Detección de fraudes
- ✓ Ruteo de quejas
- ✓ Reconocimiento de Voz
- ✓ Segmentación de Clientes en Marketing
- ✓ Clasificación de Noticias
- ✓ etc...

# Algortimo Naïve Bayes



# Naïve Bayes



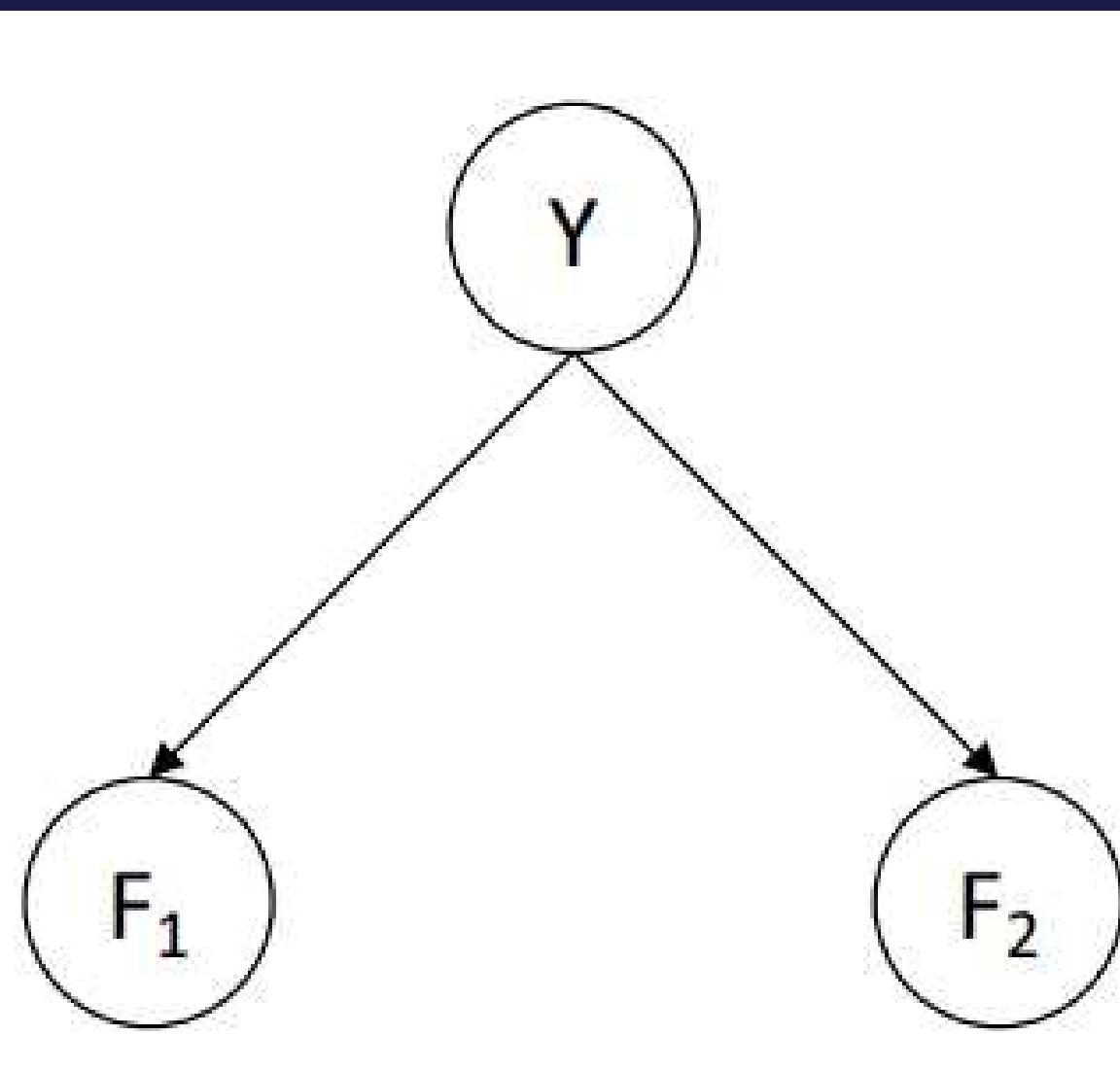
Para el correo electrónico sabemos que hay dos clasificaciones esperadas: **SPAM y NO SPAM**

También para este ejemplo en los datos sabremos dos características:

- Quién envía el correo ( $F_1$ )
- Número de veces que aparece la palabra *FREE* ( $F_2$ )

Consideramos ambas *features* independientes entre ellas

# Naïve Bayes



Y: The label (spam or ham)	
Y	P(Y)
ham	?
spam	?

F <sub>1</sub> : A feature (do I know the sender?)		
F <sub>1</sub>	Y	P(F <sub>1</sub>  Y)
yes	ham	?
no	ham	?
yes	spam	?
no	spam	?

F <sub>2</sub> : Another feature (# of occurrences of FREE)		
F <sub>2</sub>	Y	P(F <sub>2</sub>  Y)
0	ham	?
1	ham	?
2	ham	?
0	spam	?
1	spam	?
2	spam	?

# Naïve Bayes

## DATOS DE ENTRENAMIENTO

Training Data		
#	Email Text	Label
1	Attached is my portfolio.	ham
2	Are you <b>free</b> for a meeting tomorrow?	ham
3	<b>Free</b> unlimited credit cards!!!!	spam
4	Mail \$10,000 check to this address	spam
5	Sign up now for 1 <b>free</b> Bitcoin	spam
6	<b>Free</b> money <b>free</b> money	spam

Total de correos: **6**

Veces que aparece ham: **2**

$$P(\text{ham}) = 2/6 = 1/3 = 0.3333$$

Veces que aparece *FREE* en un correo etiquetado como ham:

- 0 veces: **1**  $P(\text{f2}|\text{ham}) = 1/2 = 0.5$
- 1 veces: **1**  $P(\text{f2}|\text{ham}) = 1/2 = 0.5$
- 2 veces: **0**  $P(\text{f2}|\text{ham}) = 0$

# Naïve Bayes

## DATOS DE ENTRENAMIENTO

Training Data		
#	Email Text	Label
1	Attached is my portfolio.	ham
2	Are you <b>free</b> for a meeting tomorrow?	ham
3	<b>Free</b> unlimited credit cards!!!!	spam
4	Mail \$10,000 check to this address	spam
5	Sign up now for 1 <b>free</b> Bitcoin	spam
6	<b>Free</b> money <b>free</b> money	spam

Total de correos: **6**

Veces que aparece spam: 4

$$P(\text{spam}) = 4/6 = 2/3 = 0.6666$$

Veces que aparece *FREE* en un correo etiquetado como spam:

- 0 veces: **1**  $P(f2|\text{spam}) = 1/4 = 0.25$
- 1 veces: **2**  $P(f2|\text{spam}) = 2/4 = 0.5$
- 2 veces: **1**  $P(f2|\text{spam}) = 1/4 = 0.25$

# Naïve Bayes

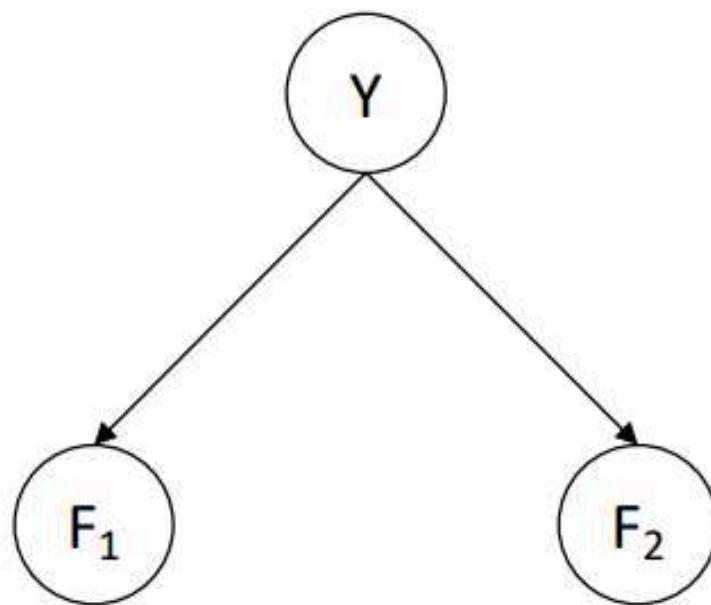
$F_2$ : # of occurrences of FREE

$F_2$	Y	$P(F_2 Y)$
0	ham	0.5
1	ham	0.5
2	ham	0.0
0	spam	0.25
1	spam	0.50
2	spam	0.25

Training Data

#	Email Text	Label
1	Attached is my portfolio.	ham
2	Are you <b>free</b> for a meeting tomorrow?	ham
3	<b>Free</b> unlimited credit cards!!!!	spam
4	Mail \$10,000 check to this address	spam
5	Sign up now for 1 <b>free</b> Bitcoin	spam
6	<b>Free</b> money <b>free</b> money	spam

# Naïve Bayes



Y: The label (spam or ham)	
Y	P(Y)
ham	0.6
spam	0.4

F1: A feature (do I know the sender?)		
F1	Y	P(F1 Y)
yes	ham	0.7
no	ham	0.3
yes	spam	0.1
no	spam	0.9

F2: Another feature (# of occurrences of FREE)		
F2	Y	P(F2 Y)
0	ham	0.85
1	ham	0.07
2	ham	0.08
0	spam	0.75
1	spam	0.12
2	spam	0.13

# Naïve Bayes

Suponga el mensaje conociendo la persona que envió (contacto):

**“Free food in Soda 430 today”**

Identifique si es spam o no spam

Sabemos que:

F1 = yes

F2 = 1

## Probabilidad Conjunta

Que el correo sea Spam, sabiendo quien lo envio y que tiene la palabra Free una vez en el mensaje

$$P(Y = \text{spam}, F1 = \text{yes}, F2 = 1) = P(Y = \text{spam}) * P(F1 = \text{yes} | Y = \text{spam})$$

$$* P(F2 = 1 | Y = \text{spam}) = 0.4 * 0.1 * 0.12 = \mathbf{0.0048}$$

Que el correo sea No Spam, sabiendo quien lo envio y que tiene la palabra Free una vez en el mensaje

$$P(Y = \text{ham}, F1 = \text{yes}, F2 = 1) = P(Y = \text{ham}) * P(F1 = \text{yes} | Y = \text{ham}) *$$

$$P(F2 = 1 | Y = \text{ham}) = 0.6 * 0.7 * 0.07 = \mathbf{0.0294}$$

# Naïve Bayes

Suponga el mensaje conociendo la persona que envió (contacto):

**“Free food in Soda 430 today”**

Identifique si es spam o no spam

Sabemos que:

F1 = yes

F2 = 1

## Normalizar

$$P(Y = \text{spam} | F1 = \text{yes}, F2 = 1) = 0.0048 / (0.0048 + 0.0294) = \mathbf{0.14}$$

$$P(Y = \text{ham} | F1 = \text{yes}, F2 = 1) = 0.0294 / (0.0048 + 0.0294) = \mathbf{0.86}$$

14% de posibilidades de que el correo electrónico sea spam.

**86% de posibilidades de que sea ham.**

# Maximum likelihood en Naïve Bayes



# Maximum Likelihood

El algoritmo de Naive Bayes utiliza el principio de Máxima Verosimilitud (Maximum Likelihood) para estimar las probabilidades necesarias para realizar la clasificación.

El objetivo es encontrar la clase **C** que maximice la probabilidad posterior dado un conjunto de características **X=(x<sub>1</sub>,x<sub>2</sub>,...,x<sub>n</sub>)**:

$$P(C | X) = \frac{P(X | C) \cdot P(C)}{P(X)}$$

Dado que P(X) es constante para todas las clases, basta con maximizar el numerador (o sea no necesitamos normalizar para identificar la clasificación, basta con encontrar el mayor valor en el numerador):

$$P(C | X) \propto P(X | C) \cdot P(C)$$

# Over-fitting



# Over-fitting

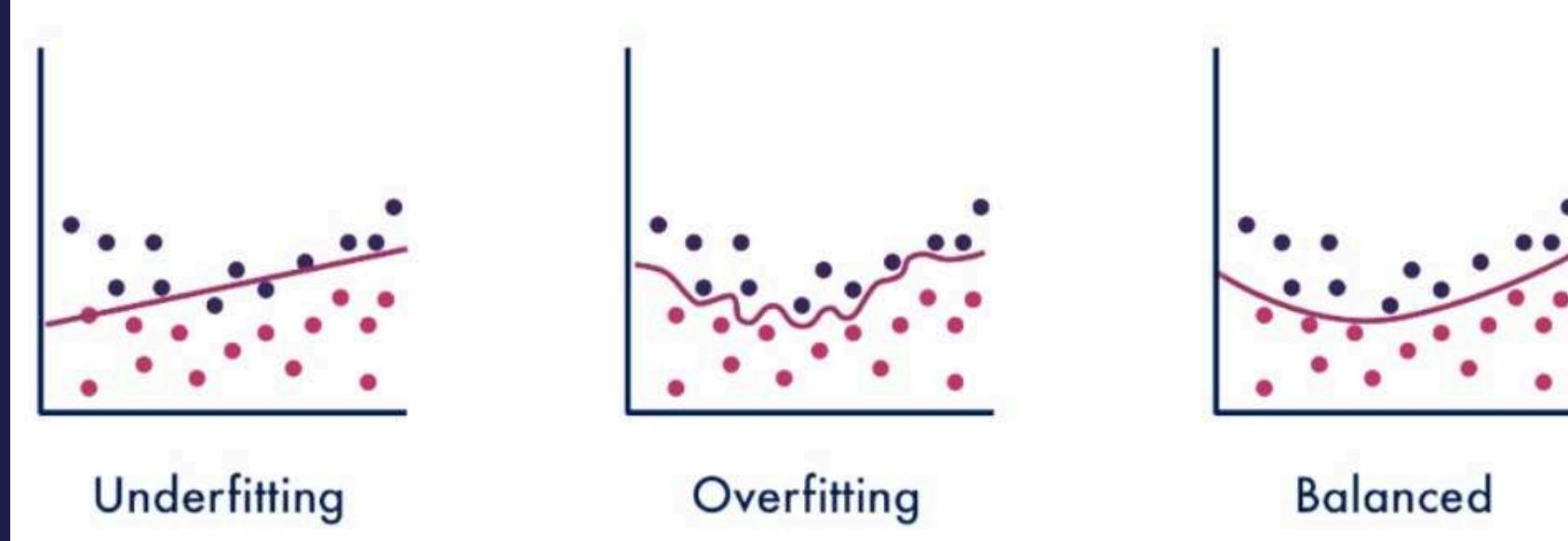
Ocurre cuando un modelo de machine learning se ajusta demasiado bien a los datos de entrenamiento, capturando tanto los patrones reales como el ruido o las anomalías presentes en el conjunto de datos.

Como resultado, el modelo funciona muy bien con los datos de entrenamiento pero falla al generalizar con nuevos datos

## ¿Por qué ocurre el overfitting?

- Modelo demasiado complejo
- Demasiadas características
- Pocos datos de entrenamiento
- Entrenamiento excesivo

# ¿Cómo detectar el overfitting?



- **Diferencia de rendimiento entre entrenamiento y prueba:**
  - Alta exactitud (accuracy) en entrenamiento (por ejemplo, 98%).
  - Baja exactitud (accuracy) en prueba (por ejemplo, 60%).
- **Curvas de aprendizaje:**
  - La precisión de entrenamiento sigue aumentando, pero la precisión de validación se estanca o disminuye.

# ¿Cómo evitar el overfitting?

Aunque el algoritmo de Naive Bayes generalmente es menos propenso al overfitting en comparación con modelos más complejos como redes neuronales, aún puede ocurrir en ciertos escenarios, para esto se puede aplicar las siguientes técnicas:

- **Suavizado de Laplace (Laplace Smoothing)**
- **Selección de características:** remover features irrelevantes o ruidosas que hagan aprender patrones falsos
- **Eliminación de características redundantes o correlacionadas** (features realmente dependientes)
- **Limpieza y preprocessamiento de datos:** Eliminar valores atípicos (outliers) y normalizar los datos puede reducir el riesgo.

# ¿Cómo evitar el overfitting?

- **Validación cruzada (Cross-Validation):** Realizar una validación cruzada, como **K-Fold\***, ayuda a evaluar el modelo en diferentes subconjuntos de datos para detectar el overfitting de manera temprana.
- **Aumento de datos (Data Augmentation):** Si tienes pocos datos, el modelo puede memorizar ejemplos específicos. Generar ejemplos adicionales, especialmente en tareas de clasificación de imágenes o texto, puede ayudar a mejorar la generalización.
- **Priorización de clases balanceadas:** Si el conjunto de datos está muy desbalanceado, el modelo podría sobreajustarse a la clase mayoritaria.

# K-Fold Cross-Validation

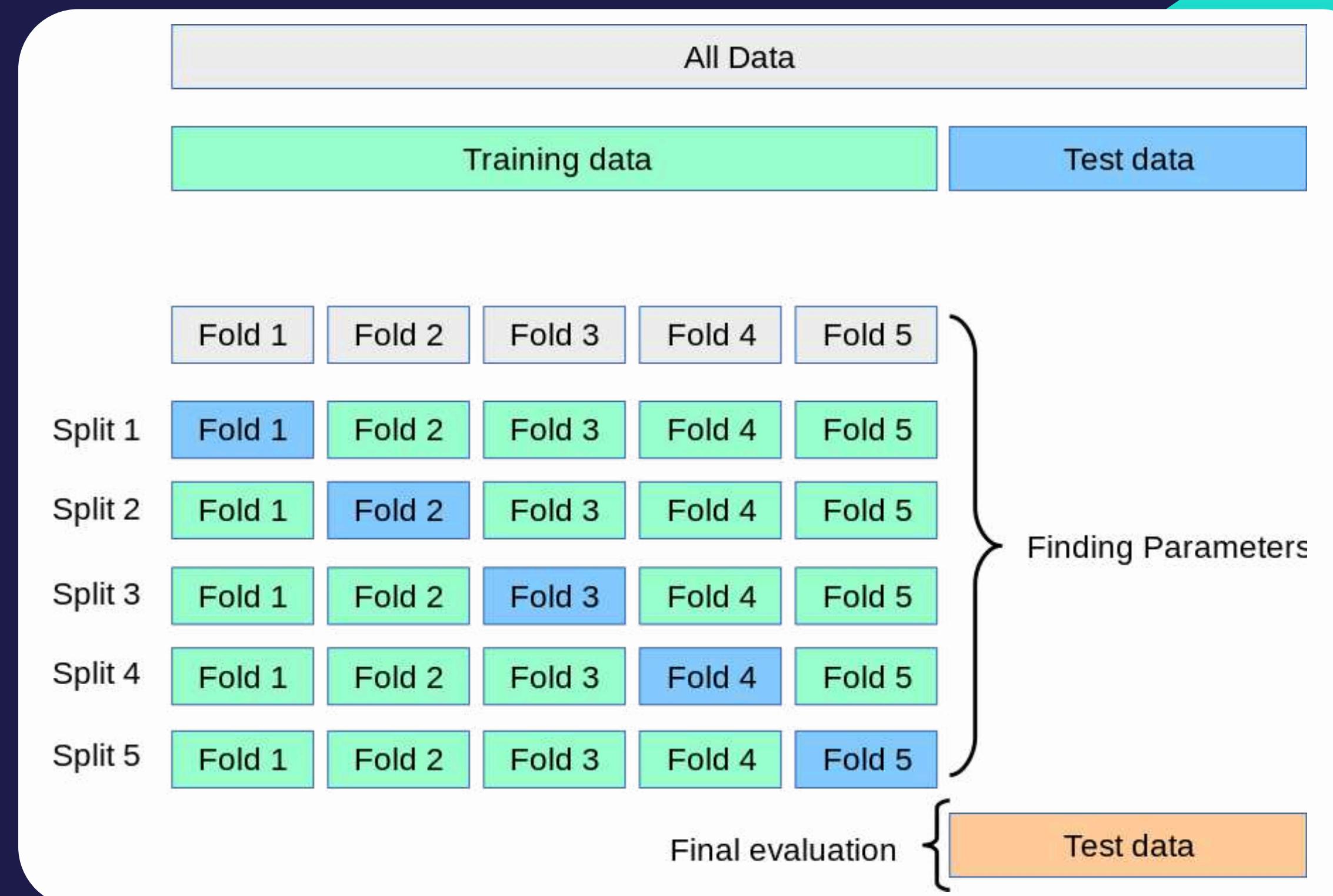
# K-Fold Cross-Validation

Es una técnica utilizada para evaluar el rendimiento de un modelo de machine learning y evitar problemas como el overfitting. Su objetivo principal es asegurarse de que el modelo generalice bien a datos nuevos

## Pasos

1. **Dividir los datos en K grupos (o "folds") de tamaño aproximadamente igual.**
  - a. Por ejemplo, si se tienen 1000 datos y se elige K=5, cada grupo tendrá 200 datos
2. **Entrenamiento y prueba repetidos:**
  - a. Se realizan K iteraciones en total.
  - b. En cada iteración:
    - i. (K-1) folds se utilizan para entrenar el modelo.
    - ii. 1 fold se utiliza para probar el modelo.
3. **Promedio de resultados:**
  - a. Después de entrenar y evaluar el modelo K veces, se promedian las métricas (como precisión, recall, F1-score, etc.) para obtener una estimación más robusta del rendimiento.

# K-Fold Cross-Validation



# Laplace Smoothing



# Laplace Smoothing

Es una técnica que se utiliza en el algoritmo Naive Bayes para evitar el problema de las probabilidades nulas

## ¿Cuál es el problema?

Sabemos que: 
$$P(\text{clase} \mid \text{característica}) = \frac{P(\text{característica} \mid \text{clase}) \cdot P(\text{clase})}{P(\text{característica})}$$

El mayor problema surge al calcular  $P(\text{característica} \mid \text{clase})$ , ya que si alguna característica no aparece en una clase, el resultado será cero.

Si la **frecuencia** es cero, la probabilidad será cero, y eso hará que toda la probabilidad de la clase también sea cero.

# Laplace Smoothing

El suavizado de Laplace añade un valor constante (generalmente 1) a cada recuento para asegurarse de que ninguna probabilidad sea cero.

$$P(\text{característica} \mid \text{clase}) = \frac{\text{Frecuencia}(\text{característica} \mid \text{clase}) + 1}{\text{Total de características en la clase} + V}$$

- $V$  es el número total de características posibles.
- El valor "1" garantiza que incluso las características no observadas tengan una pequeña probabilidad.

# Ejemplo Bag of Words usando Laplace Smoothing

Supongamos que tenemos un conjunto de datos de correos electrónicos clasificados como spam y no spam.

Email	Spam (1) / No Spam (0)	Palabras en el email
"oferta exclusiva"	1	oferta, exclusiva
"gana dinero fácil"	1	gana, dinero, facil
"saludos cordiales"	0	saludos, cordiales
"reunión de trabajo"	0	reunion, de, trabajo

# Ejemplo Bag of Words usando Laplace Smoothing

Contemos las frecuencias de las palabras en cada clase (spam y no spam)

## Spam:

- "oferta": 1
- "exclusiva": 1
- "gana": 1
- "dinero": 1
- "fácil": 1

## No Spam:

- "saludos": 1
- "cordiales": 1
- "reunión": 1
- "trabajo": 1

Total de palabras en Spam: 5

Total de palabras en No Spam: 4

# Ejemplo Bag of Words usando Laplace Smoothing

Probabilidades sin suavizado

Palabra	Frecuencia	P(Palabra)
"oferta"	1	1/5
"exclusiva"	1	1/5
"gana"	1	1/5
"dinero"	1	1/5
"fácil"	1	1/5

# Ejemplo Bag of Words usando Laplace Smoothing

Probabilidades sin suavizado

Palabra	Frecuencia	P(Palabra)
"saludos"	1	1/4
"cordiales"	1	1/4
"reunión"	1	1/4
"trabajo"	1	1/4

Vocabulario Total:

$$V = \{"oferta", "exclusiva", "gana", "dinero", "fácil", "saludos", "cordiales", "reunión", "trabajo"\}$$

9 palabras unicas

# Ejemplo Bag of Words usando Laplace Smoothing

Aplicando Suavizado

$$P(\text{palabra} \mid \text{clase}) = \frac{\text{Frecuencia de la palabra} + 1}{\text{Total de palabras en la clase} + V}$$

**Spam:**

$$P(\text{"oferta"} \mid \text{spam}) = \frac{1 + 1}{5 + 9} = \frac{2}{14} \approx 0.1429$$

$$P(\text{"exclusiva"} \mid \text{spam}) = \frac{1 + 1}{5 + 9} = \frac{2}{14} \approx 0.1429$$

$$P(\text{"gana"} \mid \text{spam}) = \frac{1 + 1}{5 + 9} = \frac{2}{14} \approx 0.1429$$

**No Spam:**

$$P(\text{"saludos"} \mid \text{no spam}) = \frac{1 + 1}{4 + 9} = \frac{2}{13} \approx 0.1538$$

$$P(\text{"cordiales"} \mid \text{no spam}) = \frac{1 + 1}{4 + 9} = \frac{2}{13} \approx 0.1538$$

# Laplace Smoothing

Con nuestro proceso previamente calculado, tenemos la probabilidades ahora, supongamos que queremos clasificar un nuevo email: “**gana dinero**”

Queremos saber si es spam o no spam, usando Naive Bayes:

$$P(\text{Spam} \mid \text{gana, dinero}) = P(\text{Spam}) \cdot P(\text{gana} \mid \text{Spam}) \cdot P(\text{dinero} \mid \text{Spam})$$

$$P(\text{Spam}) = \frac{2}{4} = 0.5$$

$$P(\text{gana} \mid \text{Spam}) = 0.1429$$

$$P(\text{dinero} \mid \text{Spam}) = 0.1429$$

$$P(\text{Spam} \mid \text{gana, dinero}) = 0.5 \cdot 0.1429 \cdot 0.1429 \approx 0.0102$$

$$P(\text{No Spam}) = \frac{2}{4} = 0.5$$

$$P(\text{gana} \mid \text{No Spam}) = \frac{1}{13} \approx 0.0769$$

$$P(\text{dinero} \mid \text{No Spam}) = \frac{1}{13} \approx 0.0769$$

$$P(\text{No Spam} \mid \text{gana, dinero}) = 0.5 \cdot 0.0769 \cdot 0.0769 \approx 0.00295$$

Por Maximum Likelihood podemos concluir  $P(\text{Spam} \mid \text{gana, dinero}) > P(\text{No Spam} \mid \text{gana, dinero})$   
por tanto podemos concluir que el correo es SPAM

# Suma de Logaritmos



# Suma de Logaritmos

Los números muy pequeños pueden desaparecer (es decir, ser redondeados a cero) o pueden causar errores de precisión, lo que hace que los cálculos sean inexactos, el valor final puede ser tan pequeño que se pierde la información.

Una técnica comúnmente utilizada para evitar estos problemas es trabajar con logaritmos de las probabilidades en lugar de las probabilidades directamente.

## Normal

$$P(\text{clase} \mid \text{características}) = \frac{P(\text{característica}_1 \mid \text{clase}) \cdot P(\text{característica}_2 \mid \text{clase}) \cdot \dots}{P(\text{características})}$$

## Logaritmo

$$P(\text{clase} \mid \text{características}) = \log P(\text{característica}_1 \mid \text{clase}) + \log P(\text{característica}_2 \mid \text{clase}) + \dots$$

**Inteligencia Artificial**



# **Regresión Lineal y Logística**



# Regresión

- Introducción Regresión
- Regresión Lineal
  - Fundamentos Matemáticos
  - Evaluación de modelos de regresión
- Regresión Logística
  - Introducción a la Clasificación y Regresión Logística
  - Función Sísmoide
- Optimización
  - Gradiente descendente
  - Regularización



# Regresión

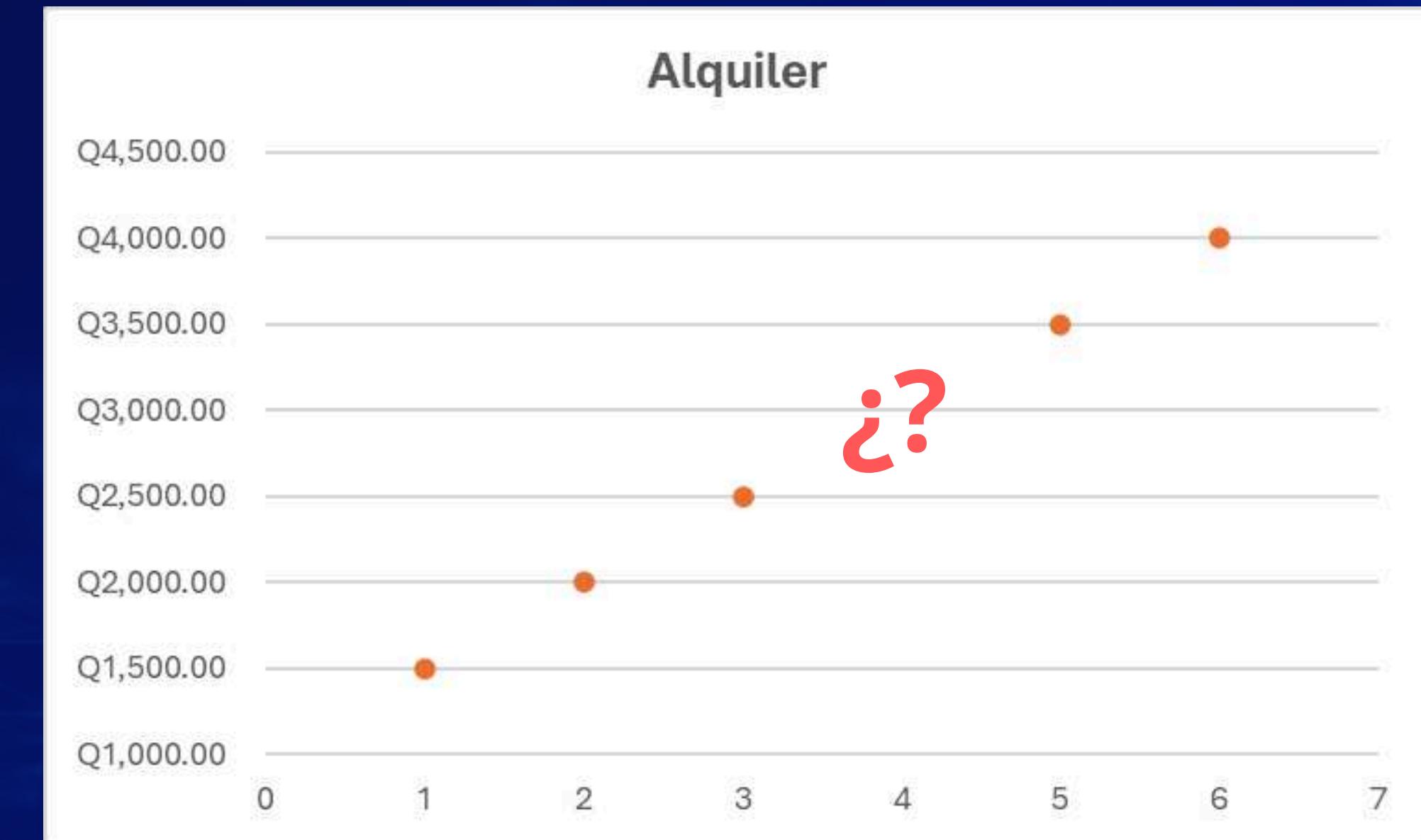
Predecir valores continuos con base en datos previos.





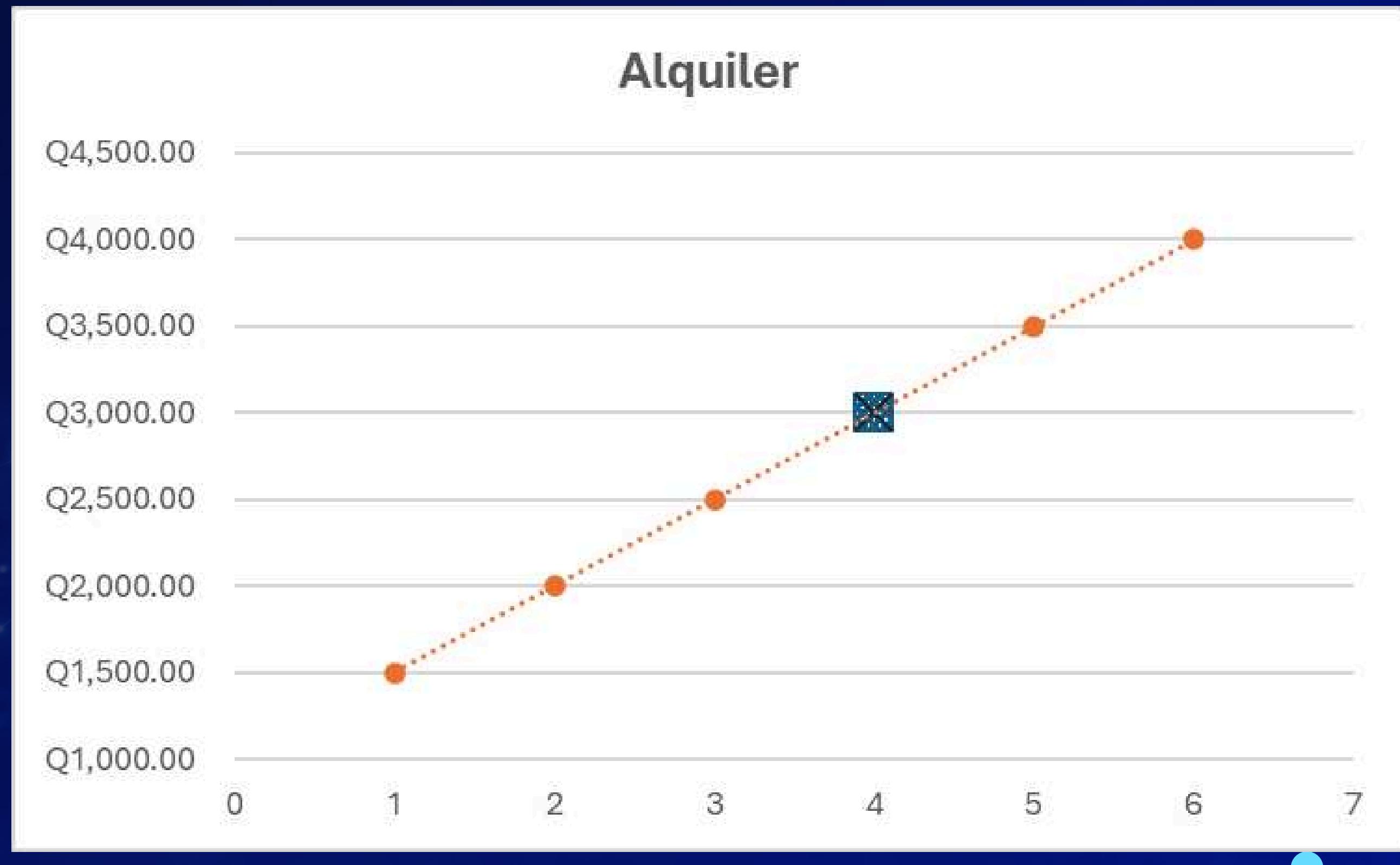
## Ejemplo: Valor de alquiler de casas en función del número de habitaciones

Habitaciones (x)	Alquiler (y)
1	1500 Q
2	2000 Q
3	2500 Q
4	?
5	3500 Q
6	4000 Q



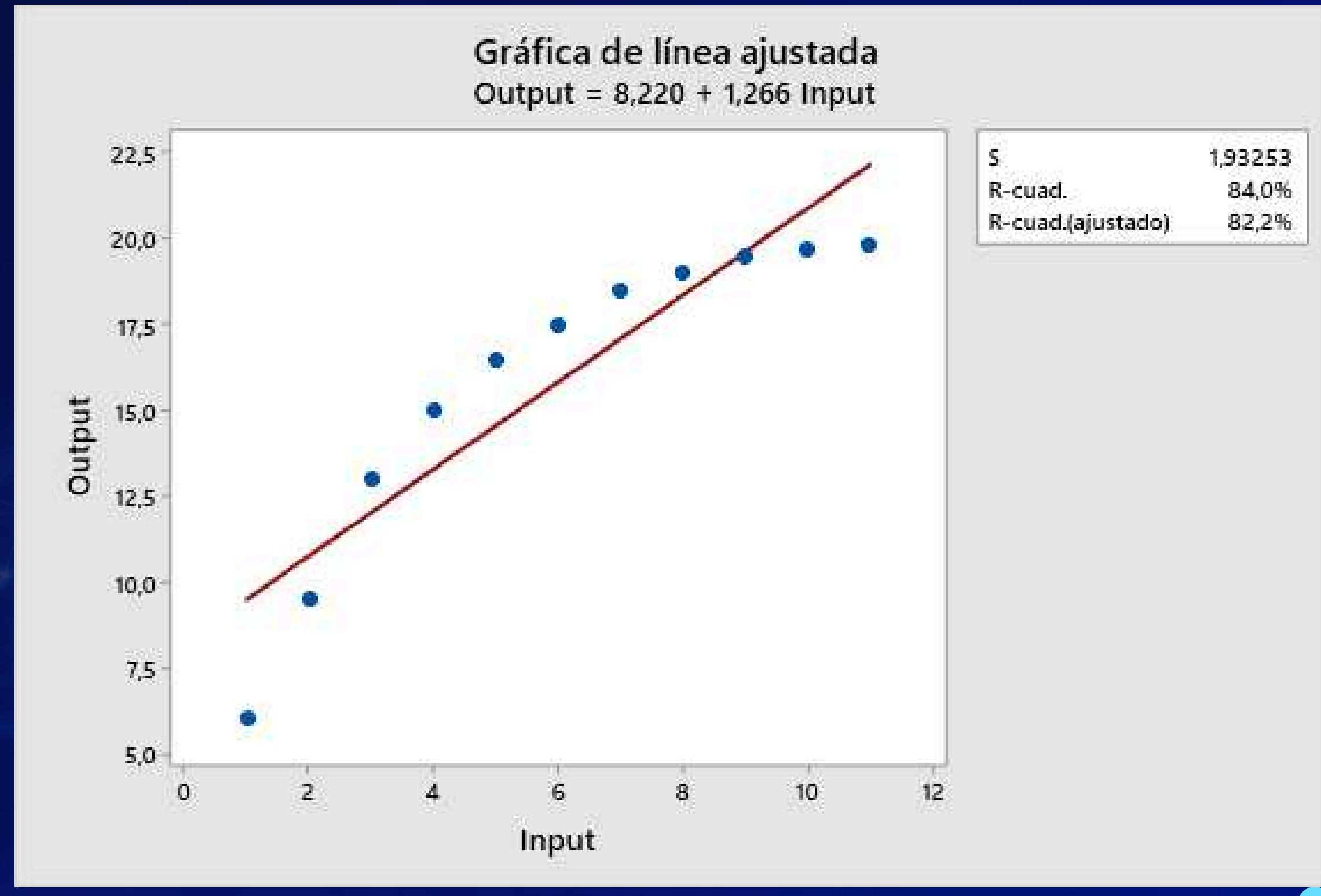


## Ejemplo: Valor de alquiler de casas en función del número de habitaciones



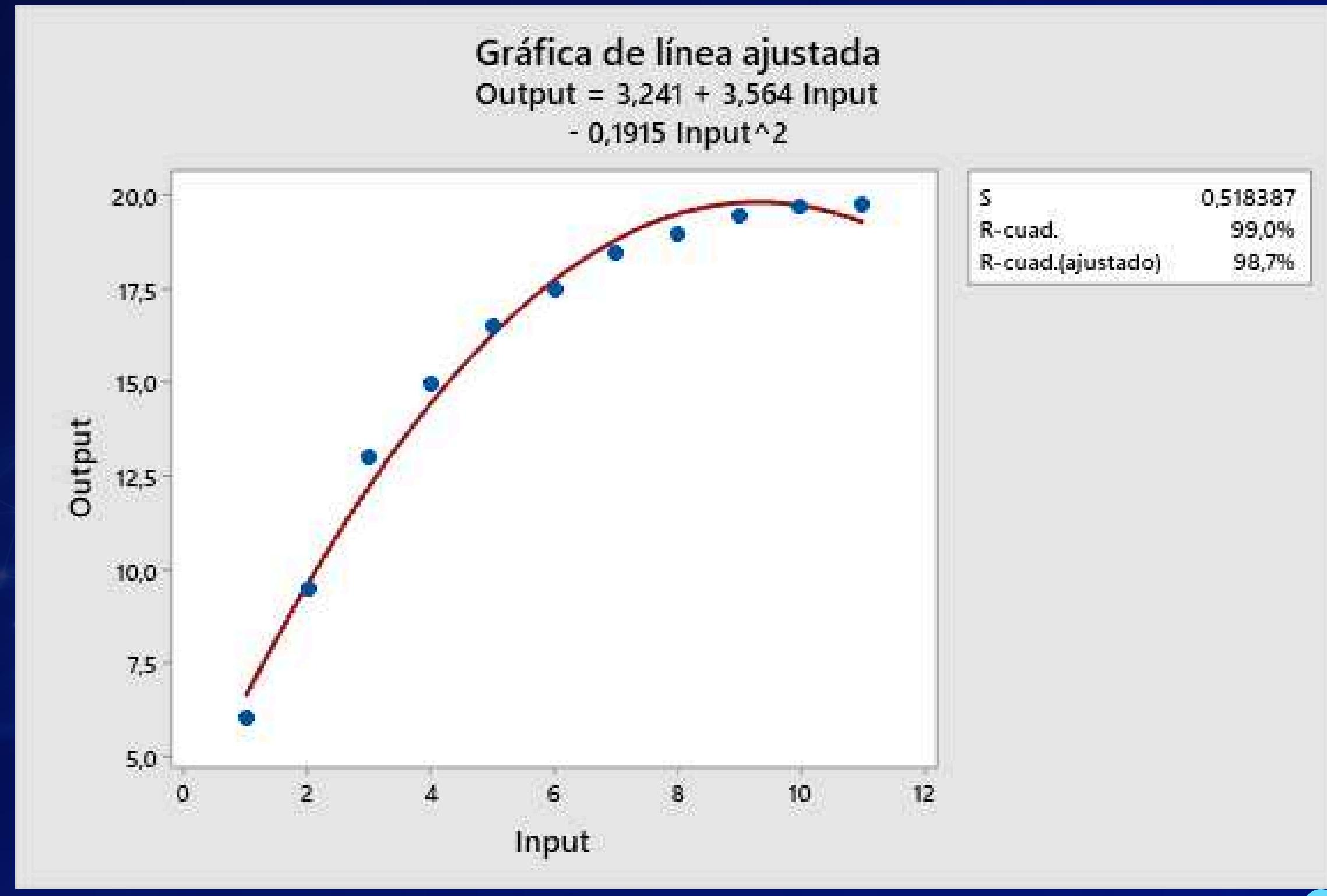


## Ejemplo





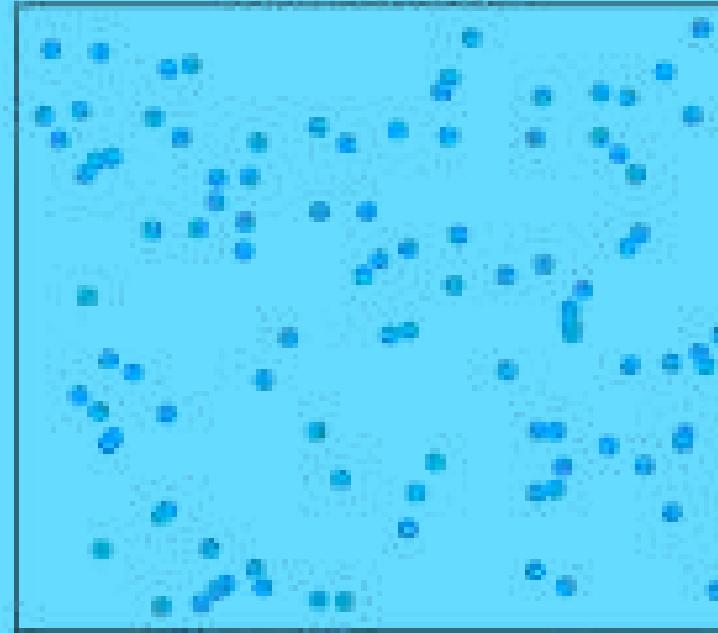
## Ejemplo



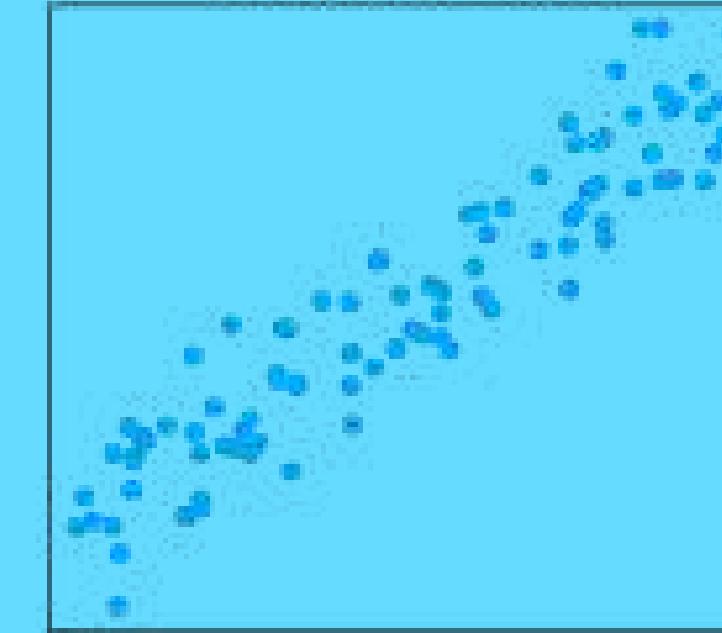


## Otros Modelos

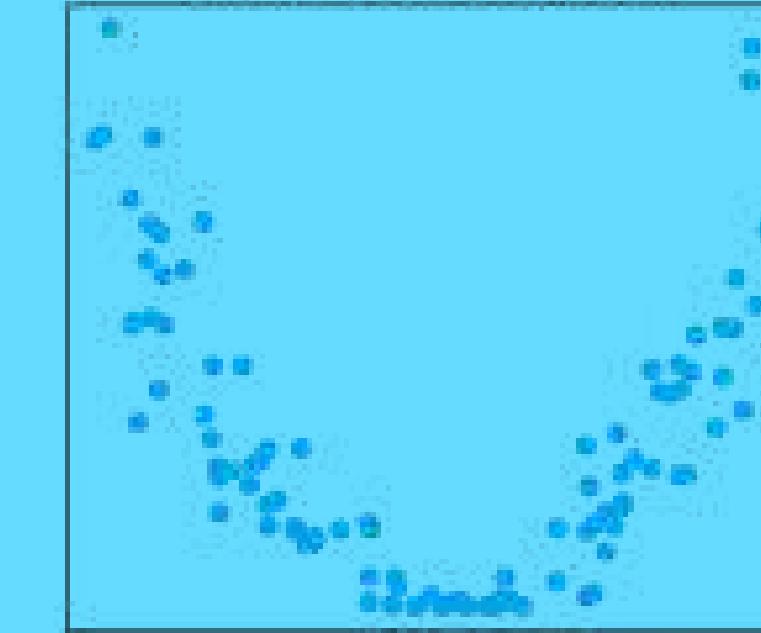
Sin relación



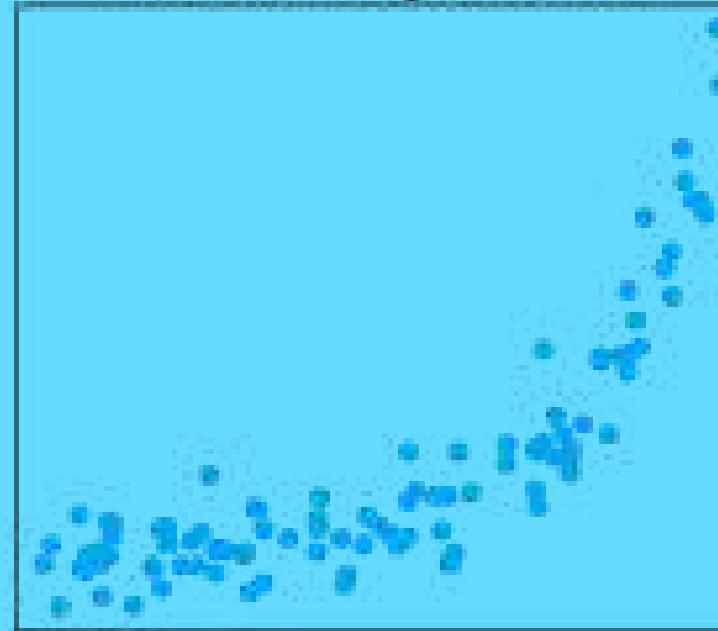
Relación lineal



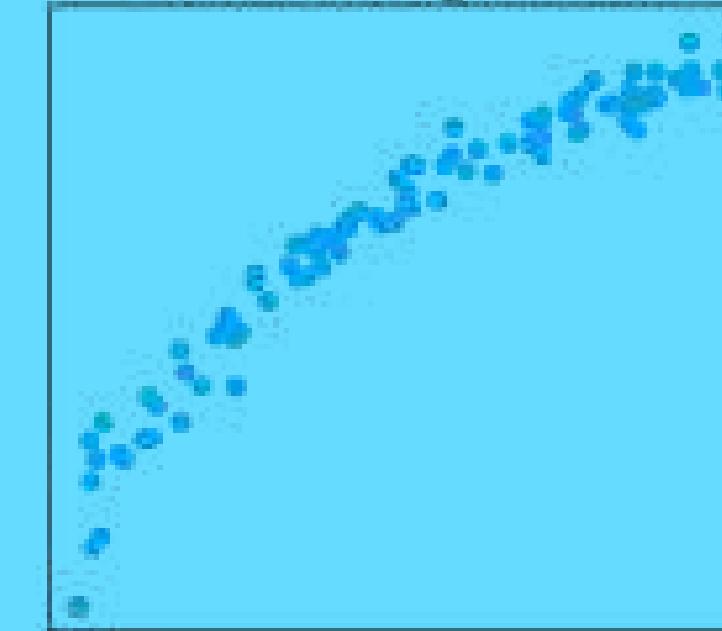
Relación cuadrática



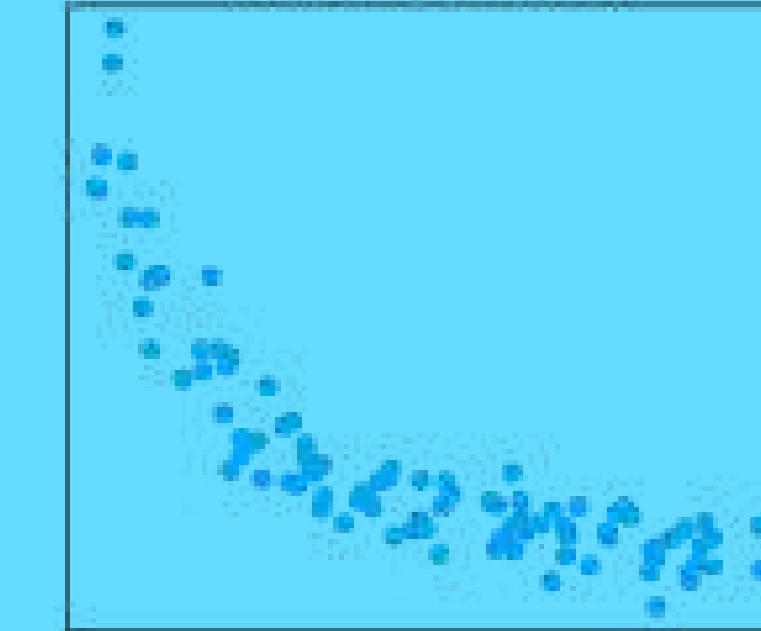
Relación exponencial



Relación logarítmica



Relación inversa





## Modelos Lineales vs Modelos No Lineales

Característica	Modelos Lineales	Modelos No Lineales
Relación entrada-salida	Lineal	No lineal
Fórmula general	$y = w_1x_1 + w_2x_2 + \dots + b$	$y = f(x)$ , donde $f$ es no lineal
Ejemplo simple	Regresión lineal	Regresión polinómica
Interpretabilidad	Fácil	Difícil
Computación	Rápida	Lenta
Flexibilidad	Baja	Alta
Tendencia al Overfitting	Menor	Mayor

# Diferencia entre Clasificación y Regresión

- **Clasificación:** La salida es booleana (0 o 1) o pertenece a una clase discreta.
  - *Ejemplo:* "¿Es spam o no?"
- **Regresión:** La salida es numérica y continua.
  - *Ejemplo:* "¿Cuál será el precio de una casa?"



## Regresión



# Problema de Regresión

En regresión, queremos encontrar una función  $f(x)$  que relacione los datos de entrada  $x$  con las salidas reales  $r$ .

- Si no hay ruido en los datos, es simplemente un problema de interpolación (ajustar la función exacta).
- Si hay ruido, se debe a variables ocultas que no observamos. Entonces, modelamos la relación como:

$$r_t = f(x_t) + \epsilon$$

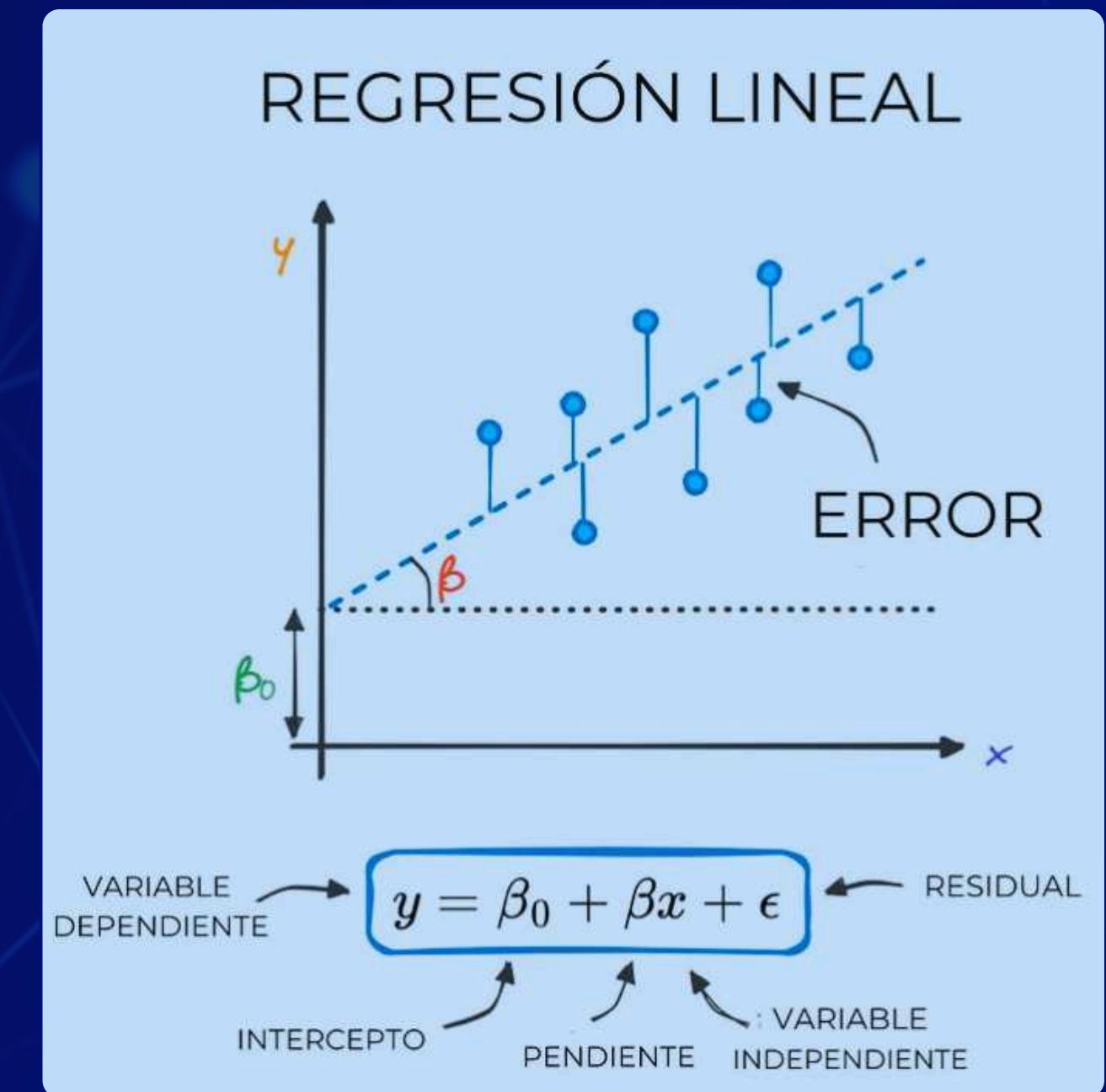
Donde  $\epsilon$  representa el ruido o factores externos que no podemos medir.

# Regresión Lineal Simple

Ecuación:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Parámetros ( $\beta_0$  y  $\beta_1$ ): Intercepto y pendiente.



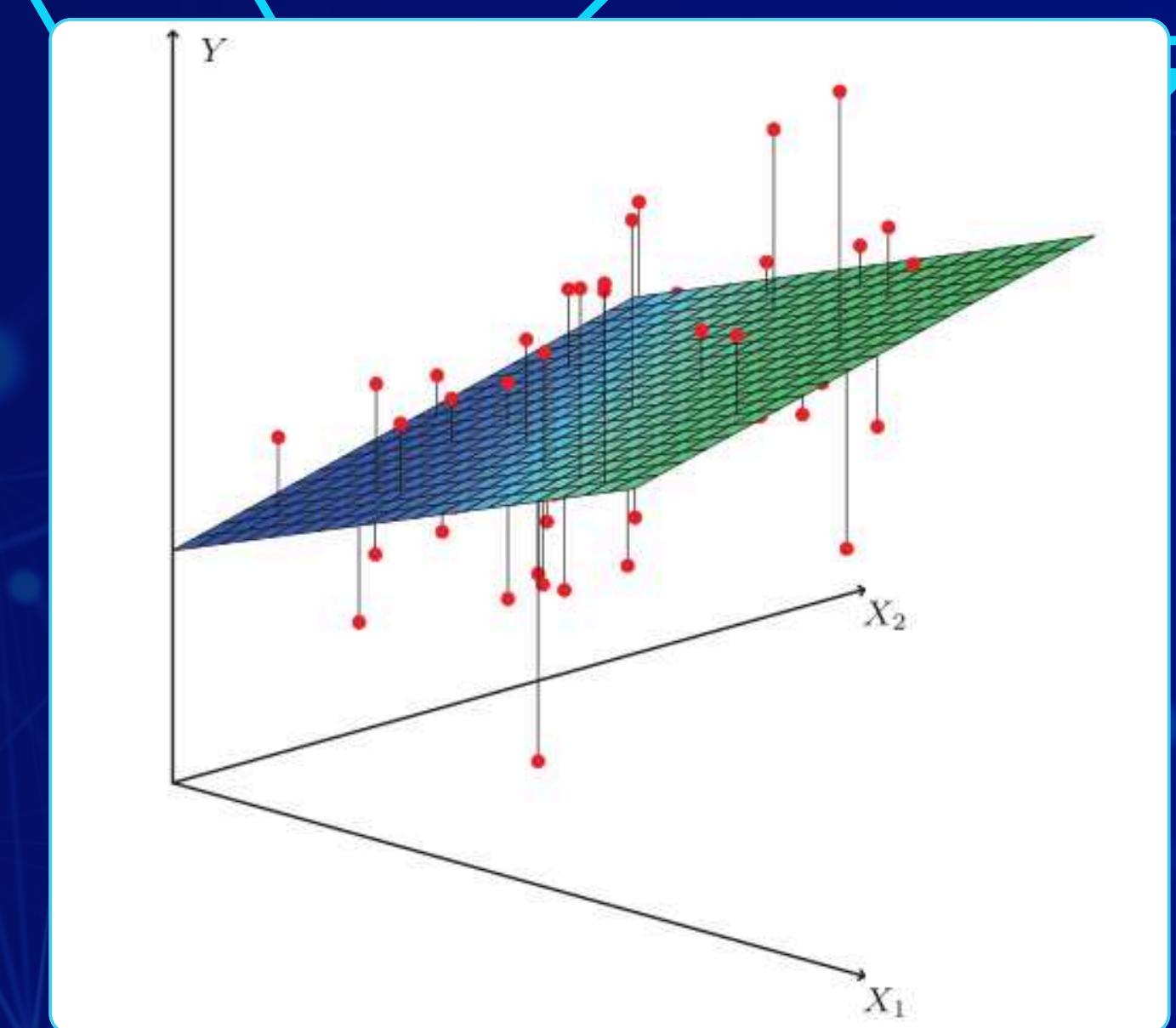
Objetivo: Encontrar  $\beta_0$  y  $\beta_1$  que minimicen el error

# Regresión Lineal Múltiple

Ecuación:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

- $y$ : Variable dependiente (objetivo).
- $\beta_0$ : Intercepto.
- $\beta_1, \beta_2, \dots, \beta_n$ : Coeficientes de las variables independientes  $x_1, x_2, \dots, x_n$ .
- $\epsilon$ : Término de error (residuo).



# Regresión Lineal Múltiple

Un modelo lineal se simplifica como:

$$y = X\beta + \epsilon$$

Donde  $y = (y_1, \dots, y_n)^T$  es un vector de outputs,  $X$  es una  $n \times k$  matriz,  $\beta = (\beta_1, \dots, \beta_n)^T$  es un vector de parámetros

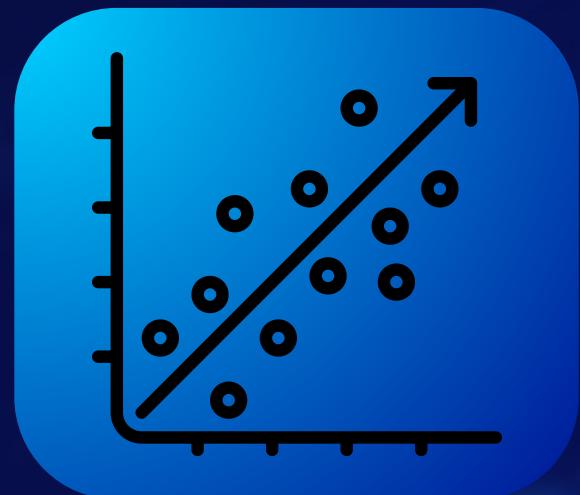
$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,k-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,k-1} \\ \vdots & \vdots & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,k-1} \end{pmatrix}$$

$$\beta = (\beta_0, \beta_1, \dots, \beta_{k-1})^T$$

$\epsilon$  es un vector de errores aleatorios.



# Supuestos de la Regresión Lineal Simple



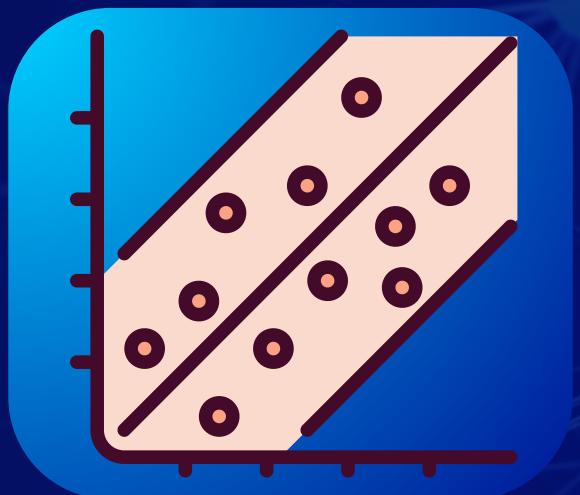
### Linealidad

Esto significa que debe existir una línea recta que pueda trazarse a través de los puntos de datos.



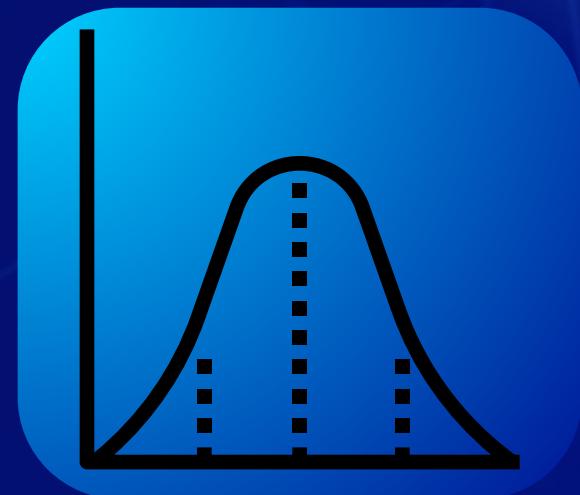
### Independencia

Cada observación en el conjunto de datos debe ser independiente de las demás. Esto significa que el valor de una observación no debe influir en el valor de otra.



### Homoscedasticidad

La variabilidad de los errores debe ser constante en todos los niveles de las variables independientes. Esto significa que la cantidad de la variable independiente no debe afectar cuánto varían los errores. Si la variabilidad de los errores no es constante, la regresión lineal no será precisa.



### Normalidad

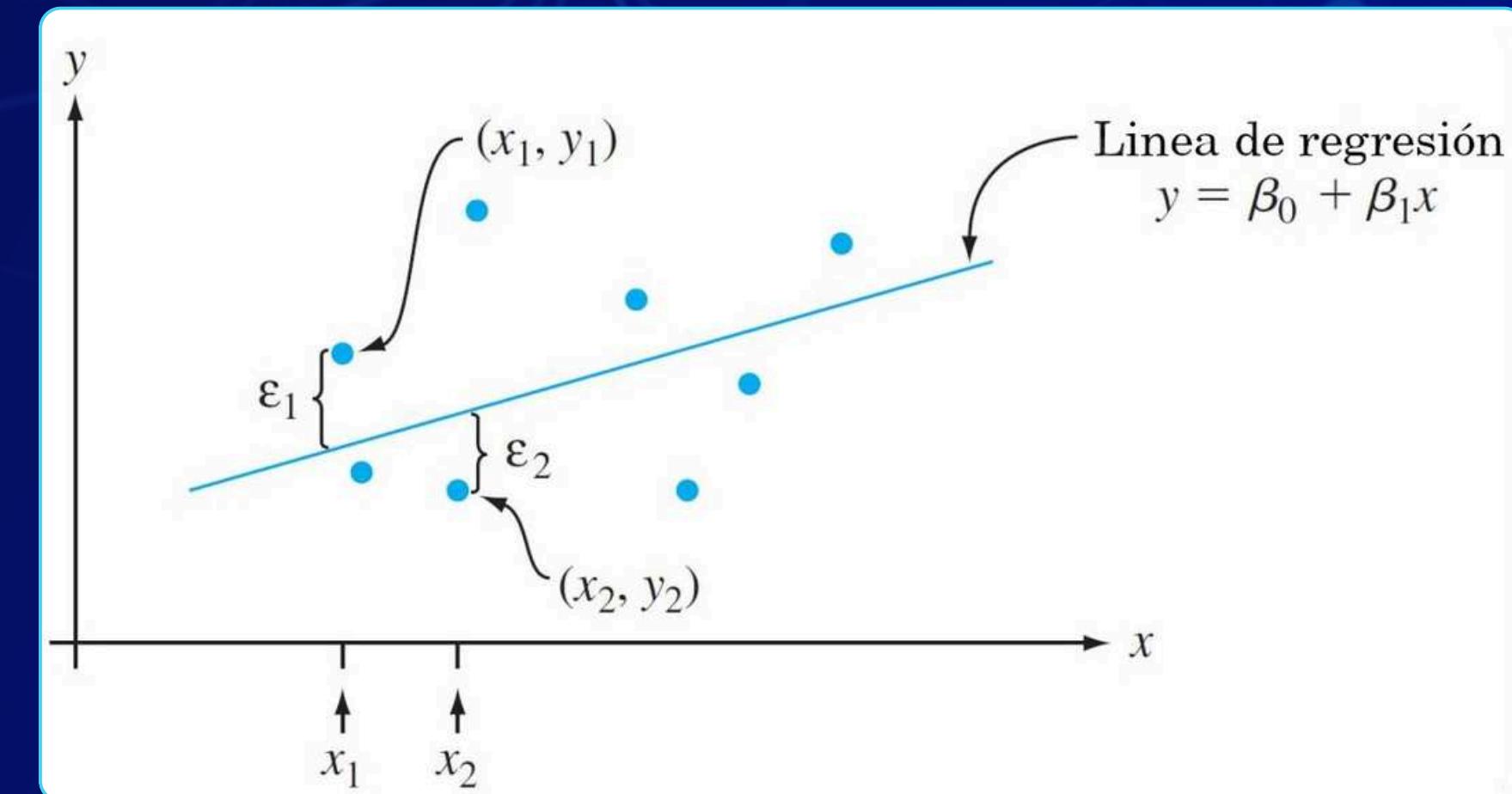
Los residuos (errores) deben seguir una distribución normal, que es una curva en forma de campana.

*Si no se dan estos supuestos, la regresión lineal no será precisa.*

# Función de Costo

La función de costo (cost function) es una función matemática que el modelo minimiza durante el entrenamiento.

Su propósito es medir el error entre las predicciones del modelo y los valores reales de los datos de entrenamiento.

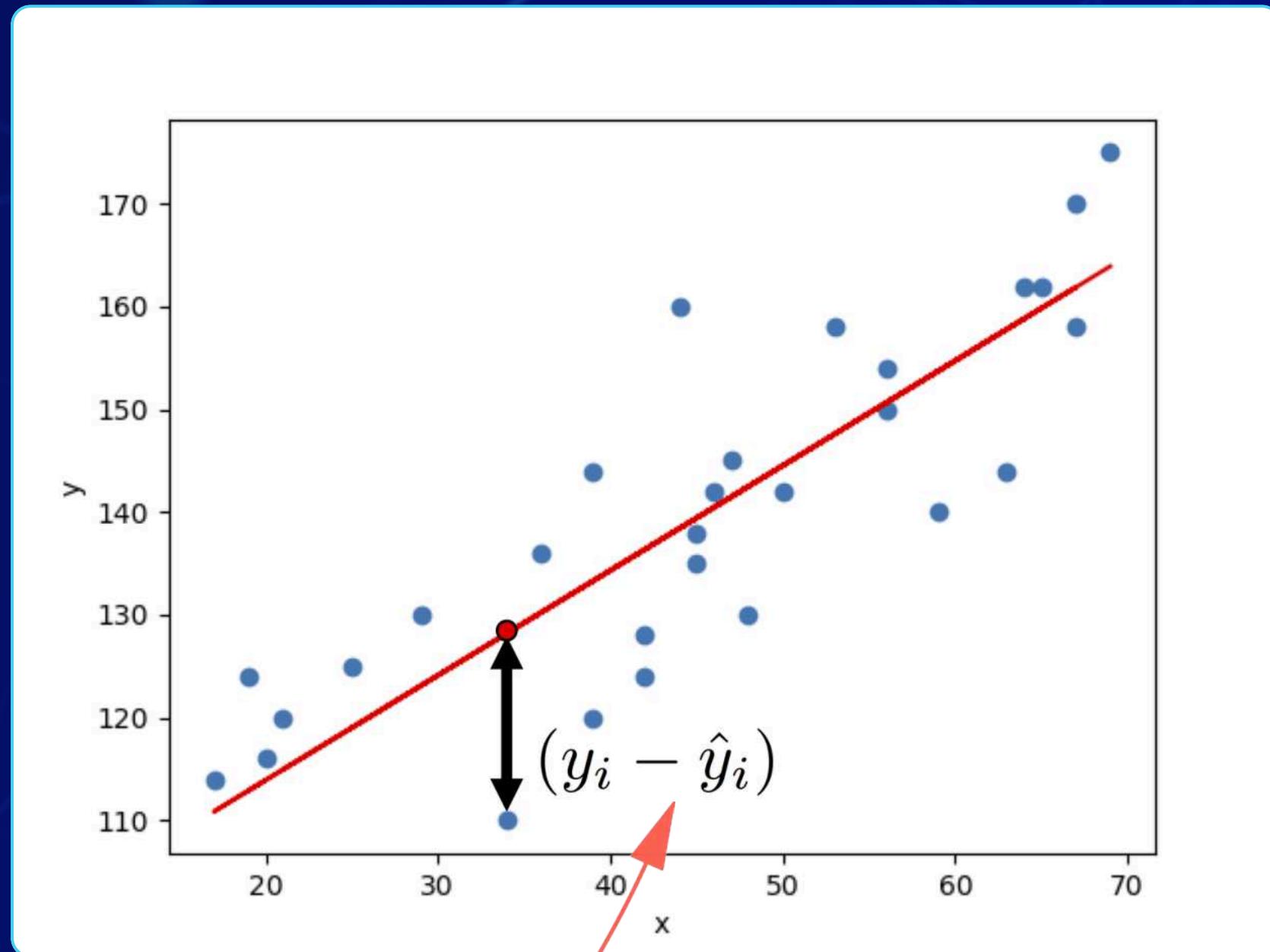


# Función de Costo

$$J(w) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{y}_i)$$

**Donde:**

- $J(w)$  es la función de costo.
- $N$  es el número total de ejemplos de entrenamiento.
- $L(y_i, \hat{y}_i)$  es la función de pérdida (cómo se mide el error para un solo dato).
- $y_i$  es el valor real.
- $\hat{y}_i$  es el valor predicho.



*¡Se vuelve un problema de Optimización!*

# Funciones de Pérdida

## Error Cuadrático (Squared Error Loss)

$$L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

- Penaliza los errores grandes de forma cuadrática.
- Base del Error Cuadrático Medio (MSE).

## Error Absoluto (Absolute Error Loss)

$$L(y_i, \hat{y}_i) = |y_i - \hat{y}_i|$$

- Penaliza errores de forma lineal.
- Más robusto a valores atípicos que el error cuadrático.

# Funciones de Pérdida

## Error Huber (Huber Loss)

$$L(y_i, \hat{y}_i) = \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2 & \text{for } |y_i - \hat{y}_i| \leq \delta, \\ \delta|y_i - \hat{y}_i| - \frac{1}{2}\delta^2 & \text{for } |y_i - \hat{y}_i| > \delta. \end{cases}$$

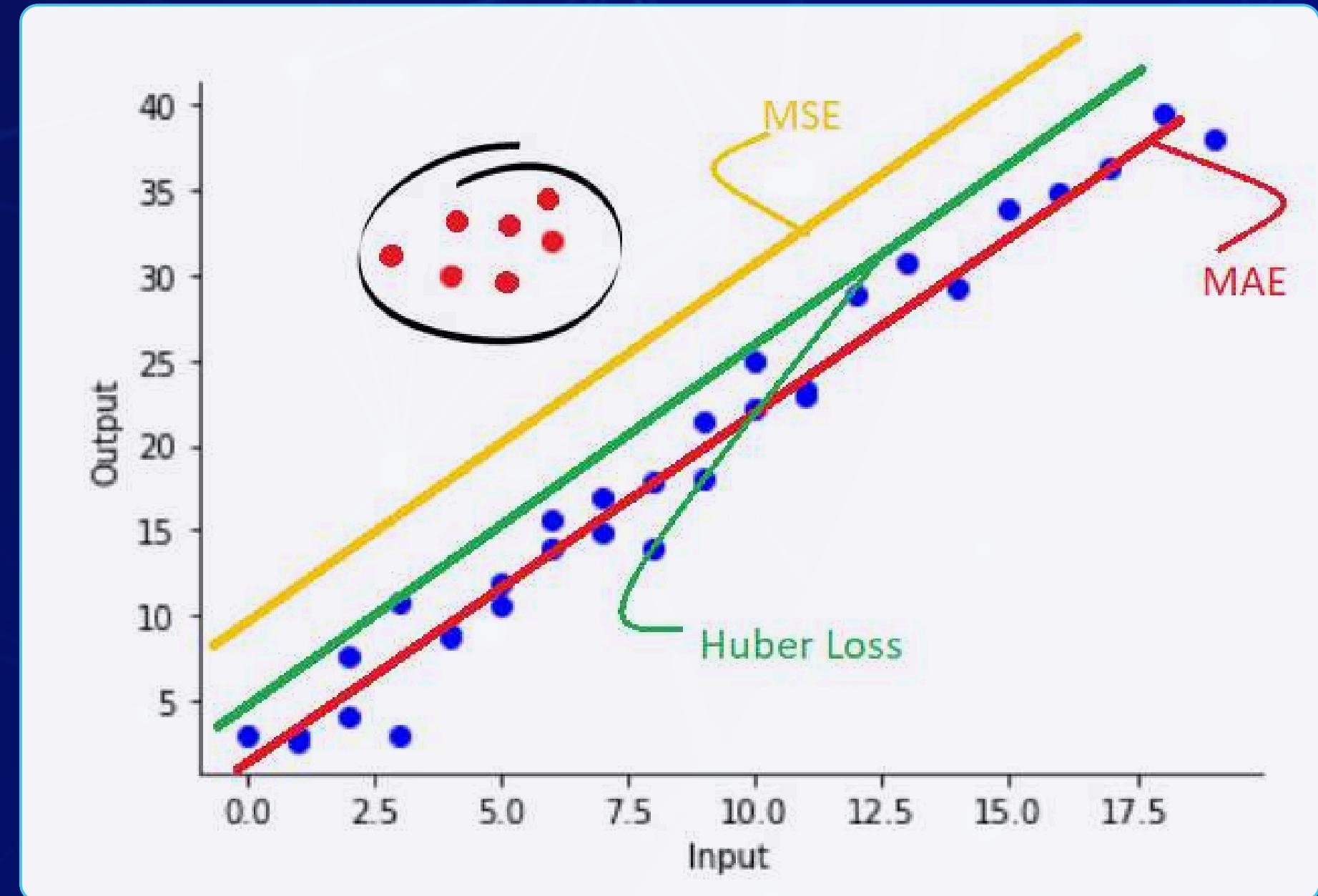
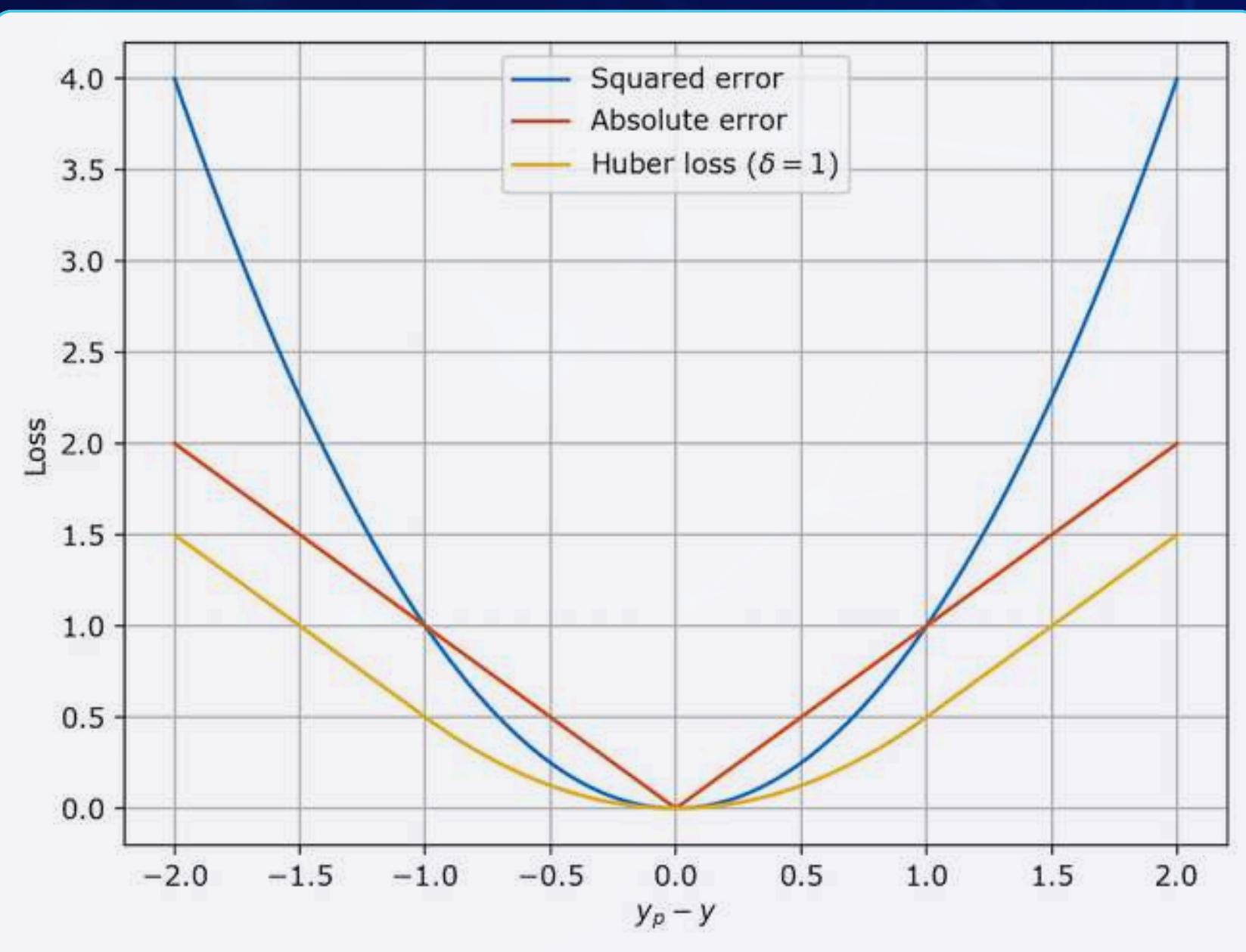
- Combina el MSE y MAE para ser menos sensible a valores atípicos.
- Para pequeños errores, actúa como MSE. Para errores grandes, como MAE.

## Error Log-Cosh (Log-Cosh Loss)

$$L(y_i, \hat{y}_i) = \log(\cosh(y_i - \hat{y}_i))$$

- Similar al Error Huber, pero más suave y diferenciable en todo su dominio.
- Se comporta como MSE para errores pequeños y MAE para errores grandes.

# Funciones de Pérdida



# Mínimos Cuadrados Ordinarios

Es un método de estimación utilizado para encontrar los parámetros de un modelo de regresión lineal.

La Suma de los Errores al Cuadrado es una función de costo que se utiliza para evaluar qué tan bien se ajusta el modelo a los datos. Cuanto menor sea la SSE, mejor será el ajuste del modelo.

$$J(w) = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Error suma de cuadrados (SSE)

# Mínimos Cuadrados Ordinarios

El objetivo es encontrar los valores de  $\beta_0$  y  $\beta_1$  que minimicen la suma de los errores al cuadrado (Mínimos Cuadrados Ordinarios, OLS):

$$S = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Reemplazamos  $y_i$  por la ecuación de la recta:

$$S = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

# Mínimos Cuadrados Ordinarios

La Suma de los Errores al Cuadrado es una función de costo que se utiliza para evaluar qué tan bien se ajusta el modelo a los datos. Cuanto menor sea la SSE, mejor será el ajuste del modelo.

$$S = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

Error suma de cuadrados (SSE)

Seguidamente debemos derivar la ecuación con respecto a  $\beta_0$  y  $\beta_1$

# Mínimos Cuadrados Ordinarios

Ahora tenemos las ecuaciones finales:

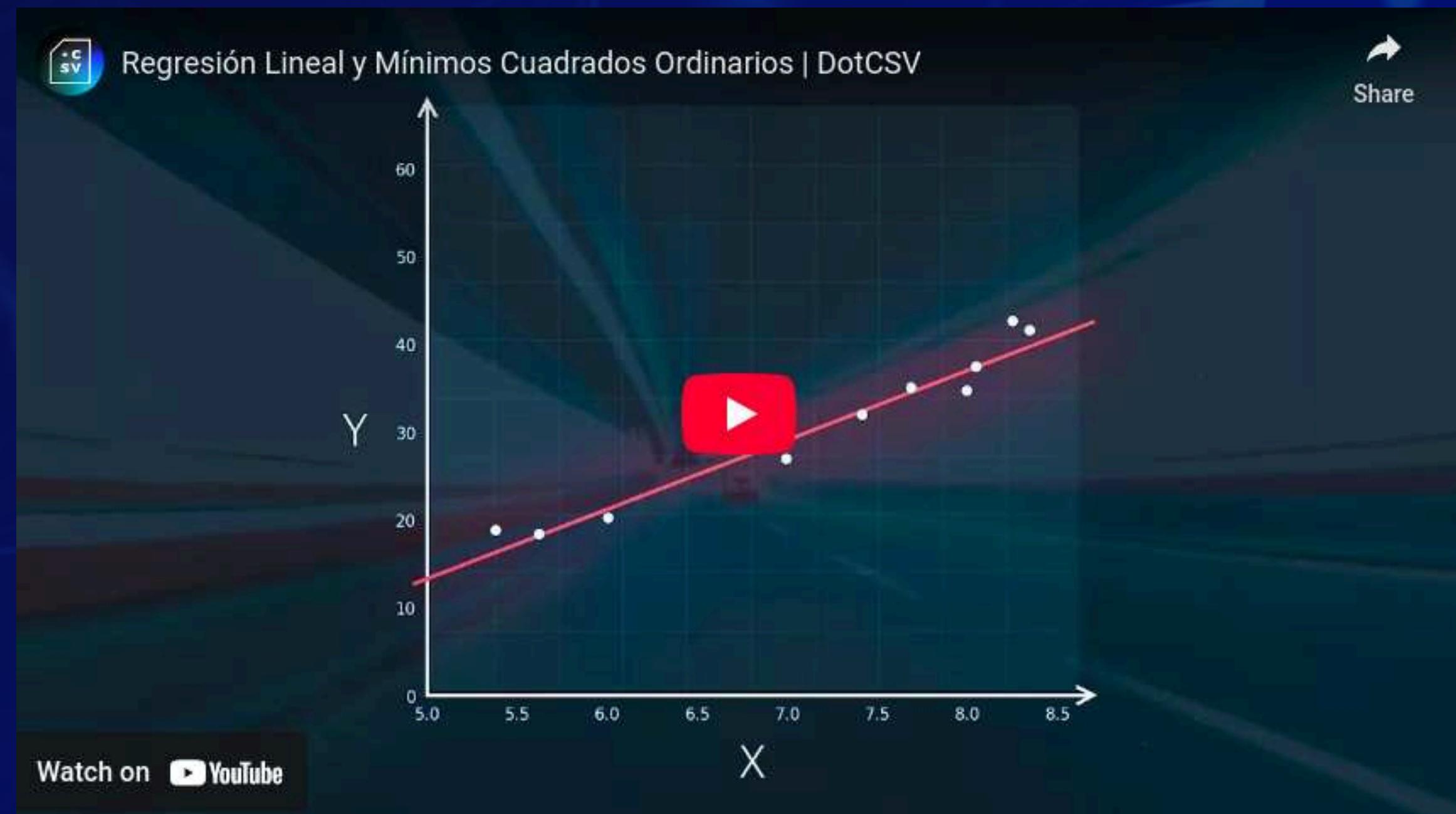
$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Donde  $\bar{x}$  y  $\bar{y}$  son los promedios de x e y, respectivamente.

- $\beta_1$  representa la pendiente de la recta: mide cómo cambia y por cada unidad de cambio en x.
- $\beta_0$  es el intercepto: el valor esperado de y cuando  $x=0$

# Resumen



# Ejemplo Regresión Lineal Simple

Habitaciones (x)	Alquiler (y)
1	1500 Q
2	2000 Q
3	2500 Q
4	3000 Q
5	3500 Q
6	4000 Q

## Paso 1: Relación entre las variables

Queremos encontrar una ecuación lineal que se ajuste a estos datos:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Donde:

- Y es el alquiler (valor dependiente).
- X es el número de habitaciones (valor independiente).
- $\beta_0$  es la intersección (cuando  $X=0$ ).
- $\beta_1$  es la pendiente (cambio en Y por cada unidad de X).
- $\epsilon$  representa el término de error.

# Ejemplo

**Paso 2:** Cálculo de promedios

Primero calculamos los promedios de X y Y:

$$\bar{X} = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

$$\bar{Y} = \frac{1500 + 2000 + 2500 + 3000 + 3500 + 4000}{6} = 2750$$

# Ejemplo

## Paso 3: Cálculo de la pendiente ( $\beta_1$ )

Para calcular la pendiente, usamos la fórmula:

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

a) Calculamos el numerador

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

b) Calculamos el denominador

$$\sum_{i=1}^n (x_i - \bar{x})^2$$

c) Calculamos  $\beta_1$

# Ejemplo

Paso 3a:

$X_i$	$Y_i$	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	1500	$1 - 3.5 = - 2.5$	$1500 - 2750 = - 1250$	$(- 2.5)(- 1250) = 3125$
2	2000	$2 - 3.5 = - 1.5$	$2000 - 2750 = - 750$	$(- 1.5)(- 750) = 1125$
3	2500	$3 - 3.5 = - 0.5$	$2500 - 2750 = - 250$	$(- 0.5)(- 250) = 125$
4	3000	$4 - 3.5 = 0.5$	$3000 - 2750 = 250$	$(0.5)(250) = 125$
5	3500	$5 - 3.5 = 1.5$	$3500 - 2750 = 750$	$(1.5)(750) = 1125$
6	4000	$6 - 3.5 = 2.5$	$4000 - 2750 = 1250$	$(2.5)(1250) = 3125$

$$\sum_{i=1}^6 (x_i - \bar{x})(y_i - \bar{y}) = 3125 + 1125 + 125 + 125 + 1125 + 3125 = 8750$$

# Ejemplo

Paso 3b:

$X_i$	$(X_i - \bar{X})^2$
1	$(-2.5)^2 = 6.25$
2	$(-1.5)^2 = 2.25$
3	$(-0.5)^2 = 0.25$
4	$(0.5)^2 = 0.25$
5	$(1.5)^2 = 2.25$
6	$(2.5)^2 = 6.25$

$$\sum_{i=1}^{6} (x_i - \bar{x})^2 = 6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25 = 17.5$$

# Ejemplo

Paso 3c:

$$\beta_1 = \frac{8750}{17.5}$$

$$\beta_1 = 500$$

# Ejemplo

## Paso 4: Cálculo de la intersección ( $\beta_0$ )

Identificando  $\beta_1 = 500$ , resolvemos:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_0 = 2750 - 500 \times 3.5$$

$$2750 - 1750$$

$$\beta_0 = 1000$$

# Ejemplo

## Paso 5: Modelo final

Usamos la fórmula:

$$y = \beta_0 + \beta_1 x + \epsilon$$

El modelo ajustado es:

$$y = 1000 + 500x$$

El ajuste de la línea busca minimizar  $\epsilon$ , así que al calcular  $\beta_0$  y  $\beta_1$ , estamos optimizando el modelo para que  $\epsilon$  sea lo más pequeño posible.

En este sentido,  $\epsilon$  no desaparece, sino que ya está considerado como parte del proceso de ajuste.

Con este modelo, podemos predecir el alquiler para un número dado de habitaciones.

# Ejemplo Regresión Lineal Múltiple

El objetivo es encontrar una relación lineal entre la calificación final del curso de IA ( $Y$ ) y las variables independientes: la calificación del examen parcial ( $x_1$ ) y el número de clases perdidas ( $x_2$ ).

<b>Estudiante</b>	<b><math>Y</math> (Calificación de IA)</b>	<b><math>x_1</math> (Calificación del examen)</b>	<b><math>x_2</math> (Clases perdidas)</b>
1	85	65	1
2	74	50	7
3	76	55	5
4	90	65	2
5	85	55	6
6	87	70	3
7	94	65	2
8	98	70	1
9	81	55	5
10	91	70	3
11	76	50	1
12	74	55	4

# Ejemplo

**Paso 1:** Relación entre las variables

Queremos encontrar una ecuación lineal que se ajuste a estos datos:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Donde:

- $\beta_0$ : Intersección (término constante).
- $\beta_1$ : Coeficiente asociado a la calificación del parcial
- $\beta_2$ : Coeficiente asociado al número de clases perdidas
- $\epsilon$  representa el término de error.

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,k-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,k-1} \\ \vdots & \vdots & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,k-1} \end{pmatrix}$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{k-1})^T$$

$\epsilon$  es un vector de errores aleatorios.

# Ejemplo

**Matriz de variables independientes (X)**

$$X = \begin{bmatrix} 1 & 65 & 1 \\ 1 & 50 & 7 \\ 1 & 55 & 5 \\ 1 & 65 & 2 \\ 1 & 55 & 6 \\ 1 & 70 & 3 \\ 1 & 65 & 2 \\ 1 & 70 & 1 \\ 1 & 55 & 5 \\ 1 & 70 & 3 \\ 1 & 50 & 1 \\ 1 & 55 & 4 \end{bmatrix}$$

**Vector de Valores observados (Y)**

$$Y = \begin{bmatrix} 85 \\ 74 \\ 76 \\ 90 \\ 85 \\ 87 \\ 94 \\ 98 \\ 81 \\ 91 \\ 76 \\ 74 \end{bmatrix}$$

# Ejemplo

**Paso 2:** Fórmula para calcular el vector  $\beta$

Usamos la fórmula de mínimos cuadrados en forma matricial:

$$\beta = (X^T X)^{-1} X^T Y$$

Donde:

- $X^T$ : Transpuesta de la matriz  $X$ .
- $(X^T X)^{-1}$ : Inversa de la matriz  $X^T X$ .
- $\beta$ : Vector que contiene los coeficientes  $\beta_0, \beta_1, \beta_2$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,k-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,k-1} \\ \vdots & \vdots & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,k-1} \end{pmatrix} \quad \beta = (\beta_0, \beta_1, \dots, \beta_{k-1})^T$$

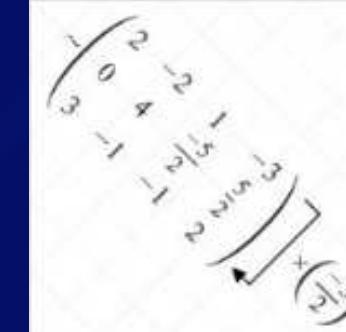
# Ejemplo

**Paso 2:** Fórmula para calcular  $\beta$

Usamos la fórmula de mínimos cuadrados en forma matricial:

$$\beta = (X^T X)^{-1} X^T Y$$

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 65 & 50 & 55 & 65 & 55 & 70 & 65 & 70 & 55 & 70 & 50 & 55 \\ 1 & 7 & 5 & 2 & 6 & 3 & 2 & 5 & 4 & 3 & 1 & 4 \end{pmatrix}$$



## Calculadora de Matrices

Cálculo de suma de matrices, de diferencia de matrices, de producto de matrices, matriz inversa, de determinante, de matriz transpuesta; Reducir...

matrixcalc

# Ejemplo

Paso 3a: Calcular  $X^T X$

$$\left( \begin{array}{cccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 65 & 50 & 55 & 65 & 55 & 70 & 65 & 70 & 55 & 70 \\ 1 & 7 & 5 & 2 & 6 & 3 & 2 & 5 & 4 & 3 \\ \vdots & & & & & & & & & \end{array} \right) \cdot \left( \begin{array}{ccc} 1 & 65 & 1 \\ 1 & 50 & 7 \\ 1 & 55 & 5 \\ 1 & 65 & 2 \\ 1 & 55 & 6 \\ 1 & 70 & 3 \\ 1 & 65 & 2 \\ 1 & 70 & 5 \\ 1 & 55 & 4 \\ 1 & 70 & 3 \\ 1 & 50 & 1 \\ 1 & 55 & 4 \\ \vdots & & \end{array} \right) = \left( \begin{array}{ccc} 12 & 725 & 43 \\ 725 & 44475 & 2540 \\ 43 & 2540 & 195 \\ \vdots & & \end{array} \right)$$

# Ejemplo

Paso 3b: Calcular  $(X^T X)^{-1}$

$$\begin{pmatrix} 12 & 725 & 43 \\ 725 & 44475 & 2540 \\ 43 & 2540 & 195 \end{pmatrix}^{-1} = \begin{pmatrix} 7,654748 & -0,110822 & -0,244443 \\ -0,110822 & 0,001692 & 0,002395 \\ -0,244443 & 0,002395 & 0,027830 \end{pmatrix}$$

# Ejemplo

Paso 3c: Calcular  $X^T Y$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 65 & 50 & 55 & 65 & 55 & 70 & 65 & 70 & 55 & 70 & 50 & 55 \\ 1 & 7 & 5 & 2 & 6 & 3 & 2 & 5 & 4 & 3 & 1 & 4 \end{pmatrix} \cdot \begin{pmatrix} 85 \\ 74 \\ 76 \\ 90 \\ 85 \\ 87 \\ 94 \\ 98 \\ 81 \\ 91 \\ 76 \\ 74 \end{pmatrix} = \begin{pmatrix} 1011 \\ 61685 \\ 3581 \end{pmatrix}$$

# Ejemplo

Paso 3d: Calcular  $\beta$

$$\begin{pmatrix} 88841 & -6431 & -2837 \\ \hline 11606 & 58030 & 11606 \\ -6431 & 491 & 139 \\ \hline 58030 & 290150 & 58030 \\ -2837 & 139 & 323 \\ \hline 11606 & 58030 & 11606 \end{pmatrix} \cdot \begin{pmatrix} 1011 \\ 61685 \\ 3581 \end{pmatrix} = \begin{pmatrix} 27,546700 \\ 0,921678 \\ 0,284250 \end{pmatrix}$$

La solución nos dará los coeficientes  $\beta_0$ ,  $\beta_1$ , y  $\beta_2$ , que representan:

- $\beta_0$ : El término constante.
- $\beta_1$ : El impacto de la calificación del parcial en la nota final del curso de IA.
- $\beta_2$ : El impacto de las clases perdidas en la nota final del curso de IA.

# Ejemplo

## Paso 4: Modelo final

==== MODELO DE REGRESIÓN LINEAL MÚLTIPLE ===

Intercepto ( $\beta_0$ ): 27.5467

Coeficiente para Examen ( $\beta_1$ ): 0.9217

Coeficiente para Clases\_Perdidass ( $\beta_2$ ): 0.2842

Ecuación del modelo:

$$Estadistica = 27.55 + 0.92*Examen + 0.28*Clases_Perdidass$$



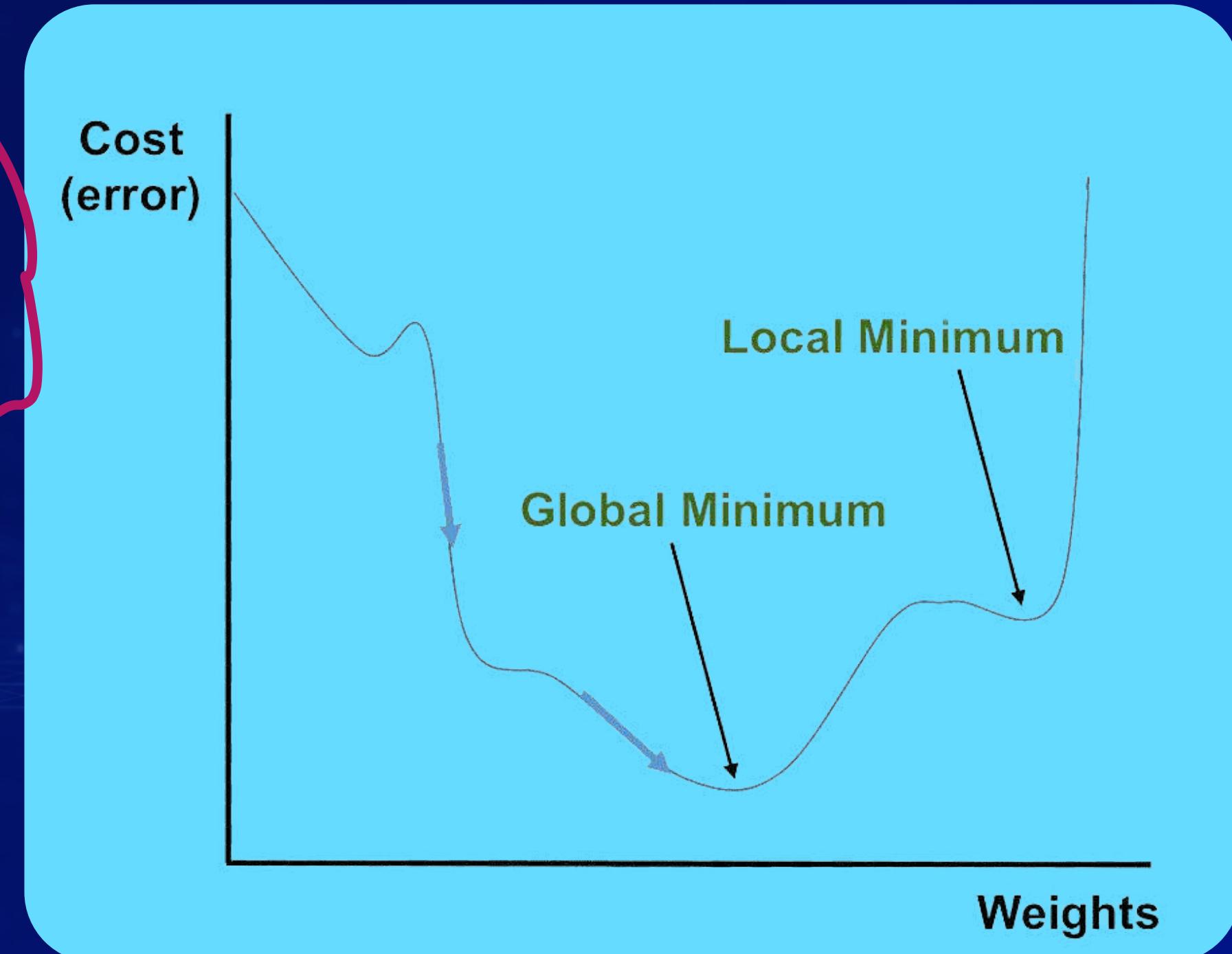
### Calculadora de Matrices

Cálculo de suma de matrices, de diferencia de matrices, de producto de matrices, matriz inversa, de determinante, de matriz transpuesta; Reducir...

matrixcalc

# Gradiente Descendente

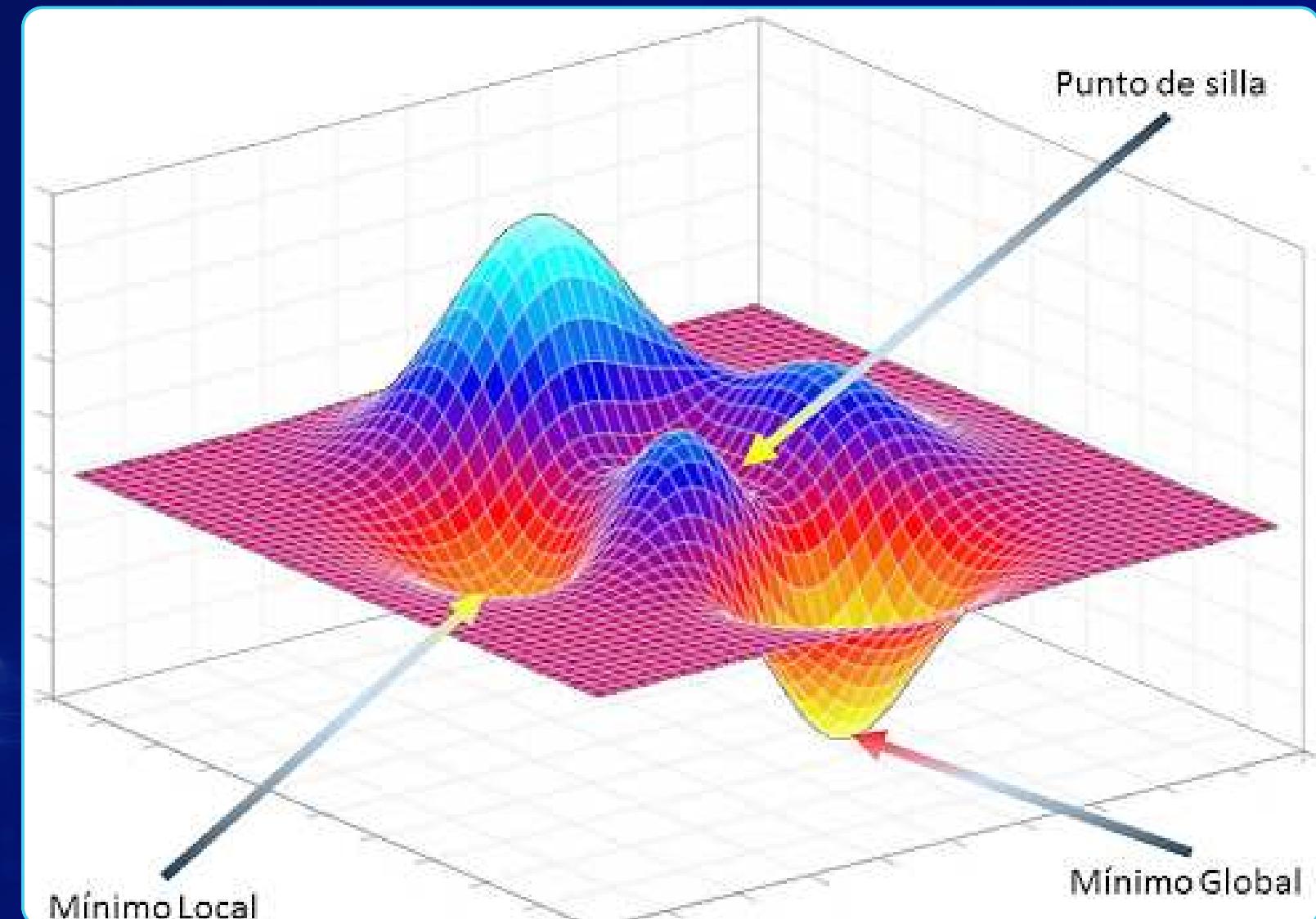
Este es un enfoque numérico utilizado para minimizar funciones de costo, especialmente en modelos complejos o con grandes volúmenes de datos.



# Gradiente Descendente

Minimización iterativa de la función de costo.

El gradiente indica la dirección de mayor aumento de la función de costo. En el gradiente descendente, nos movemos en la dirección opuesta (por eso "descendente") para encontrar el mínimo de  $J$ .



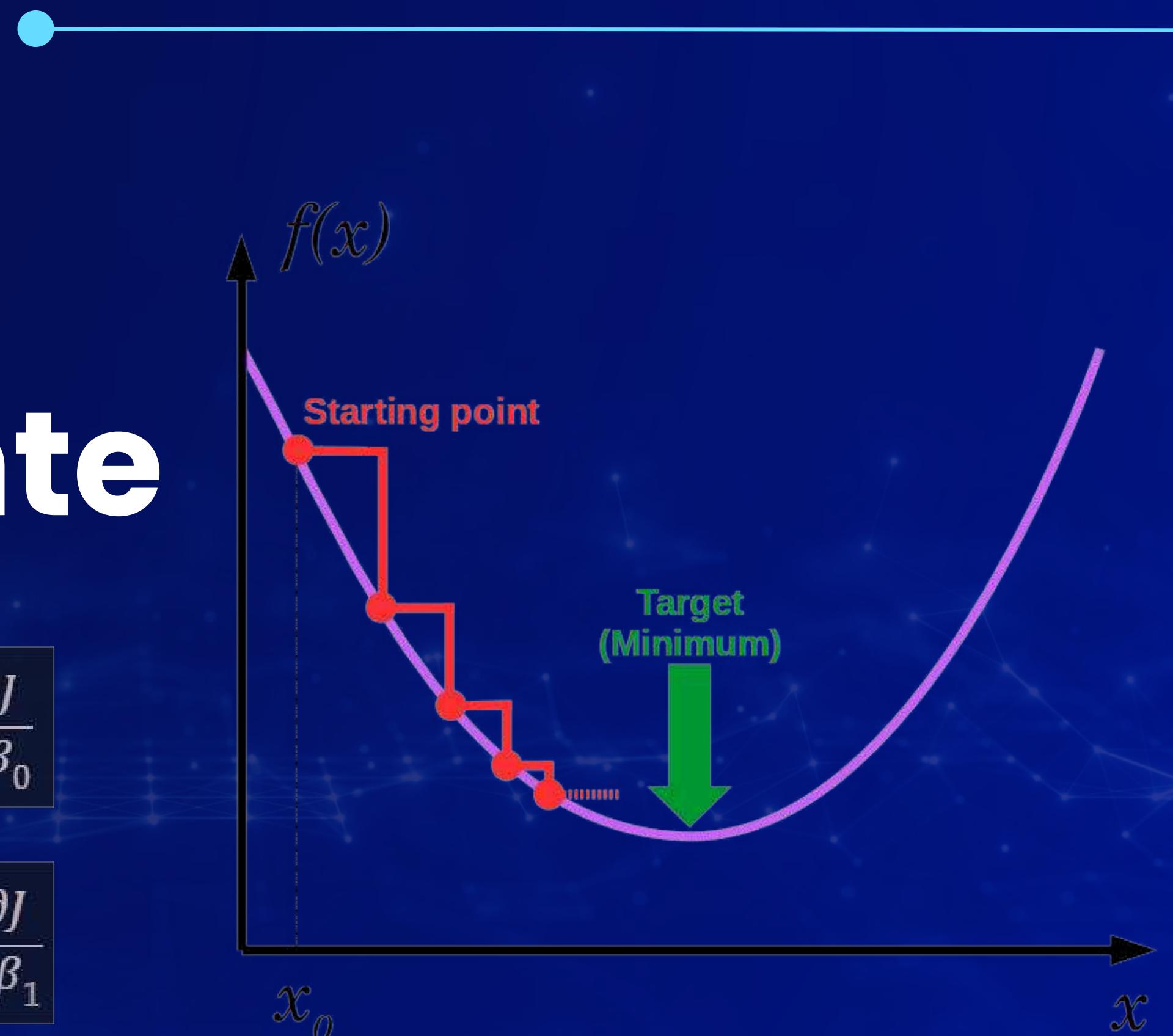
# Gradiente Descendente

Para  $\beta_0$  (intersección):

$$\beta_0 := \beta_0 - \alpha \cdot \frac{\partial J}{\partial \beta_0}$$

Para  $\beta_1$  (pendiente):

$$\beta_1 := \beta_1 - \alpha \cdot \frac{\partial J}{\partial \beta_1}$$

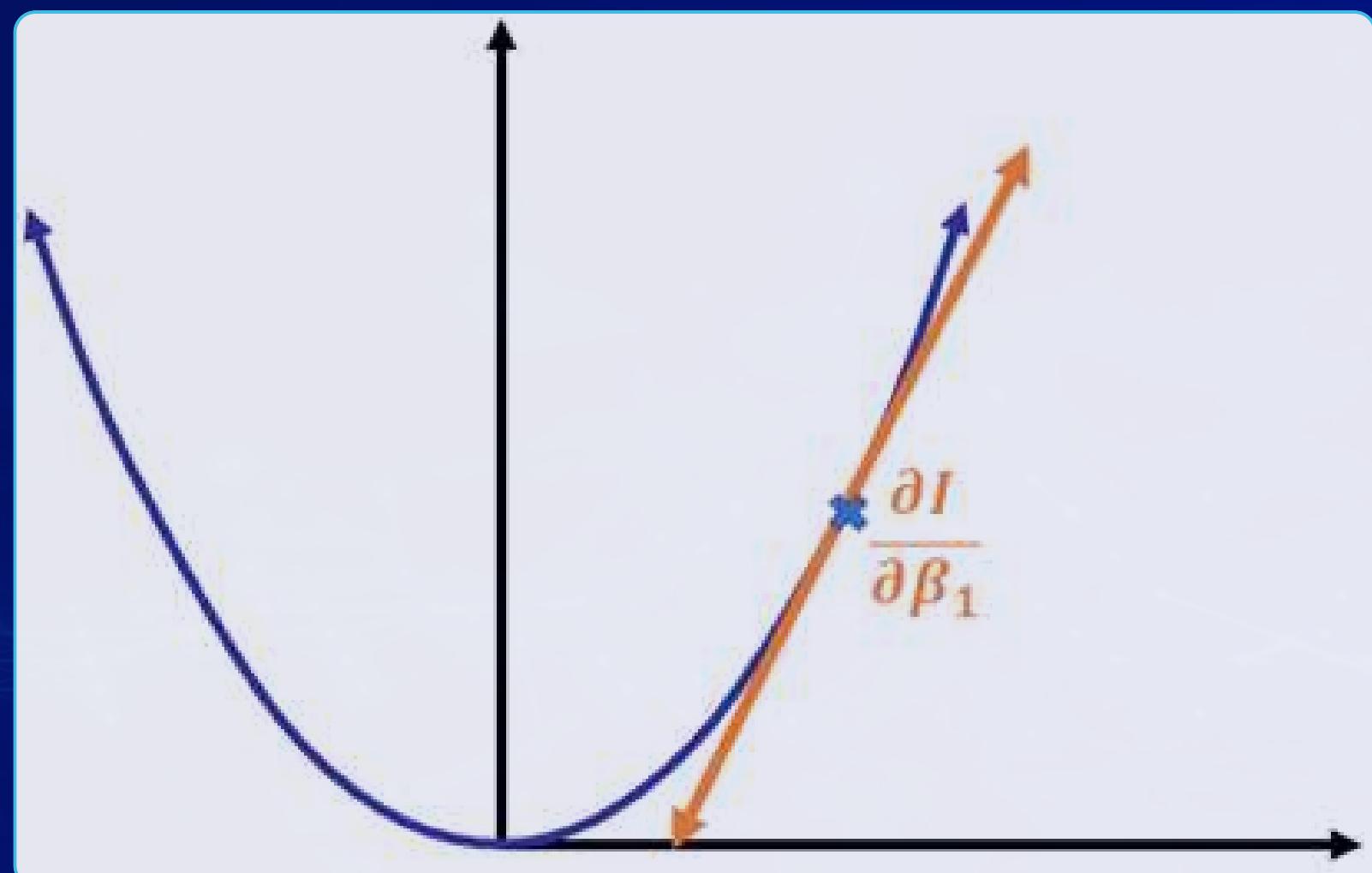


# Gradiente Descendente

La actualización de los parámetros se hace utilizando las derivadas parciales de  $J$ :

$$\frac{\partial J}{\partial \beta_0} \quad \text{y} \quad \frac{\partial J}{\partial \beta_1}$$

Estas derivadas nos indican cómo cambian los errores (el costo) cuando se ajustan  $\beta_0$  y  $\beta_1$ .

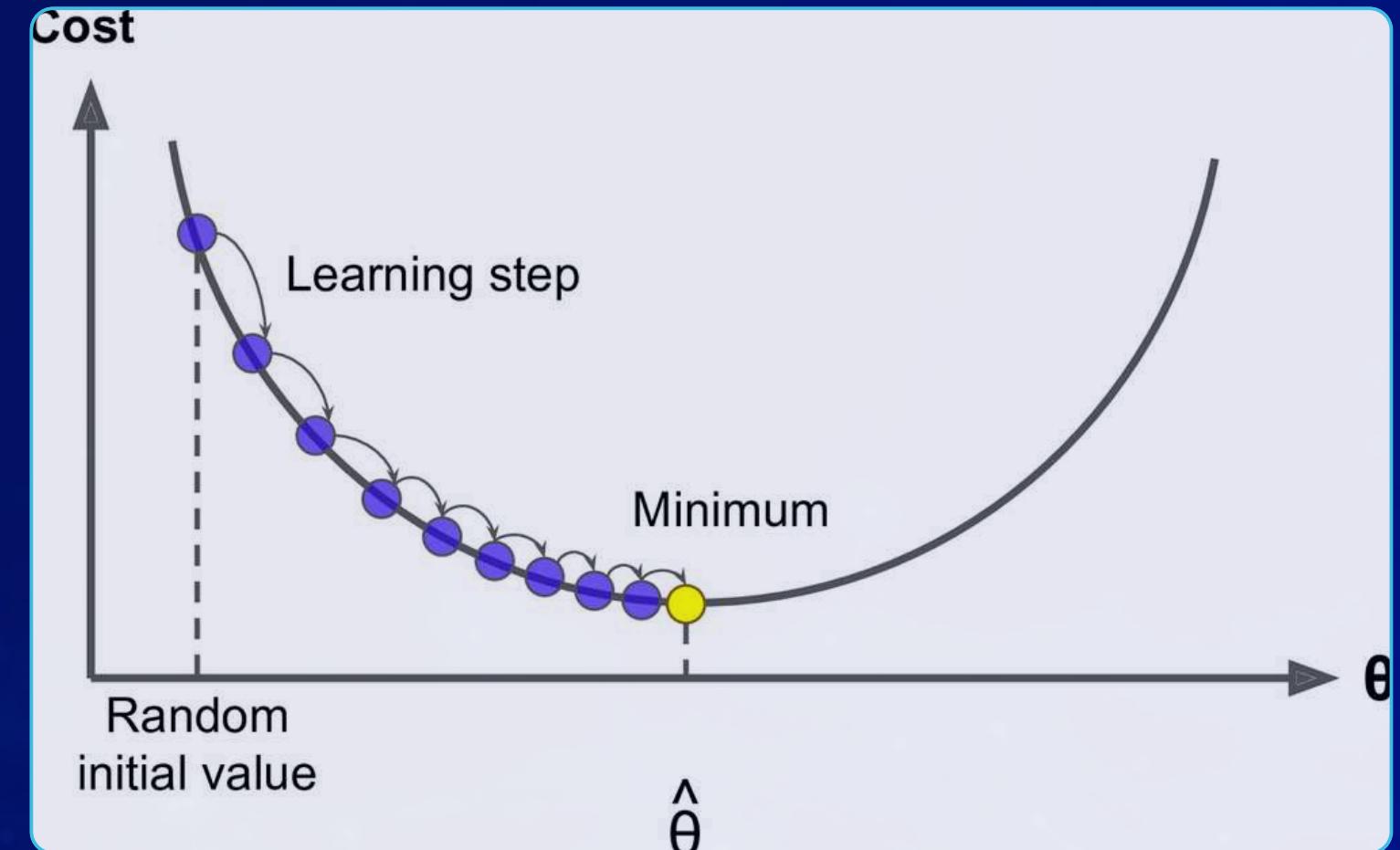




# Gradiente Descendente

Donde:

- $\alpha$ : Tasa de aprendizaje, un número pequeño (como 0.01) que controla qué tan rápido avanzamos.

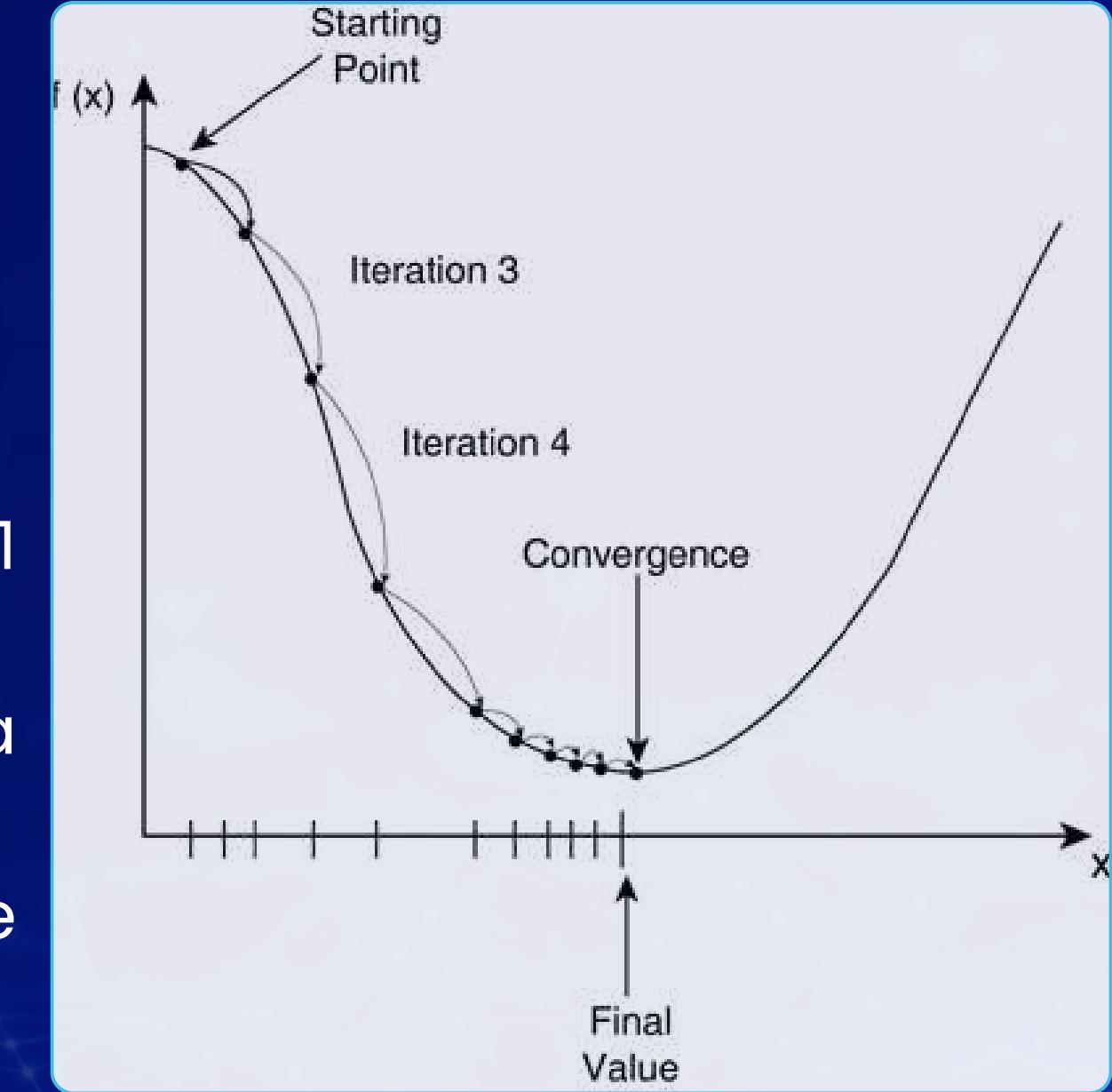


$$\frac{\partial J}{\partial \beta_0} = -\frac{1}{m} \sum_{i=1}^m (Y_i - (\beta_0 + \beta_1 \cdot X_i))$$
$$\frac{\partial J}{\partial \beta_1} = -\frac{1}{m} \sum_{i=1}^m (Y_i - (\beta_0 + \beta_1 \cdot X_i)) \cdot X_i$$



# Proceso Iterativo

- 1. Inicialización:** Asigna valores iniciales a  $\beta_0$  y  $\beta_1$  (pueden ser 0 o valores aleatorios).
- 2. Calcular el costo inicial:** Usar los valores iniciales para calcular  $J$ .
- 3. Actualizar los parámetros:** Usar las fórmulas de actualización para  $\beta_0$  y  $\beta_1$ .
- 4. Repetir:** Volver a calcular el costo  $J$  y repetir hasta que:
  - $J$  sea lo suficientemente pequeño.
  - Ó el número de iteraciones alcance un límite predefinido.





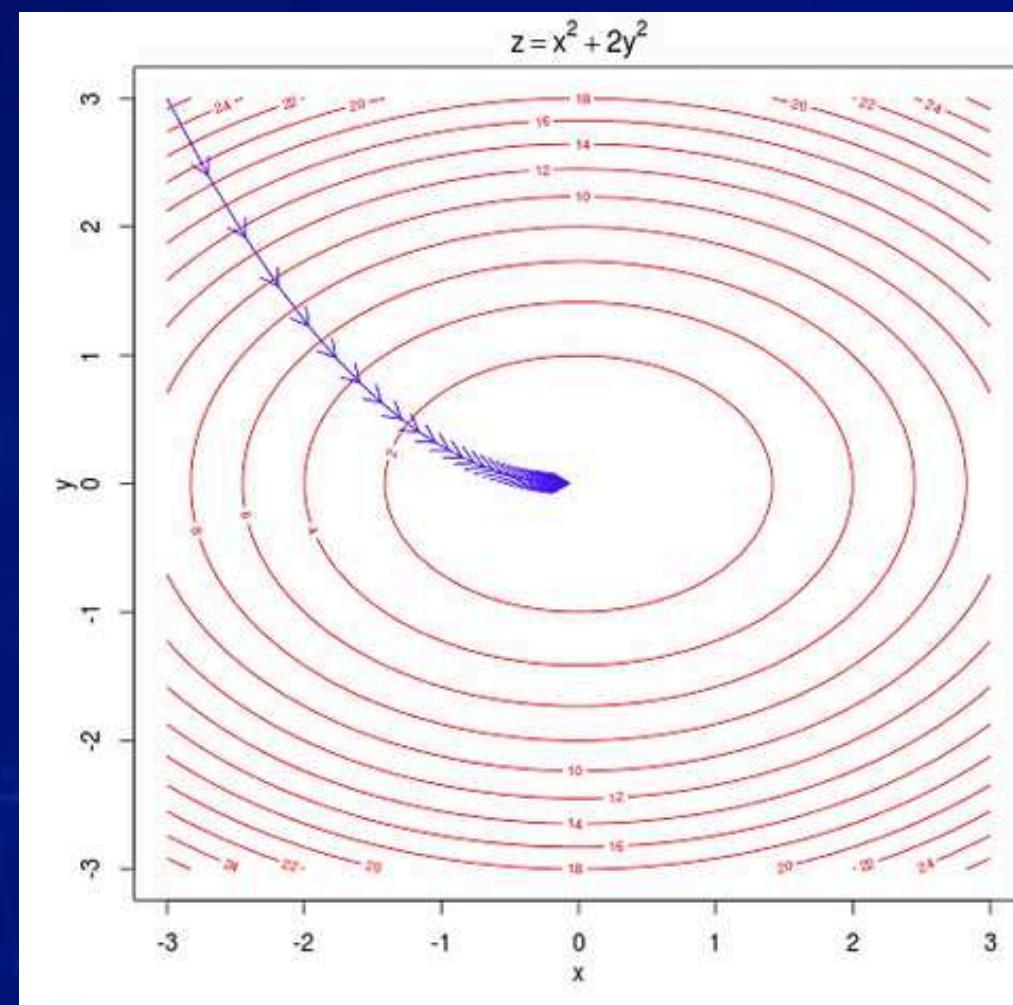
# Ejemplo intuitivo

Imagina que tienes un modelo inicial:

$$Y_{predicho} = \beta_0 + \beta_1 x$$

Si tu predicción inicial está lejos de los valores reales ( $Y_{real}$ ), el gradiente descendente ajustará  $\beta_0$  y  $\beta_1$  paso a paso para alinear mejor las predicciones con los datos.

Después de suficientes iteraciones, los valores de  $\beta_0$  y  $\beta_1$  convergerán a los mismos que encontrarías usando un método más exacto como los mínimos cuadrados.



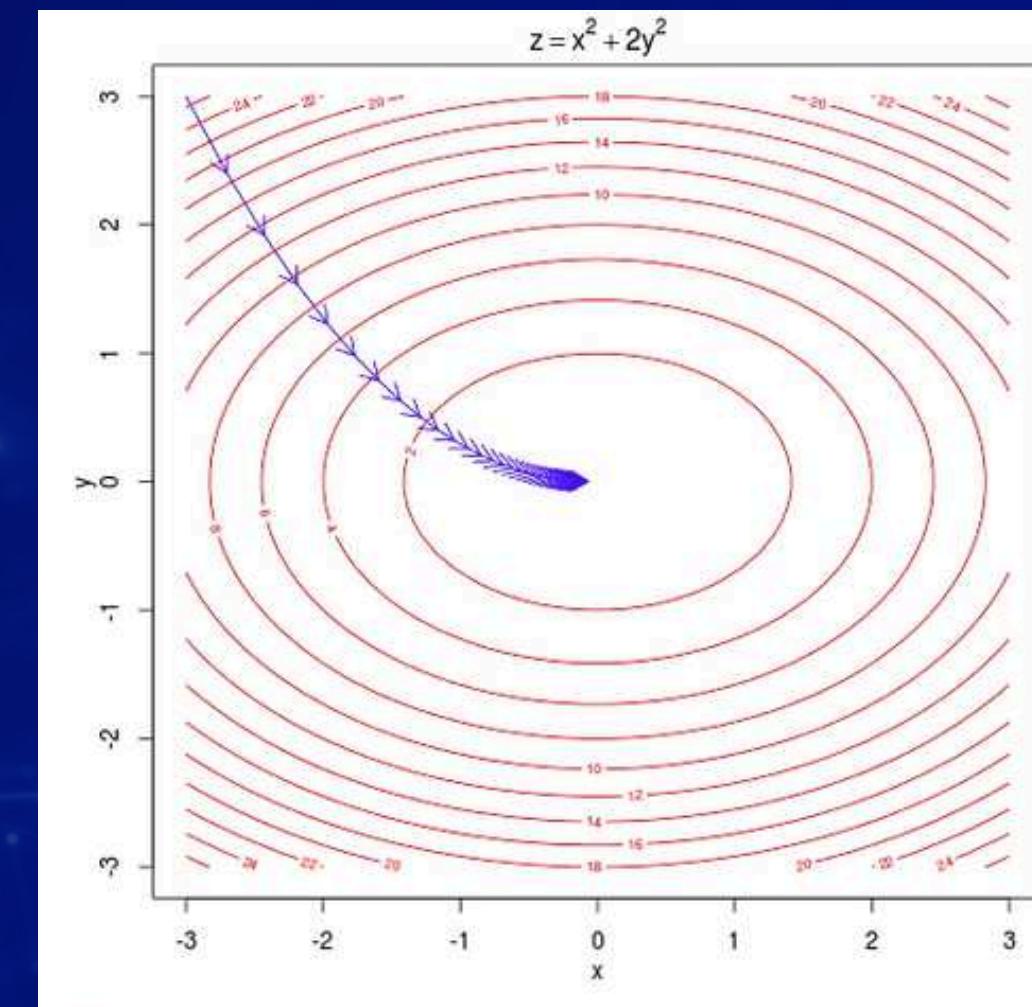




# Cuando usar...

### mínimos cuadrados ordinarios:

- Cuando el número de datos es mayor que el número de variables y la matriz  $X^T X$  es invertible
- Cuando queremos una solución exacta y rápida (cuando  $X$  es de tamaño manejable)
- Cuando no hay demasiadas características (variables), porque la inversión de la matriz  $(X^T X)^{-1}$  es computacionalmente costosa si el numero de variables es grande.



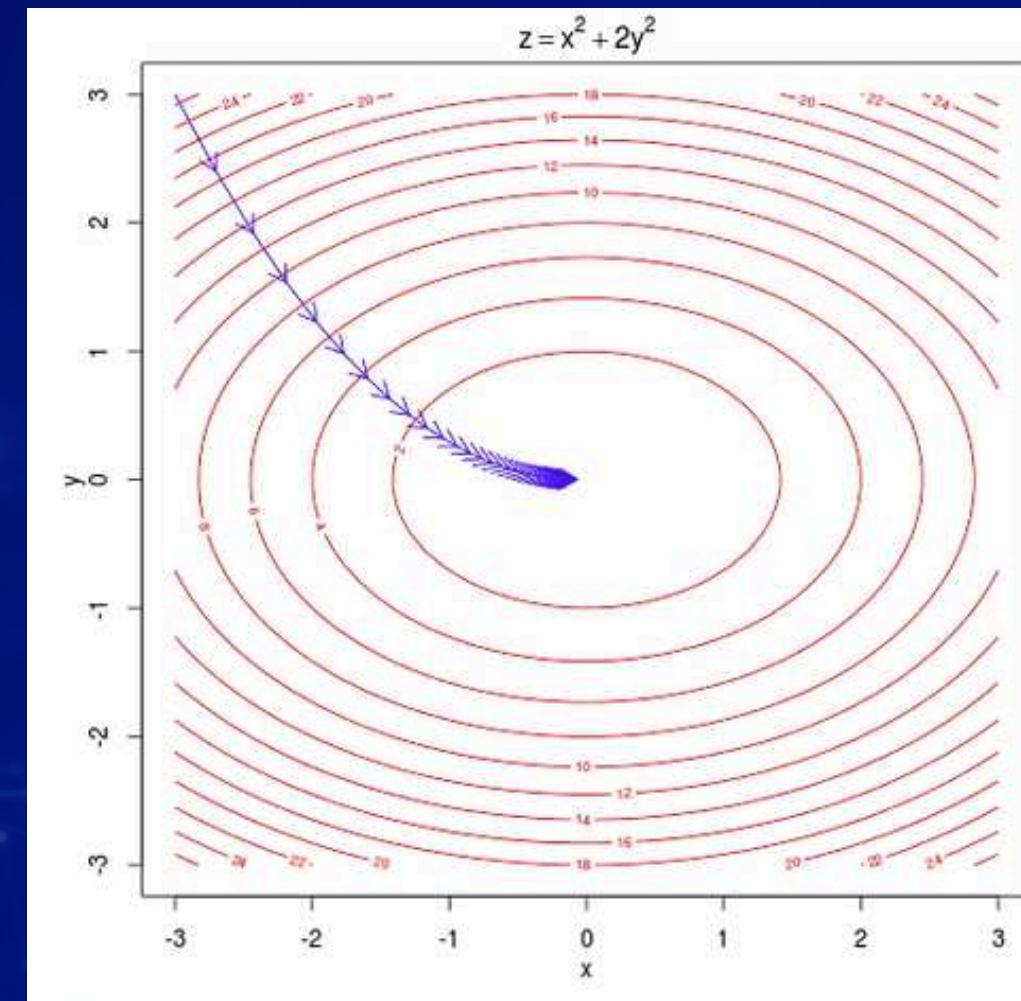
se usa cuando podemos calcular la solución exacta y tenemos una cantidad razonable de datos y características.



# Cuando usar...

### descenso del gradiente:

- Cuando el conjunto de datos es muy grande (alta dimensionalidad o muchas muestras), porque evitar calcular  $(X^T X)^{-1}$  ahorra memoria y tiempo.
- Cuando no podemos invertir  $X^T X$  porque es singular o mal condicionada.
- Cuando trabajamos con modelos de aprendizaje profundo, donde OLS no es viable debido a la no linealidad y la gran cantidad de parámetros.



es más flexible y escalable, pero requiere elegir una tasa de aprendizaje y más iteraciones para converger a la solución óptima



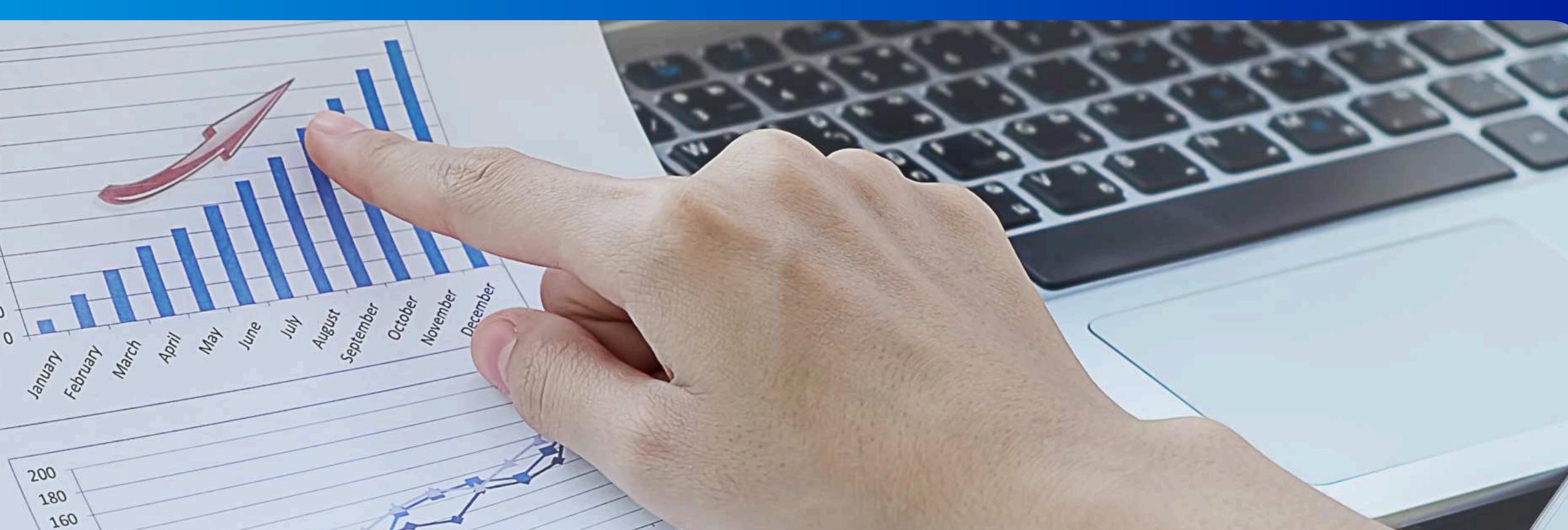
# Evaluación del modelo

Para medir qué tan buena es nuestra aproximación, usamos el error sobre los datos de entrenamiento



### Métricas principales:

- Error Cuadrático Medio (MSE): Penaliza errores grandes.
- Coeficiente de Determinación: Explica qué porcentaje de la varianza de  $y$  es explicado por  $x$ .





# Métricas

Estas métricas nos ayudan a entender qué tan bien el modelo es capaz de aproximarse a los valores reales de la variable que estamos intentando predecir.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

Where,

$\hat{y}$  – predicted value of  $y$   
 $\bar{y}$  – mean value of  $y$

# Evaluación del Modelo de Regresión

Métrica	Descripción	Ventajas	Desventajas	Cuándo usar
Error Absoluto Medio (MAE)	Promedio de las diferencias absolutas entre los valores predichos y los valores reales.	Robusto a valores atípicos. Fácil de interpretar en las unidades originales de la variable objetivo.	No penaliza los errores grandes tan fuertemente como el MSE.	Cuando los valores atípicos son comunes y se desea una métrica fácil de interpretar.
Error Cuadrático Medio (MSE)	Promedio de los cuadrados de las diferencias entre los valores predichos y los valores reales.	Penaliza más los errores grandes. Ampliamente utilizada en optimización.	Sensible a valores atípicos. No es tan fácil de interpretar en las unidades originales de la variable objetivo.	Cuando los errores grandes son especialmente indeseables y se desea una métrica matemáticamente conveniente.
Raíz del Error Cuadrático Medio (RMSE)	Raíz cuadrada del MSE.	Penaliza los errores grandes, pero es más interpretable que el MSE porque está en las unidades originales de la variable objetivo.	Sensible a valores atípicos.	Cuando los errores grandes son indeseables y se desea una métrica en las unidades originales de la variable objetivo.
Coeficiente de Determinación ( $R^2$ )	Proporción de la varianza de la variable dependiente que es predecible a partir de las variables independientes.	Indica qué tan bien se ajusta el modelo a los datos. Varía entre 0 y 1, donde 1 indica un ajuste perfecto.	No indica si el modelo es sesgado. No indica si las variables independientes son estadísticamente significativas.	Para evaluar el ajuste general del modelo y comparar diferentes modelos.

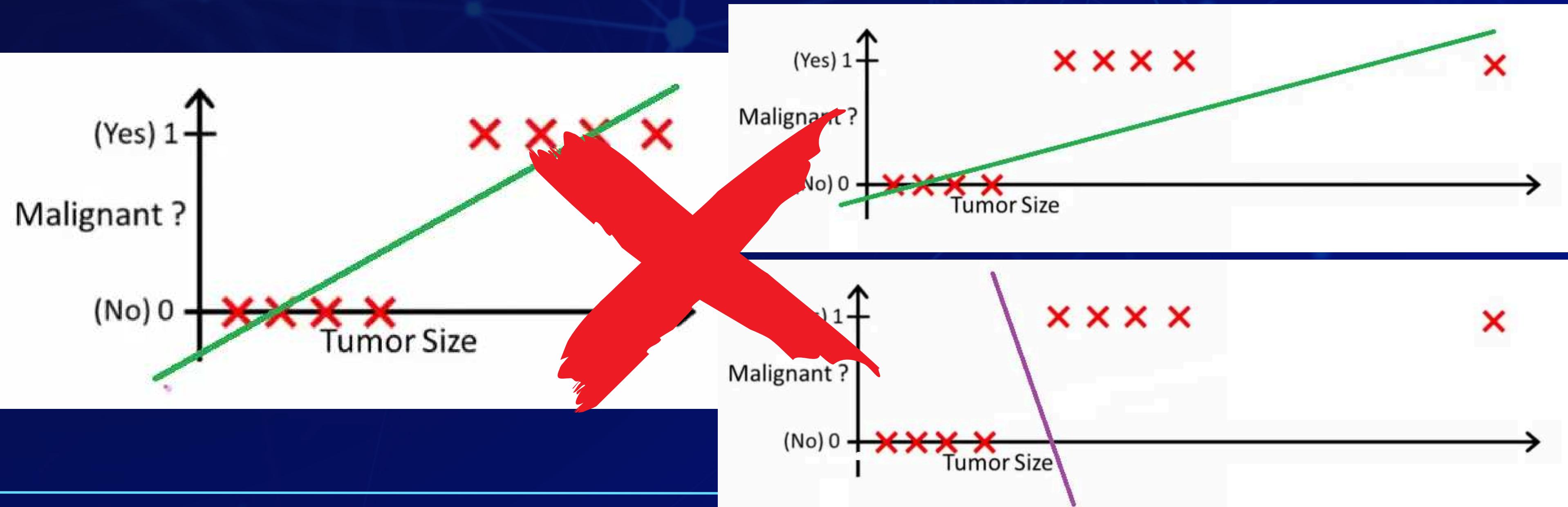


# Diferencias Claves entre Función de Coste y Métricas de Evaluación

Característica	Función de Coste	Métrica de Evaluación
Propósito	Optimizar el modelo minimizando errores.	Evaluar la calidad del modelo.
Cuándo se usa	Durante el entrenamiento.	Después del entrenamiento.
¿Afecta el modelo?	Sí, el modelo ajusta los pesos para minimizarla.	No, solo mide qué tan bueno es el modelo.
Ejemplos	MSE, MAE, Log-Cosh, Regularización.	RMSE, R <sup>2</sup> , MAPE.



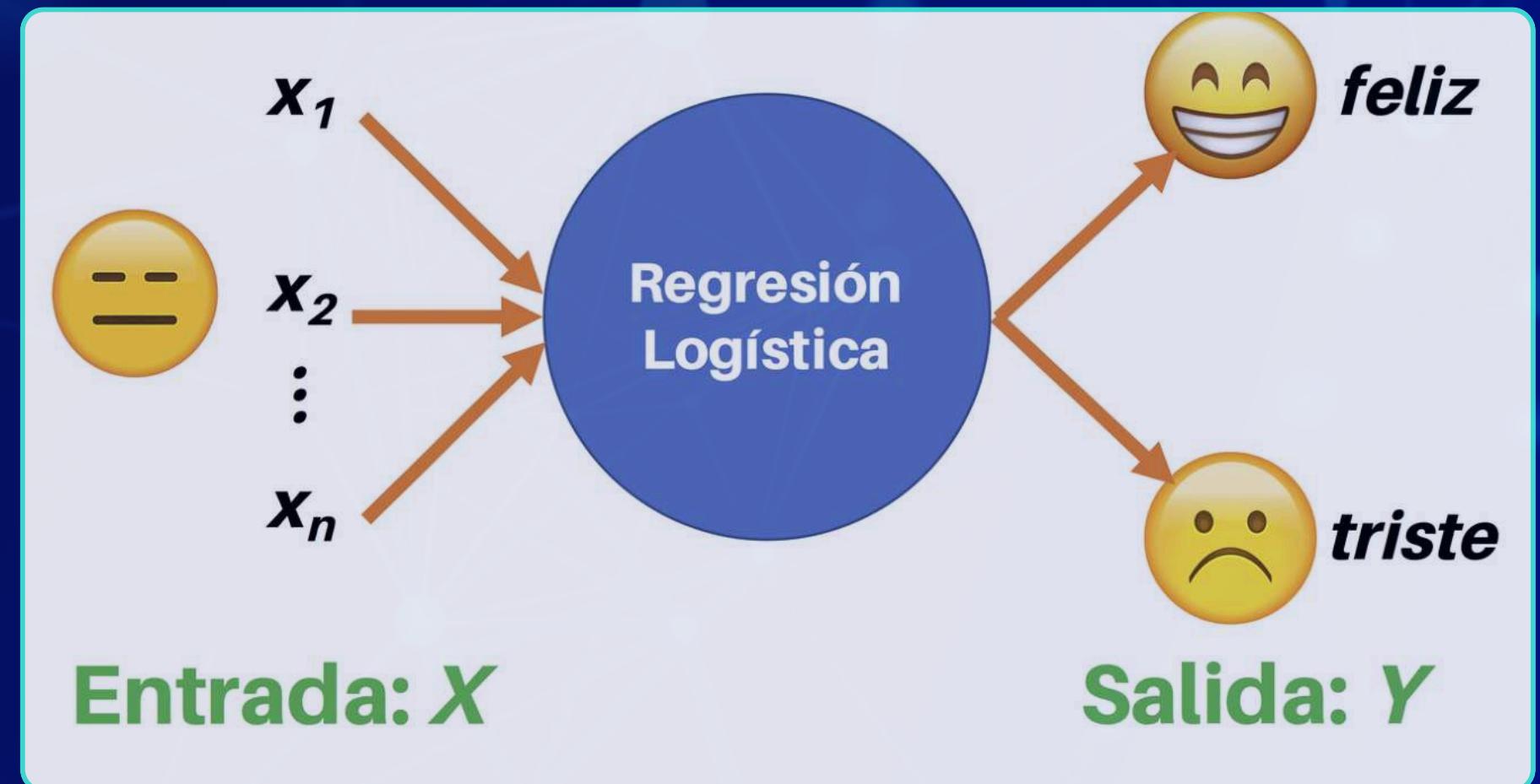
# Regresión Logística



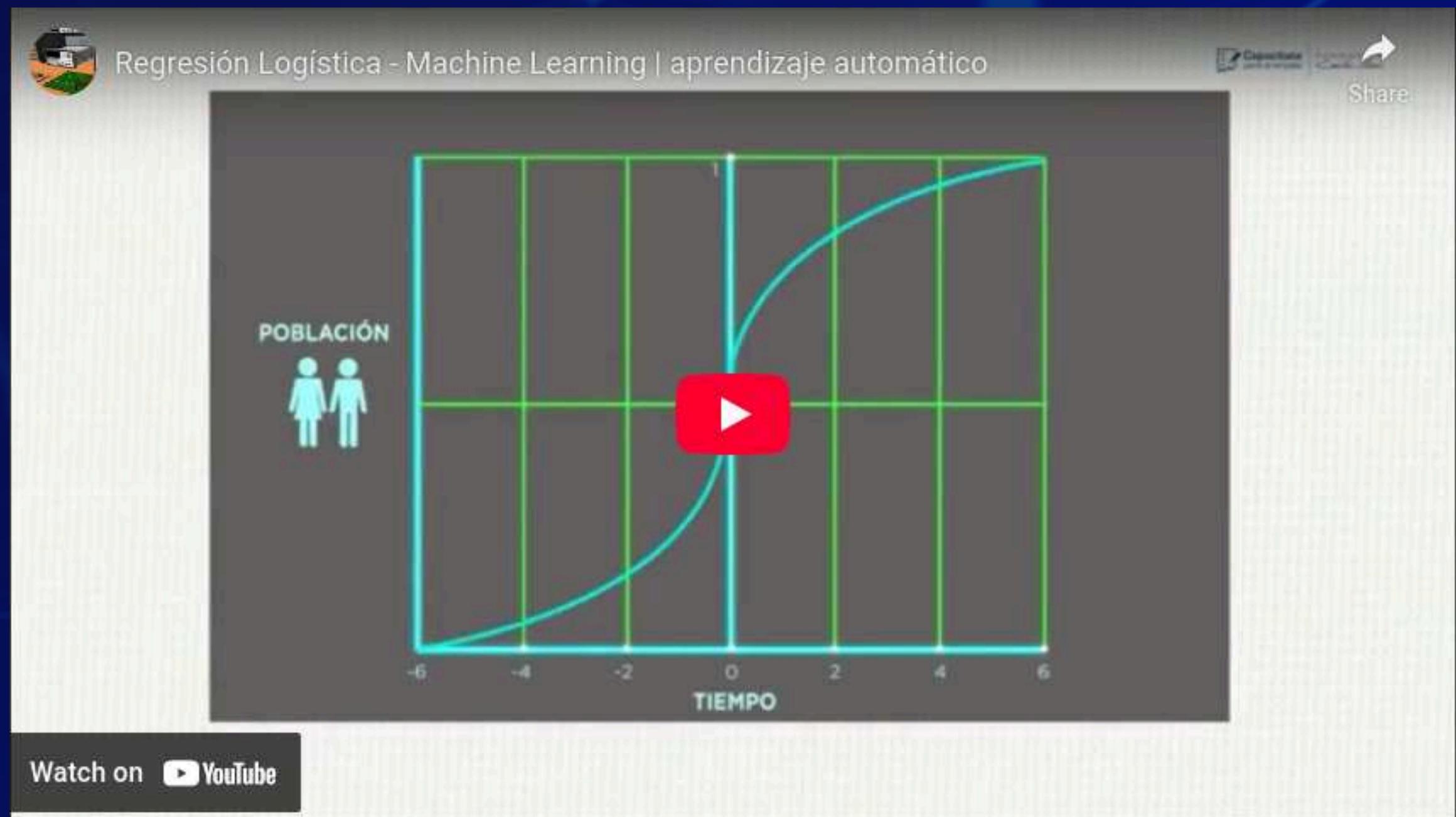


# Regresión Logística

La regresión logística es un algoritmo de aprendizaje automático supervisado que se utiliza para tareas de clasificación, cuyo objetivo es predecir la probabilidad de que una instancia pertenezca a una clase determinada.



Objetivo: Encontrar  $\beta_0$  y  $\beta_1$  que minimicen el error

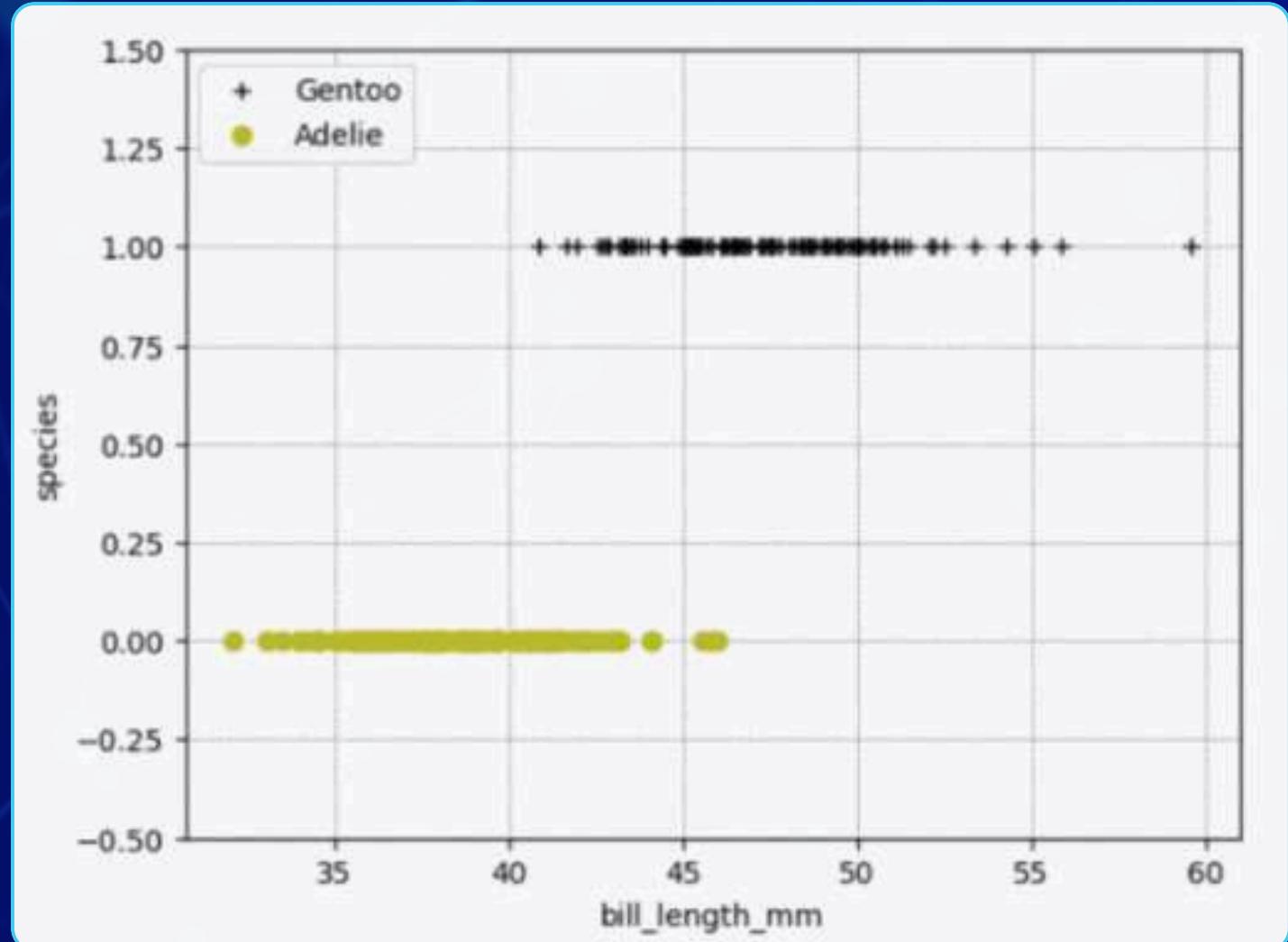




# Regresión Logística

**La regresión logística se modela con una ecuación de la misma forma solo que aplicando la función sigmoidal.**

Recordemos que estamos hablando de un modelo de clasificación y su objetivo no es predecir un valor sino encontrar la probabilidad de que una sea clase (0) o de otra (1).



Buscar una ecuación que nos permita predecir la probabilidad de que una observación pertenezca a una de dos clases.



# Regresión Logística

Para entender cómo funciona, primero recordemos el caso de la regresión lineal. En regresión lineal, modelamos una salida  $\mathcal{Y}$  como una combinación lineal de las entradas:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

Sin embargo, en problemas de clasificación binaria, necesitamos que la salida esté entre 0 y 1 para que pueda interpretarse como una probabilidad. Aquí es donde entra en juego la función logística (también conocida como sigmoide ).

# El Problema: Modelar una Probabilidad Binaria

Queremos predecir la probabilidad  $p$  de que un evento ocurra (ej.:  $y=1$ ) dado un predictor  $X$ .

Limitación:

- $p$  debe estar entre 0 y 1.
- Una recta  $p=\beta_0+\beta_1X$  puede generar valores fuera de  $[0,1]$  (ej.:  $p=-0.2$  o  $p=1.5$ ), lo cual no tiene sentido.

**Solución:** Transformar  $p$  para quitar las limitaciones

# Odds

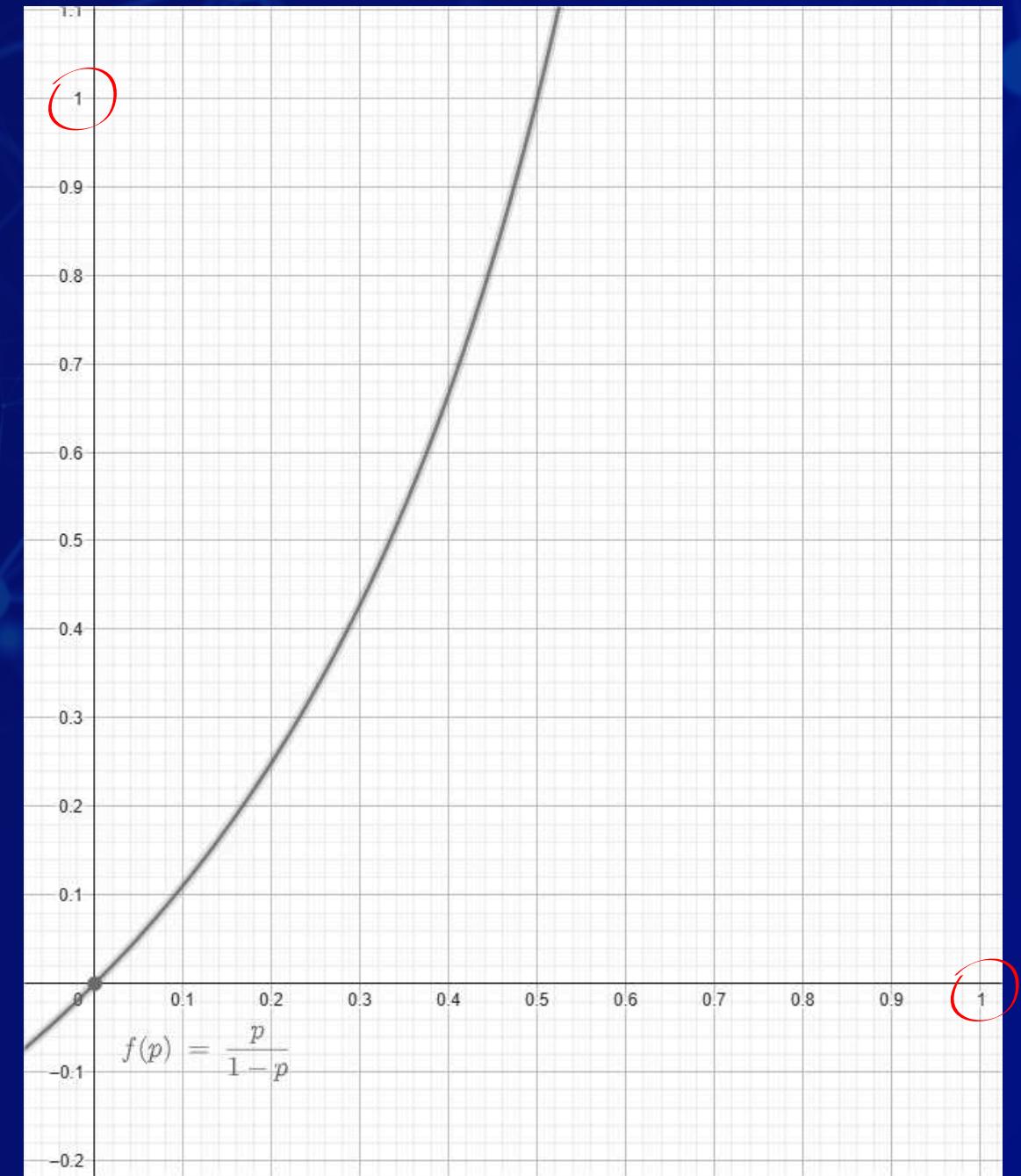
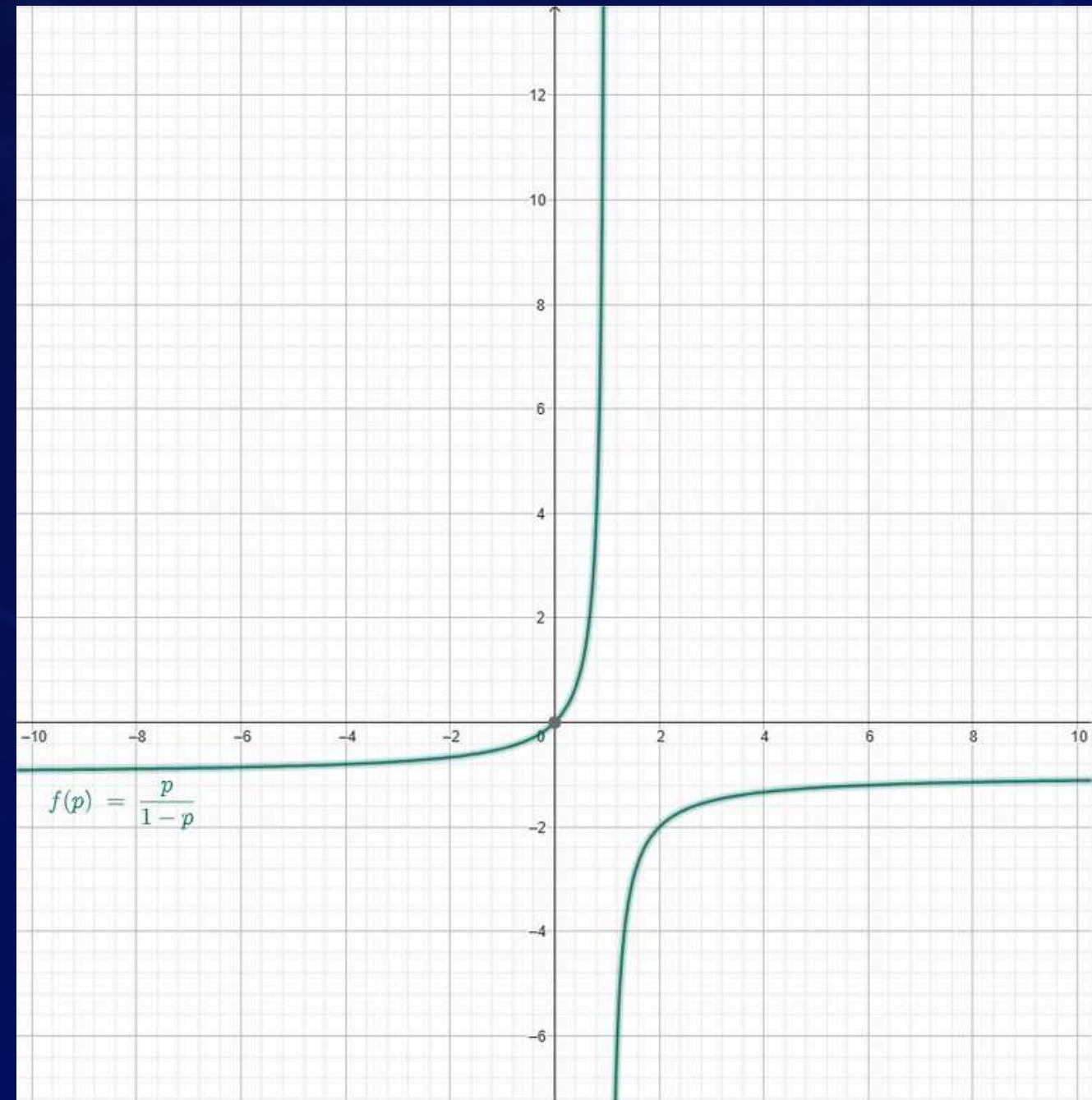
Los odds son una forma alternativa de expresar la probabilidad de un evento.

Mientras que la probabilidad ( $p$ ) indica la proporción de veces que esperamos que ocurra un evento, los odds comparan directamente las posibilidades de que ocurra el evento frente a que no ocurra.

$$Odds = \frac{p}{1 - p}$$

- Razón de probabilidades
  - Esto significa que el evento es ciertas veces tan probable como no ocurrir.
  - Para probabilidades estrictamente entre 0 y 1, es decir,  $0 < p < 1$ .

# Odds



# Logit (Log-Odds)

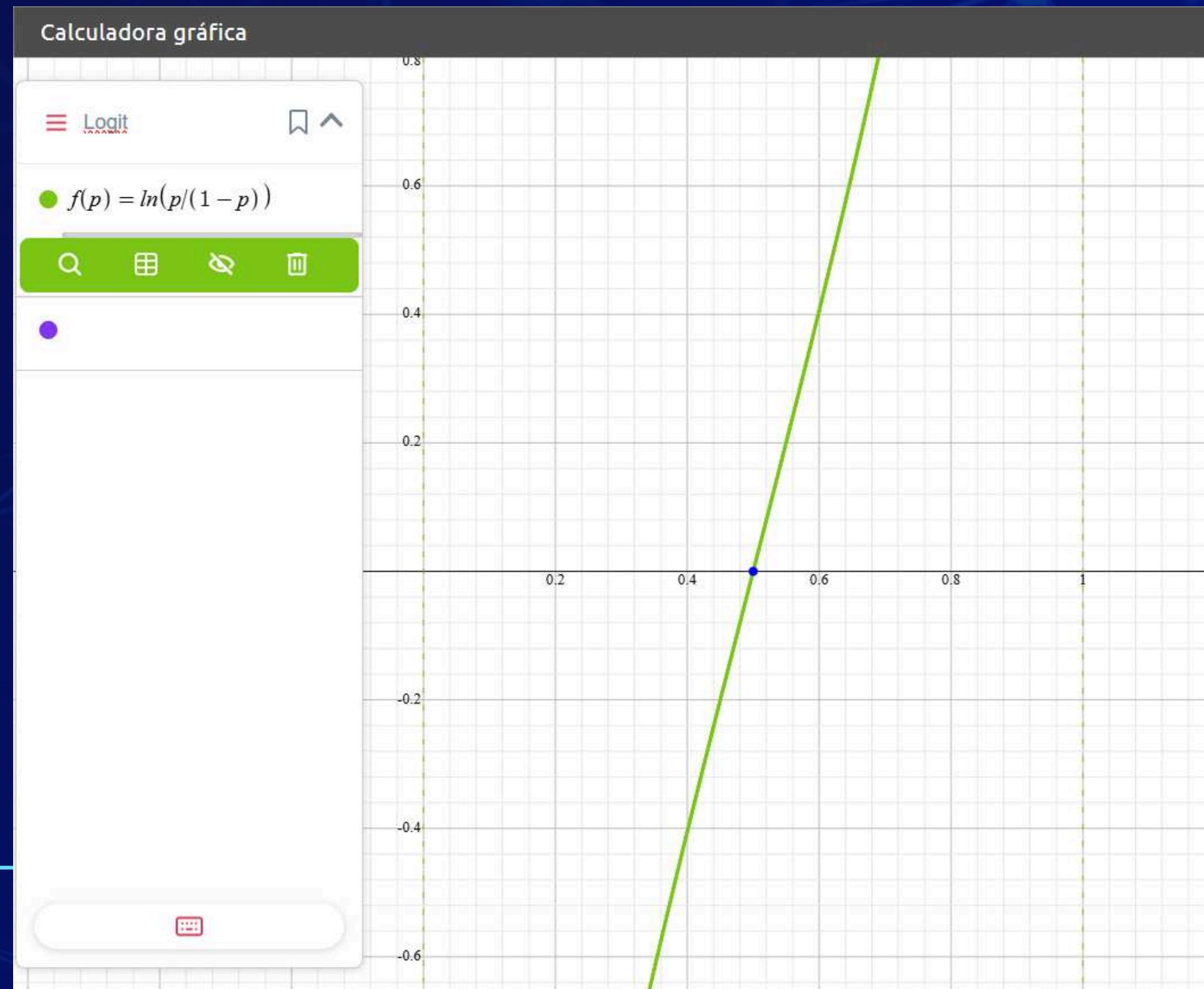
La función logit actúa como un puente entre las probabilidades y los modelos lineales, permitiéndonos trabajar con una estructura matemática más manejable.

$$\ln\left(\frac{p}{1 - p}\right)$$

◦ Asume un comportamiento lineal

"Piensa en los odds como una escala elástica: el logit la estira para que una recta pueda ajustarse sin romper las reglas de la probabilidad."

# Logit (Log-Odds)



# Logit (Log-Odds)

La función logit es clave en la regresión logística porque transforma probabilidades (que están restringidas al intervalo [0,1]) en valores que pueden variar en todo el rango real (- $\infty$  a  $\infty$ ). Esto permite modelar las probabilidades usando una función lineal.

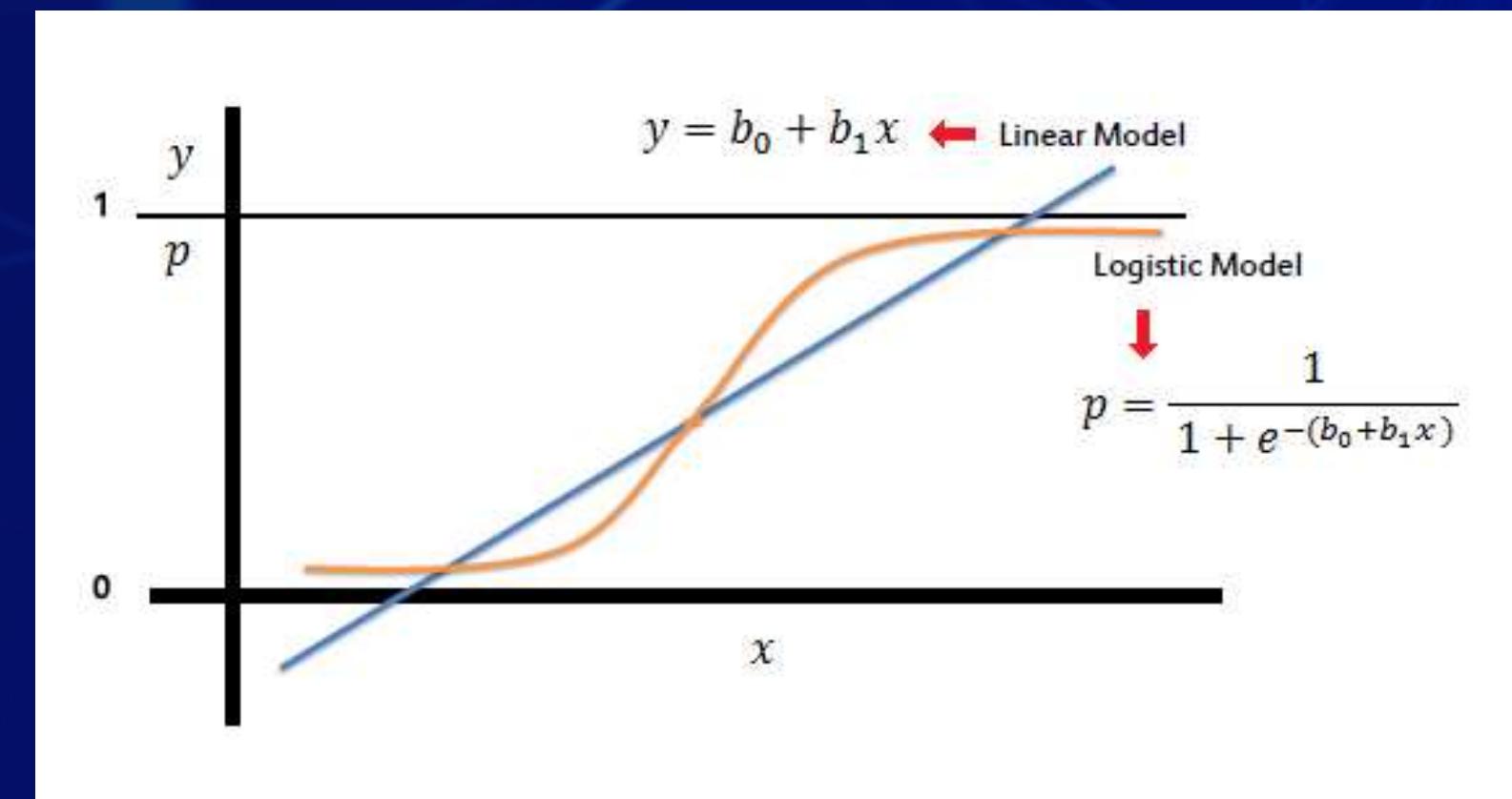
$$\ln \left( \frac{p}{1 - p} \right) = \beta_0 + \beta_1 X$$

- Es por esta razón que este método lleva de nombre “Regresión” Logística

# Función Sigmoidal

Al despejar para  $p$ , la ecuación queda:

$$p = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$



**Objetivo:** Encontrar  $\beta_0$  y  $\beta_1$  que minimicen el error

# Función Sigmoidal

Divide el numerador y denominador de la ecuación por  $e^{\beta_0 + \beta_1 x}$

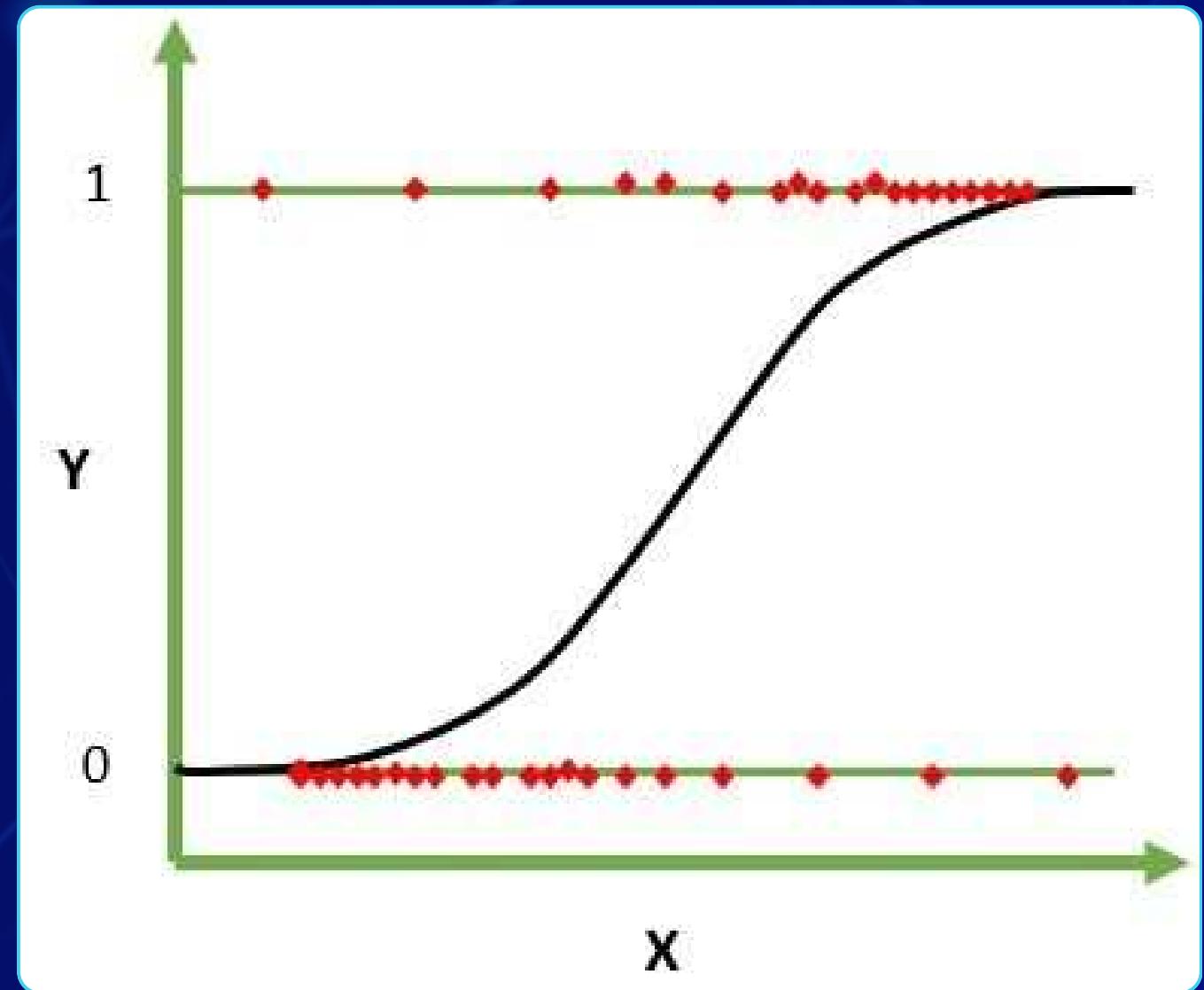
$$p = \frac{e^{(\beta_0 + \beta_1 x)} \div e^{(\beta_0 + \beta_1 x)}}{(1 + e^{(\beta_0 + \beta_1 x)}) \div e^{(\beta_0 + \beta_1 x)}}$$

# Función Sísmoide

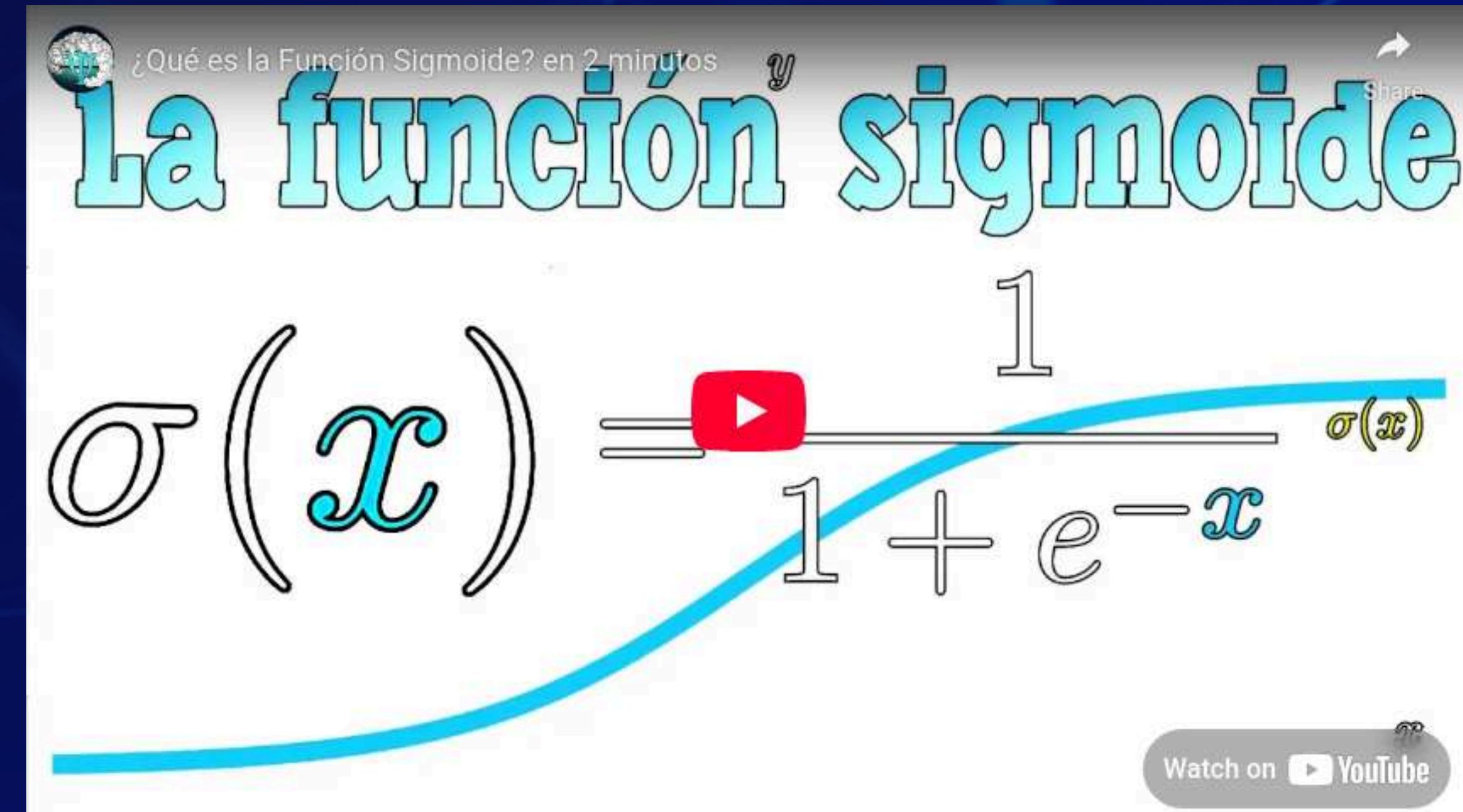
Ecuación:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Siendo  $p$  la probabilidad de que pertenecer a la clase 1 y el valor de la ecuación de la regresión lineal en  $x$ .



- Mostrar que:
  - Si  $x \rightarrow +\infty$ ,  $P(y=1) \rightarrow 1$ .
  - Si  $x \rightarrow -\infty$ ,  $P(y=1) \rightarrow 0$ .
  - En  $x=0$ ,  $P(y=1)=0.5$  (punto de decisión).



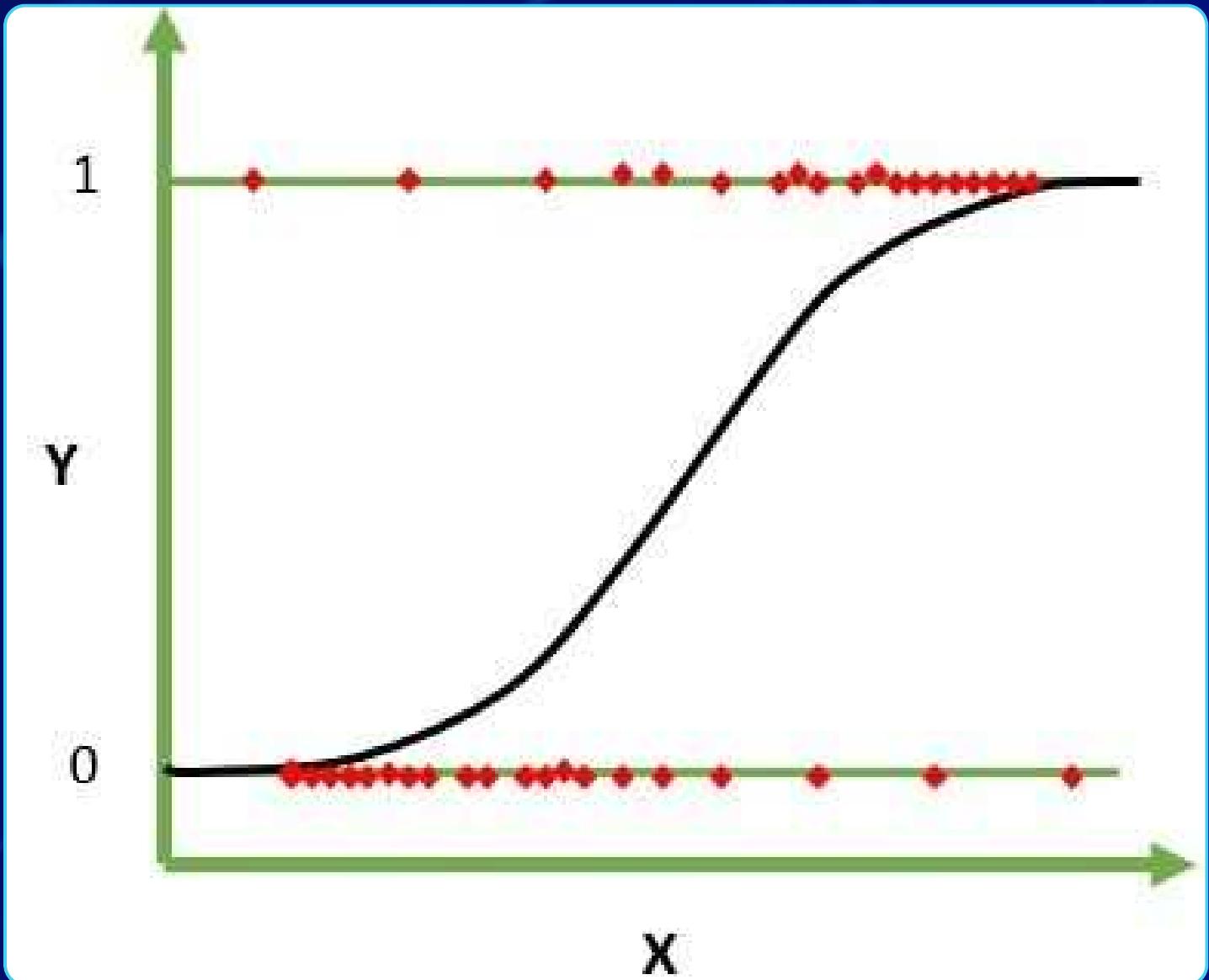
# Función de Costo

Ecuación:

$$J(\beta) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{y}_i)$$

Donde:

- $y_i$  es la etiqueta real del ejemplo (0 o 1).
- $\hat{y}_i$  es la predicción del modelo para ese ejemplo (una probabilidad entre 0 y 1).



$L(y_i, \hat{y}_i)$  Es la función de pérdida.

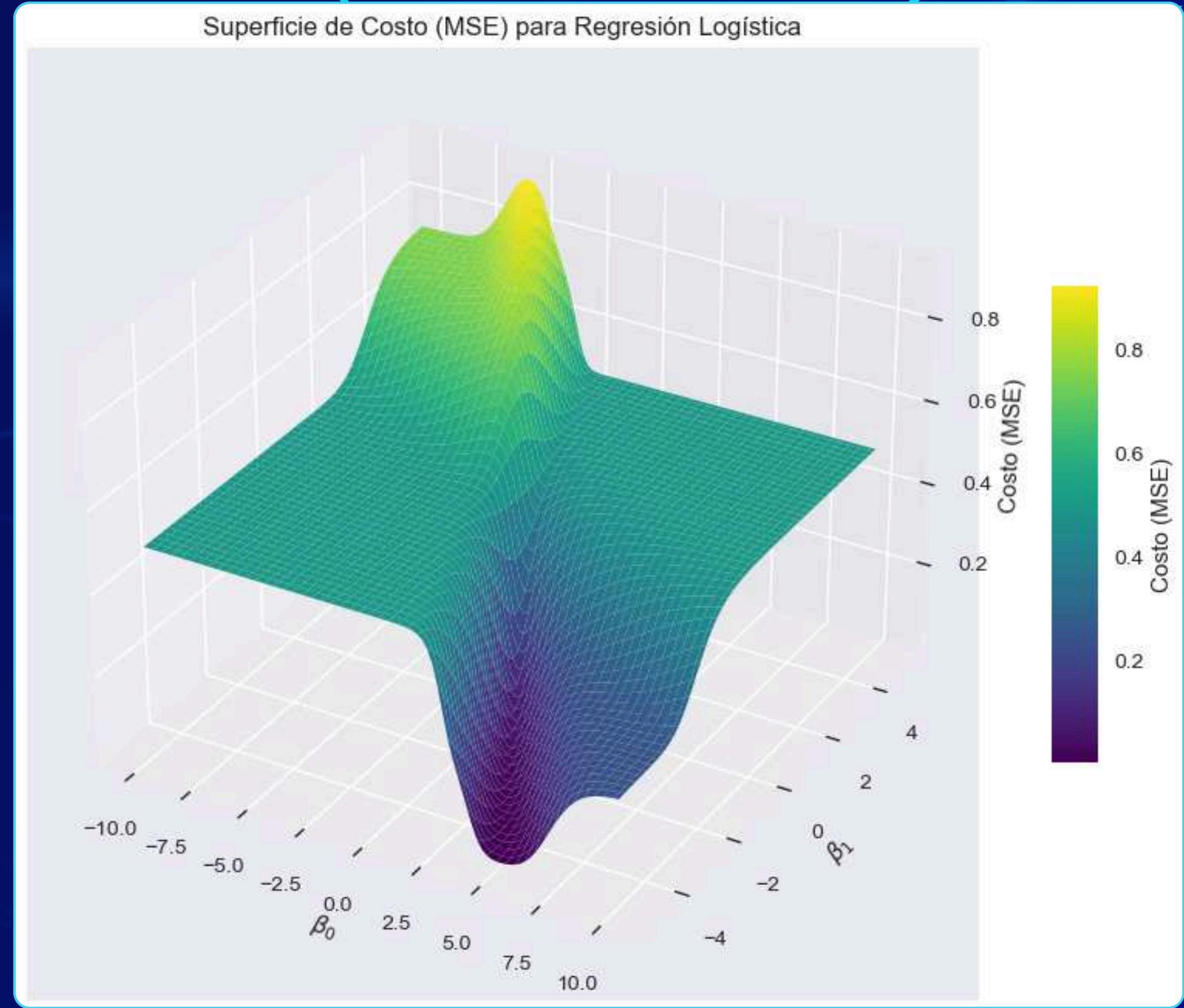
# Función de Costo

Buscar una función de pérdida adecuada:

$$L(y_i, \hat{y}_i)$$

No podemos usar las funciones de Regresión Lineal, porque:

- La superficie tiene múltiples valles y crestas, lo que indica que existen varios mínimos locales (puntos bajos donde el algoritmo podría estancarse).



El diagrama muestra claramente que el MSE es inadecuado para regresión logística debido a su no convexidad. La Entropía Cruzada es la alternativa óptima, ya que garantiza encontrar la mejor solución.

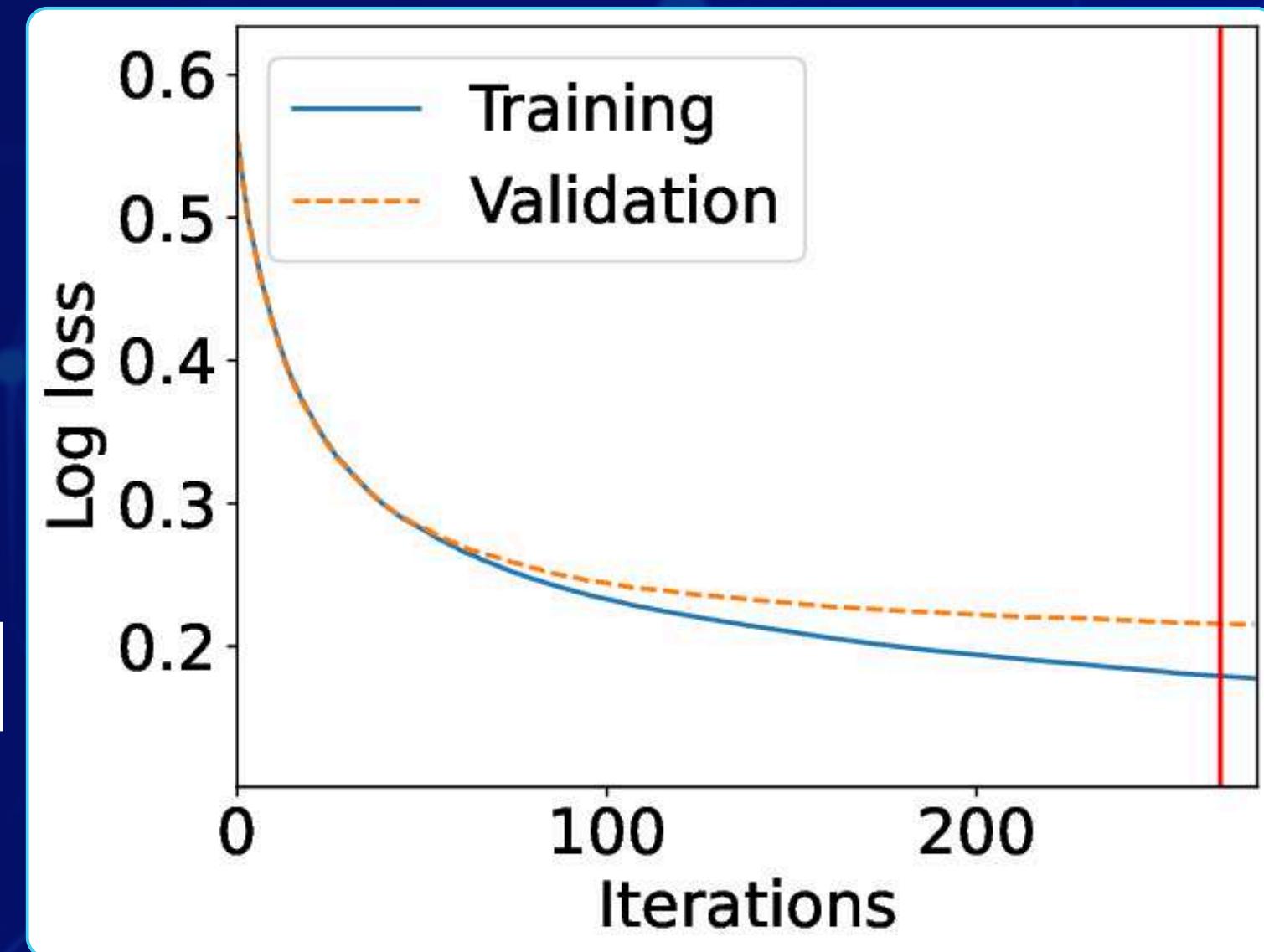
# Función de Pérdida

**Log-Loss:**

$$L = -[y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]$$

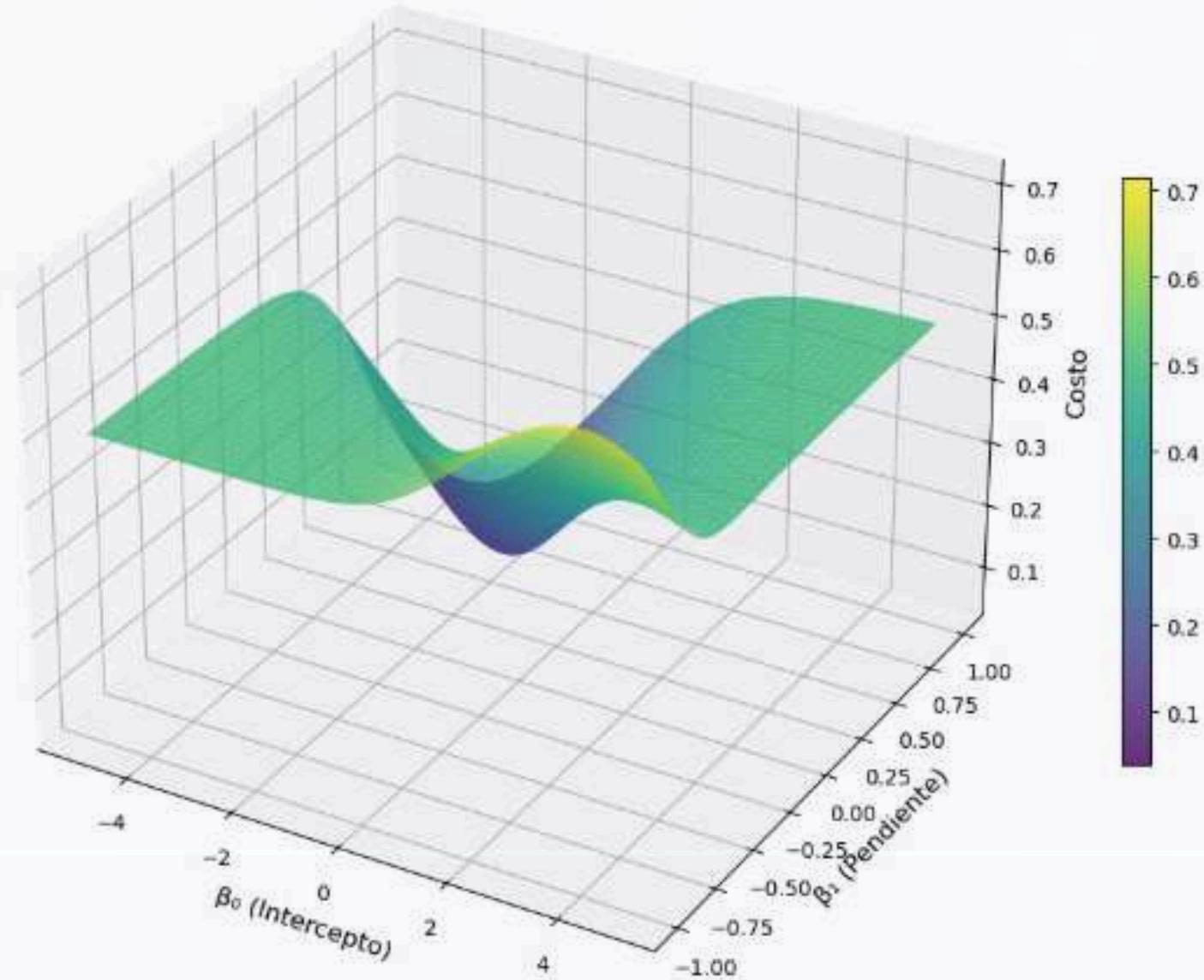
donde:

- $y_i \in \{0,1\}$ : Etiqueta real.
- $p_i$ : Probabilidad predicha de que  $y_i=1$  (salida de la sigmoide).
- Mide el error para una sola observación.
- En regresión logística, se conoce como entropía cruzada binaria (binary cross-entropy).

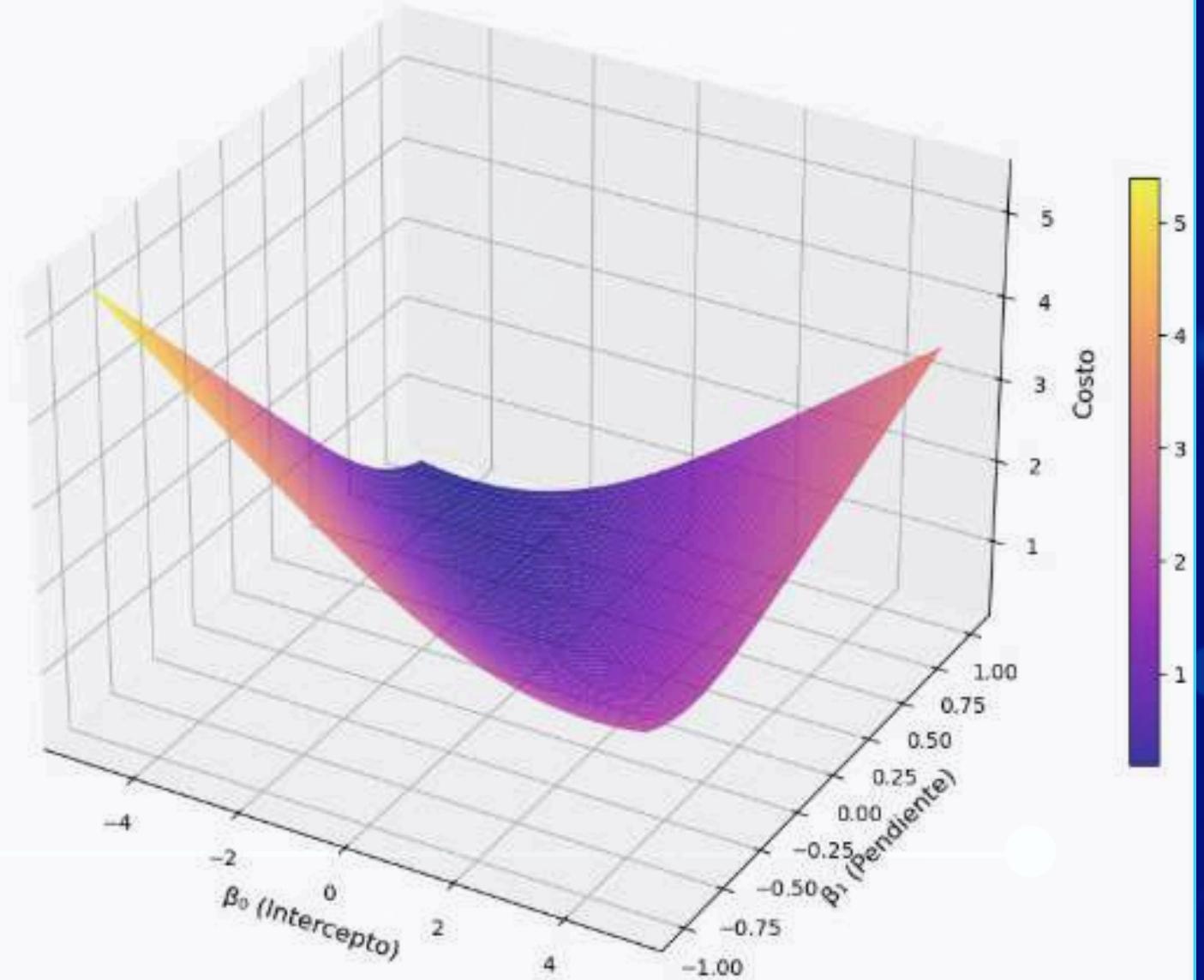


# Función de Pérdida

Función de Costo: MSE (No convexa)



Función de Costo: Entropía Cruzada (Convexa)



- A diferencia de la Entropía Cruzada (que tiene forma de tazón), el MSE no es convexo. Esto dificulta encontrar el óptimo global.

# Explicación Log-Loss

Probabilidad de  $y_i=1$

$$P(y_i = 1 | X_i) = p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})}}$$

Probabilidad de  $y_i=0$

$$P(y_i = 0 | X_i) = 1 - p_i$$

# Explicación Log-Loss

## **Maximum Likelihood Estimation:**

Es un método estadístico utilizado para encontrar los parámetros de un modelo que maximizan la probabilidad de que se observen los datos dados.

La idea es encontrar los valores de  $\beta$  que maximicen la probabilidad de observar los datos que tenemos:

$$L(\beta) = \prod_{i=1}^N P(y_i | X_i)$$

$$P(y_i | X_i) = \begin{cases} p_i & \text{si } y_i = 1, \\ 1 - p_i & \text{si } y_i = 0, \end{cases}$$

# Explicación Log-Loss

## Maximum Likelihood Estimation:

Para evitar escribir casos separados, usamos trucos matemáticos con exponentes:

$$P(y_i) = p_i^{y_i} (1 - p_i)^{1-y_i}$$

Cómo funciona:

- Si  $y_i=1$ :

$$p_i^1 (1 - p_i)^0 = p_i \times 1 = p_i$$

- Si  $y_i=0$ :

$$p_i^0 (1 - p_i)^1 = 1 \times (1 - p_i) = 1 - p_i$$

Ventaja:

- Compacta la expresión para todas las observaciones en un solo producto.

# Explicación Log-Loss

## Maximum Likelihood Estimation:

La idea es encontrar los valores de  $\beta$  que maximicen la probabilidad de observar los datos que tenemos:

$$L(\beta) = \prod_{i=1}^N (P(y_i = 1|X_i)^{y_i} \times P(y_i = 0|X_i)^{1-y_i}) = \prod_{i=1}^N p_i^{y_i} (1 - p_i)^{1-y_i}$$

Forma resumida y equivalente, usando  
propiedades de las potencias

# Optimización numérica (gradiente)

**Log-verosimilitud:**

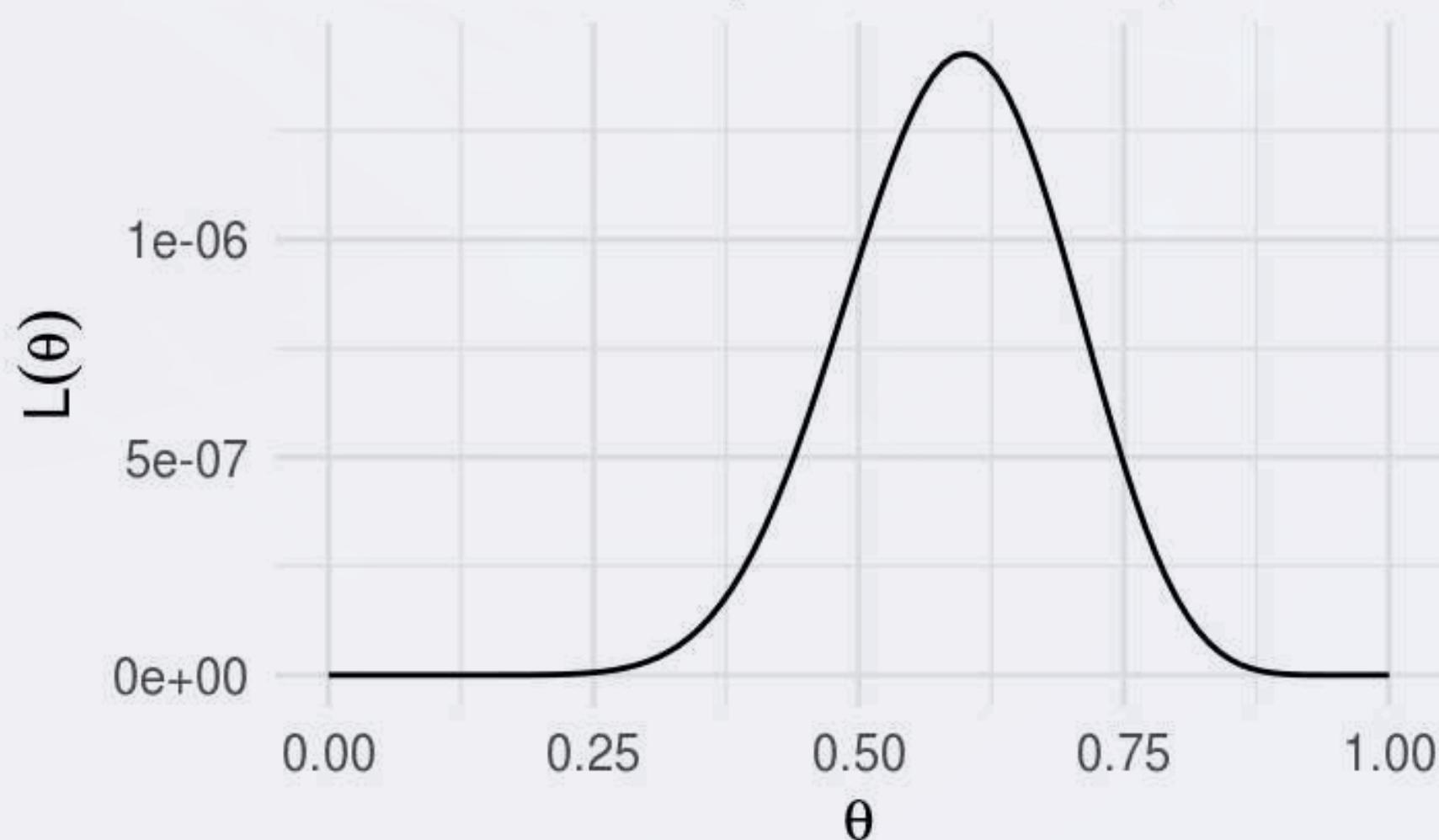
donde  $\ell(\beta)$  representa la log-verosimilitud  $\ell(\beta) = \ln(L(\beta))$

$$\ell(\beta) = \sum_{i=1}^N [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]$$

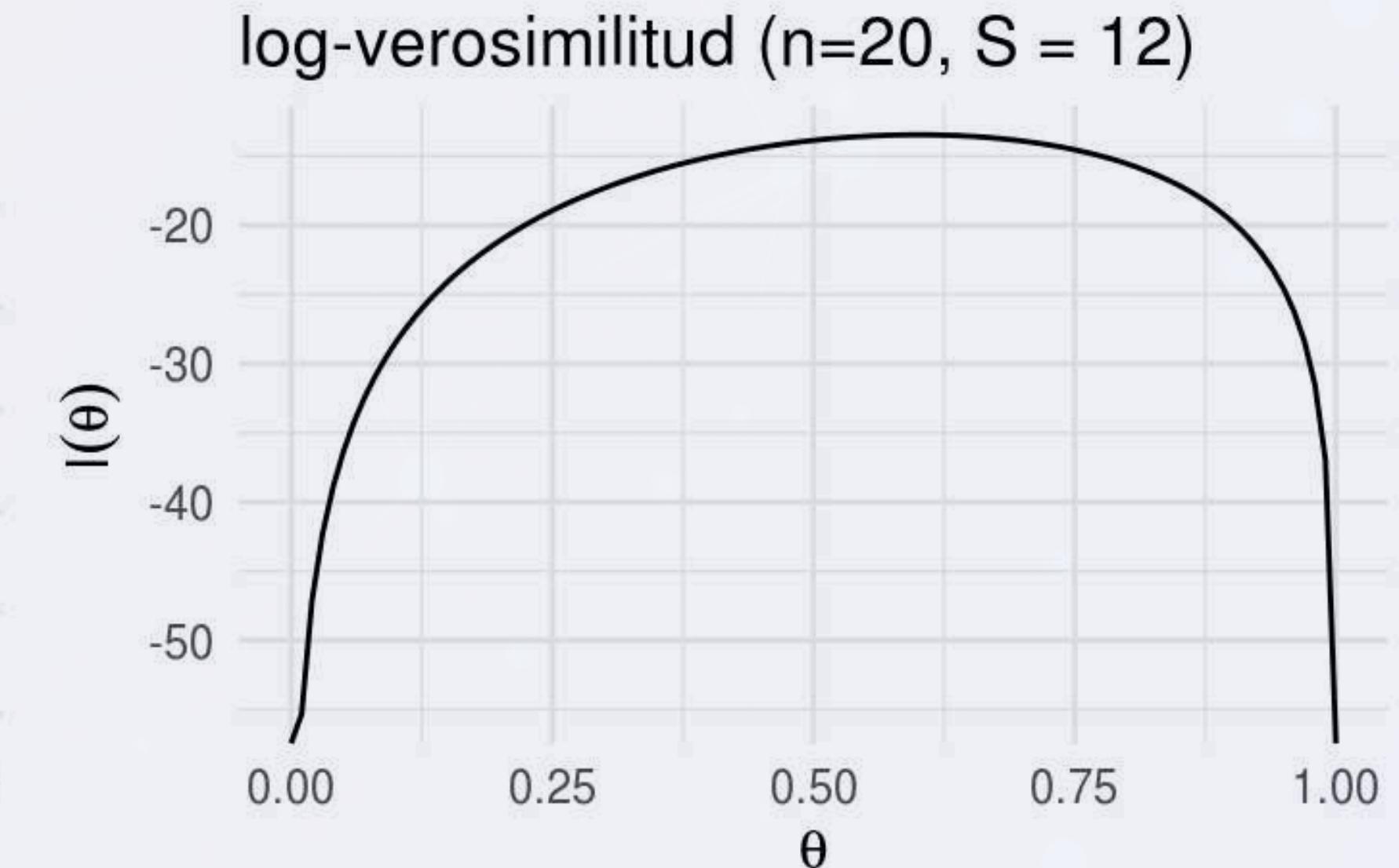
Trabajar con el logaritmo de la verosimilitud simplifica los cálculos (los productos se convierten en sumas)

# Optimización (gradiente)

Verosimilitud ( $n=20, S = 12$ )



log-verosimilitud ( $n=20, S = 12$ )



# Función de Costo

El objetivo es minimizar  $J(\beta)$ , lo que equivale a ajustar los parámetros  $\beta_0, \beta_1, \dots$  para hacer las predicciones lo más cercanas posible a las etiquetas reales  $y_i$ .

- Se basa en la log-loss, que penaliza predicciones incorrectas con alta confianza.
- Es diferenciable y adecuada para optimización mediante gradiente descendente.
- Nos permite ajustar los parámetros  $\beta_0, \beta_1$  para mejorar el desempeño del modelo.

$$J(\beta) = -\frac{1}{N} \sum_{i=1}^N [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]$$

# Estimación de los Coeficientes $\beta$

## Log-verosimilitud:

La idea es encontrar los valores de  $\beta$  que maximicen la probabilidad de observar los datos que tenemos:

$$\ell(\beta) = \sum_{i=1}^N [y_i \ln \left( \frac{1}{1 + e^{-z_i}} \right) + (1 - y_i) \ln \left( \frac{1}{1 + e^{-z_i}} \right)]$$

Donde:

$$z_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$$

Se despeja la función de Log-Verosimilitud para los *Betas* y se derivan las funciones

# Estimación de los Coeficientes $\beta$

## Gradiente Descendente: Actualización de Coeficientes

El gradiente descendente actualiza los coeficientes iterativamente:

$$\beta_j := \beta_j - \alpha \cdot \frac{\partial J}{\partial \beta_j}$$

donde:

- $\alpha$ : Tasa de aprendizaje (en este ejemplo se usará  $\alpha=0.1$ ).
- $\frac{\partial J}{\partial \beta_j}$ : Derivada parcial de la función de costo respecto a  $\beta_j$ .

$$\frac{\partial J}{\partial \beta_0} = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)$$

$$\frac{\partial J}{\partial \beta_1} = \frac{1}{N} \sum_{i=1}^N (p_i - y_i) \cdot x_i$$

# Estimación de los Coeficientes $\beta$

**Aplicar método del Gradiente Descendente:**

1. Inicializar  $\beta$  con valores aleatorios.
2. Calcular el gradiente  $\frac{\partial J}{\partial \beta_j}$
3. Actualizar  $\beta$  en dirección del gradiente

$$\beta_j := \beta_j - \alpha \cdot \frac{\partial J}{\partial \beta_j}$$

El algoritmo se detiene hasta que:

- El valor de  $J$  es suficientemente pequeño.
- Se alcanzan las iteraciones máximas permitidas.



# Regularización





# Regresión Logística: Ejemplo Paso a Paso

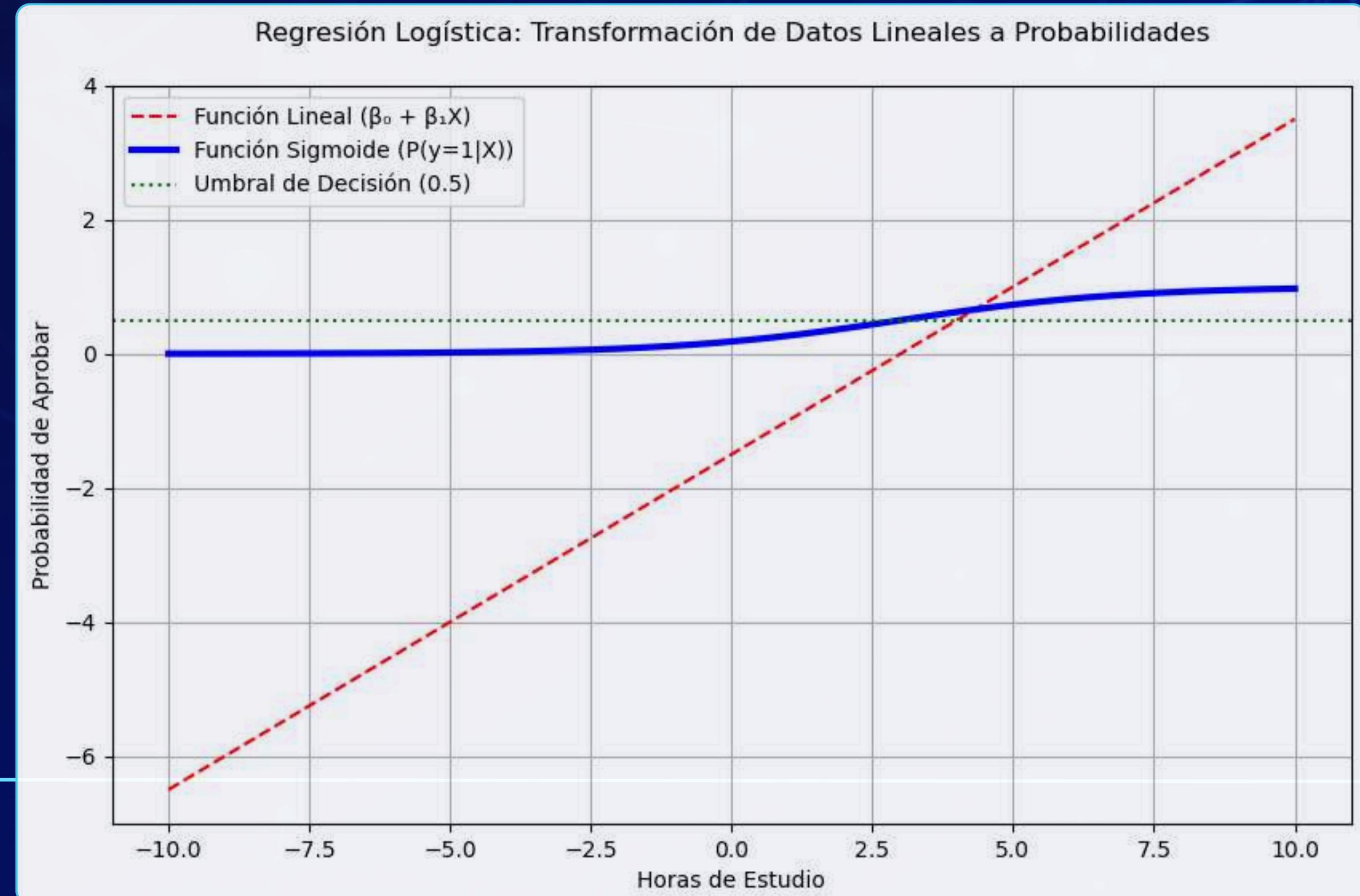
# Regresión Logística: Ejemplo Paso a Paso

## Clasificación de Alumnos Aprobados o Reprobados

- Queremos predecir si un estudiante aprueba ( $y=1$ ) o repreeba ( $y=0$ ) un examen.
- Basamos la predicción en una característica: Horas de estudio (x) .
- Usaremos un modelo de regresión logística simple.

Horas de Estudio (X)	Aprobado (y)
1	0
3	0
5	1
7	1

# Regresión Logística: Ejemplo Paso a Paso



- Muestra cómo una relación lineal se transforma en probabilidades mediante la sigmoide (Debido a esa transformación, se tiene el nombre de Regresión).
- Es complicado modelar el problema con Regresión Lineal, por eso se usa Regresión Logística.
- El umbral de decisión (0.5) separa las clases.

# Regresión Logística: Ejemplo Paso a Paso

## Modelo de Regresión Logística

Fórmula general:

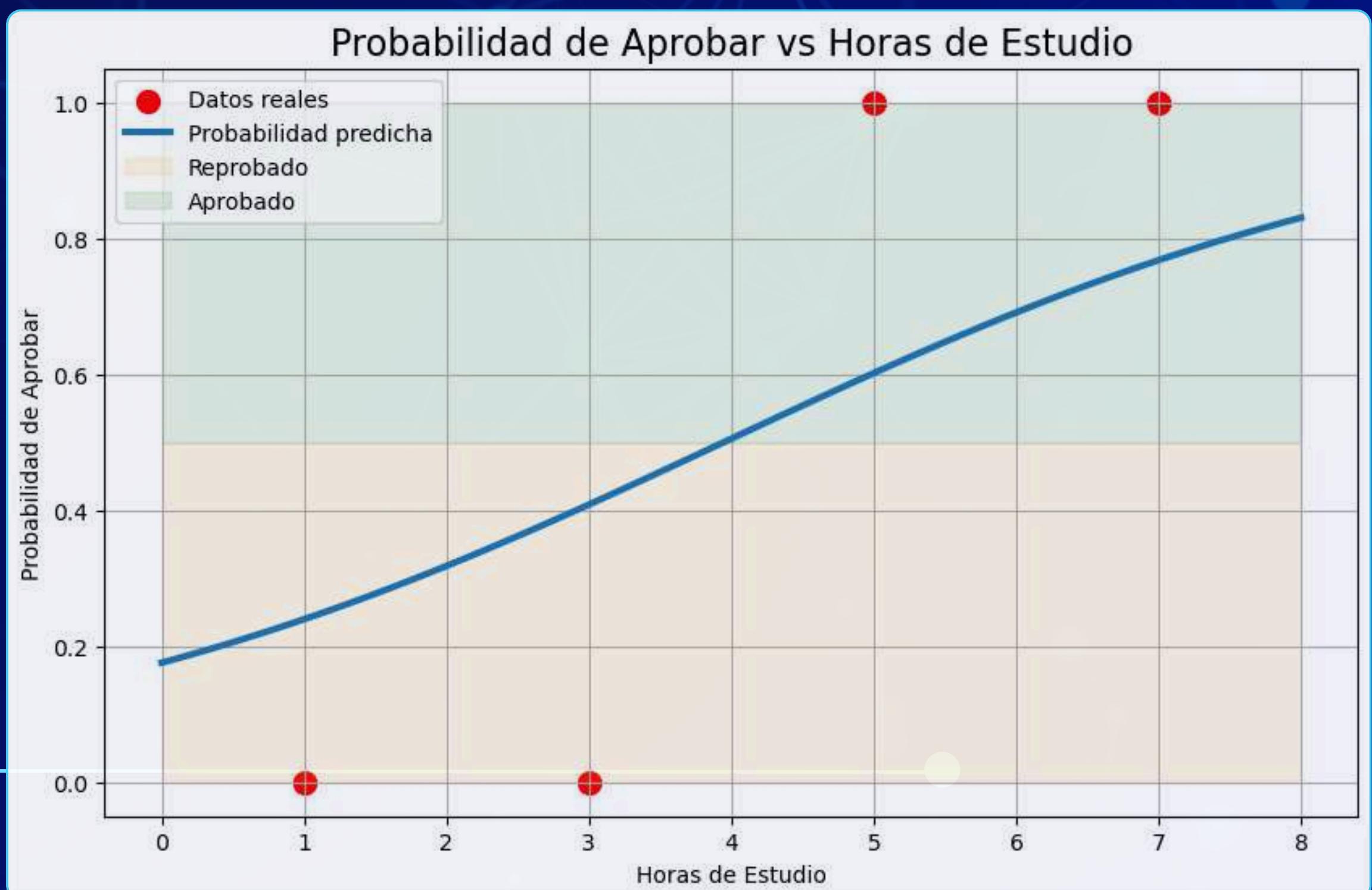
$$P(y_i = 1|X_i) = \frac{1}{1 + e^{-z}}$$

donde:

$$z = \beta_0 + \beta_1 X$$

# Regresión Logística: Ejemplo Paso a Paso

- Encontrar un modelo que asigne probabilidades a nuevos datos.
- Las áreas sombreadas representan las zonas de decisión.



# Regresión Logística: Ejemplo Paso a Paso

## Función de Costo

$$J(\beta_0, \beta_1) = -\frac{1}{4} \sum_{i=1}^4 [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]$$

Usando la ecuación Sísmoide:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}}$$

Calcular  $p_i$  para cada dato:

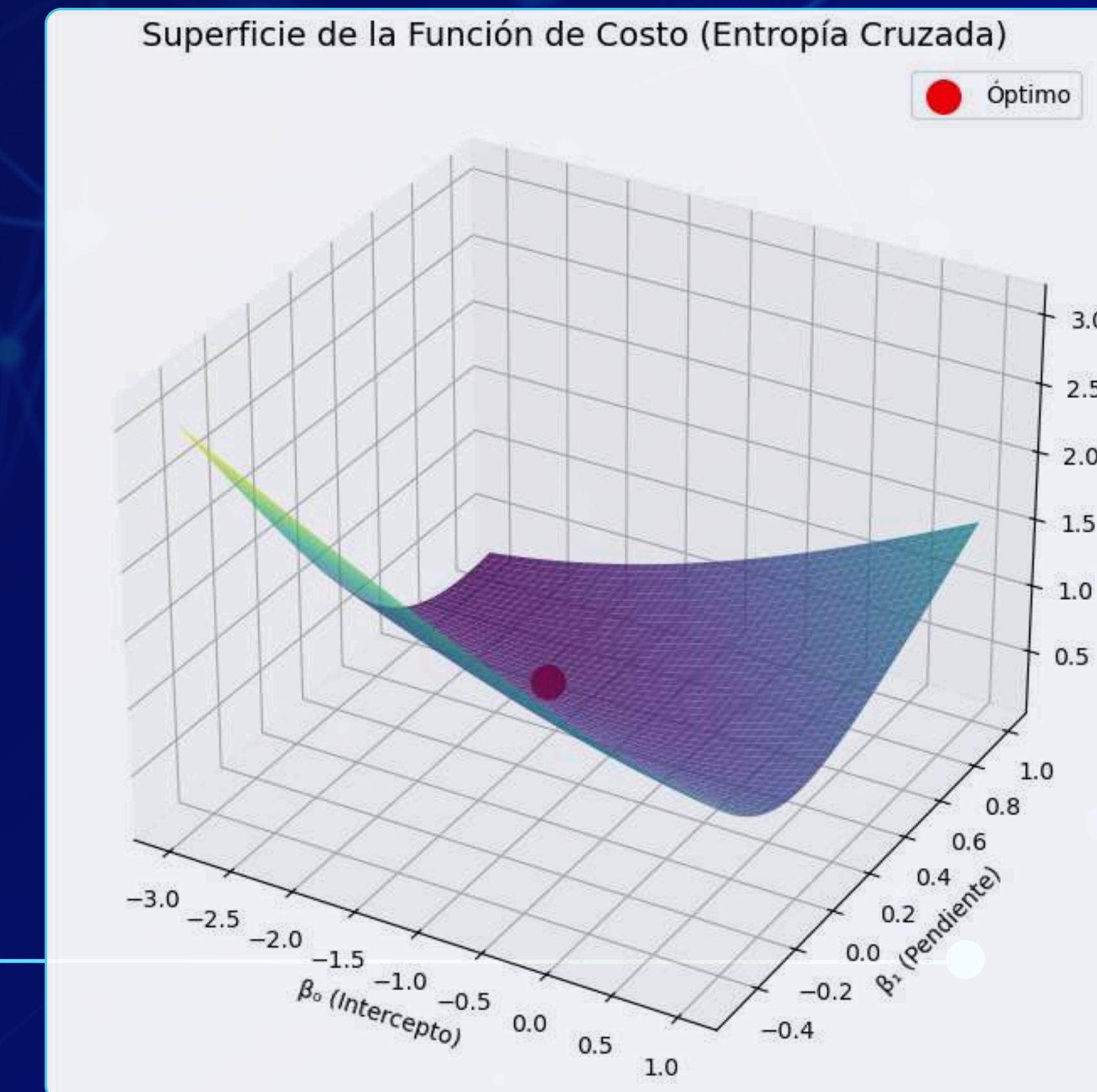
Con  $\beta_0=0$  y  $\beta_1=0$ :

$$p_i = \frac{1}{1 + e^{-(0+0X_i)}}$$

# Regresión Logística: Ejemplo Paso a Paso

## Función de Costo

- Visualiza cómo el costo varía con diferentes valores de  $\beta_0$  y  $\beta_1$ .
- El punto rojo marca el mínimo encontrado por el modelo.



# Regresión Logística: Ejemplo Paso a Paso

## Iteración #1

### Iteraciones del Gradiente Descendente

Calcular las derivadas parciales:

- Para  $\beta_0$ :

$$\frac{\partial J}{\partial \beta_0} = \frac{1}{4} \sum_{i=1}^4 (p_i - y_i)$$

Sustituyendo:

$$\frac{\partial J}{\partial \beta_0} = \frac{1}{4} [(0.5 - 0) + (0.5 - 0) + (0.5 - 1) + (0.5 - 1)] = 0$$

Horas de Estudio (X)	$p_i$	Aprobado (y)
1	0.5	0
3	0.5	0
5	0.5	1
7	0.5	1

# Regresión Logística: Ejemplo Paso a Paso

## Iteración #1

### Iteraciones del Gradiente Descendente

Calcular las derivadas parciales:

- Para  $\beta_1$ :

$$\frac{\partial J}{\partial \beta_1} = \frac{1}{4} \sum_{i=1}^4 (p_i - y_i) \cdot x_i$$

Sustituimos:

$$\frac{\partial J}{\partial \beta_1} = \frac{1}{4} [(0.5 - 0)(1) + (0.5 - 0)(3) + (0.5 - 1)(5) + (0.5 - 1)(7)] = -1.0$$

Horas de Estudio (X)	$p_i$	Aprobado (y)
1	0.5	0
3	0.5	0
5	0.5	1
7	0.5	1

# Regresión Logística: Ejemplo Paso a Paso

Actualizar coeficientes (con tasa de aprendizaje  $\alpha=0.1$ )

$$\begin{aligned}\beta_0^{nuevo} &= 0 - 0.1 \cdot 0 = 0 \\ \beta_1^{nuevo} &= 0 - 0.1 \cdot (-1.0) = 0.1\end{aligned}$$

Calcular el Costo (Usar Función de Entropía Cruzada/Log-Loss)

$$J(\beta) = -\frac{1}{4} \sum_{i=1}^4 [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)] \approx 0.693\dots$$

Resultados después de Iteración 1:

$$\beta_0=0, \beta_1=0.1, J=0.6931$$

# Regresión Logística: Ejemplo Paso a Paso

Iteración #2

Calcular nuevas  $p_i$

$$p_i = \frac{1}{1 + e^{-(0+0.1X_i)}}$$

Horas de Estudio (X)	$p_i$	Aprobado (y)
1	0.5249...	0
3	0.5744...	0
5	0.6224...	1
7	0.6681...	1

# Regresión Logística: Ejemplo Paso a Paso

## Iteración #2

### Iteraciones del Gradiente Descendente

Calcular las derivadas parciales:

- Para  $\beta_0$ :

$$\frac{\partial J}{\partial \beta_0} = \frac{1}{4} \sum_{i=1}^4 (p_i - y_i)$$

Sustituyendo:

$$\frac{\partial J}{\partial \beta_0} = \frac{1}{4} [(0.525 - 0) + (0.574 - 0) + (0.622 - 1) + (0.668 - 1)] \approx -0.09725$$

Horas de Estudio (X)	$p_i$	Aprobado (y)
1	0.5249...	0
3	0.5744...	0
5	0.6224...	1
7	0.6681...	1

# Regresión Logística: Ejemplo Paso a Paso

## Iteración #2

### Iteraciones del Gradiente Descendente

Calcular las derivadas parciales:

- Para  $\beta_1$ :

$$\frac{\partial J}{\partial \beta_1} = \frac{1}{4} \sum_{i=1}^4 (p_i - y_i) \cdot x_i$$

Sustituimos:

$$\frac{\partial J}{\partial \beta_1} = \frac{1}{4} [(0.525 - 0)(1) + (0.574 - 0)(3) + (0.622 - 1)(5) + (0.668 - 1)(7)] \approx -0.49175$$

Horas de Estudio (X)	$p_i$	Aprobado (y)
1	0.5249...	0
3	0.5744...	0
5	0.6224...	1
7	0.6681...	1

# Regresión Logística: Ejemplo Paso a Paso

Actualizar coeficientes (con tasa de aprendizaje  $\alpha=0.1$ )

$$\beta_0^{nuevo} = 0 - 0.1 \cdot (0.09725) \approx -0.0097$$
$$\beta_1^{nuevo} = 0.1 - 0.1 \cdot (-0.49175) \approx 0.1492$$

Calcular el Costo (Usar Función de Entropía Cruzada/Log-Loss)

$$J(\beta) = -\frac{1}{4} \sum_{i=1}^4 [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i))] \approx 0.619\dots$$

Resultados después de Iteración 2:

$$\beta_0=-0.0097, \beta_1=0.1492, J=0.6190$$

# Regresión Logística: Ejemplo Paso a Paso

Iteración #3

Calcular nuevas  $p_i$

$$p_i = \frac{1}{1 + e^{-( -0.0097 + 0.1492X_i)}}$$

Horas de Estudio (X)	$p_i$	Aprobado (y)
1	0.5348...	0
3	0.6077...	0
5	0.6761...	1
7	0.7378...	1

# Regresión Logística: Ejemplo Paso a Paso

## Iteración #3

### Iteraciones del Gradiente Descendente

Calcular las derivadas parciales:

- Para  $\beta_0$ :

$$\frac{\partial J}{\partial \beta_0} = \frac{1}{4} \sum_{i=1}^4 (p_i - y_i)$$

Horas de Estudio (X)	$p_i$	Aprobado (y)
1	0.5348...	0
3	0.6077...	0
5	0.6761...	1
7	0.7378...	1

Sustituyendo:

$$\frac{\partial J}{\partial \beta_0} = \frac{1}{4} [(0.534 - 0) + (0.607 - 0) + (0.676 - 1) + (0.737 - 1)] \approx 0.1385$$

# Regresión Logística: Ejemplo Paso a Paso

## Iteración #3

### Iteraciones del Gradiente Descendente

Calcular las derivadas parciales:

- Para  $\beta_1$ :

$$\frac{\partial J}{\partial \beta_1} = \frac{1}{4} \sum_{i=1}^4 (p_i - y_i) \cdot x_i$$

Sustituimos:

$$\frac{\partial J}{\partial \beta_1} = \frac{1}{4} [(0.534 - 0)(1) + (0.607 - 0)(3) + (0.676 - 1)(5) + (0.737 - 1)(7)] \approx -0.2765$$

Horas de Estudio (X)	$p_i$	Aprobado (y)
1	0.5348...	0
3	0.6077...	0
5	0.6761...	1
7	0.7378...	1

# Regresión Logística: Ejemplo Paso a Paso

Actualizar coeficientes (con tasa de aprendizaje  $\alpha=0.1$ )

$$\beta_0^{nuevo} = -0.0097 - 0.1 \cdot (0.1385) \approx 0.0236$$
$$\beta_1^{nuevo} = 0.1492 - 0.1 \cdot (-0.2765) \approx 0.176$$

Calcular el Costo (Usar Función de Entropía Cruzada/Log-Loss)

$$J(\beta) = -\frac{1}{4} \sum_{i=1}^4 [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)] \approx 0.599\dots$$

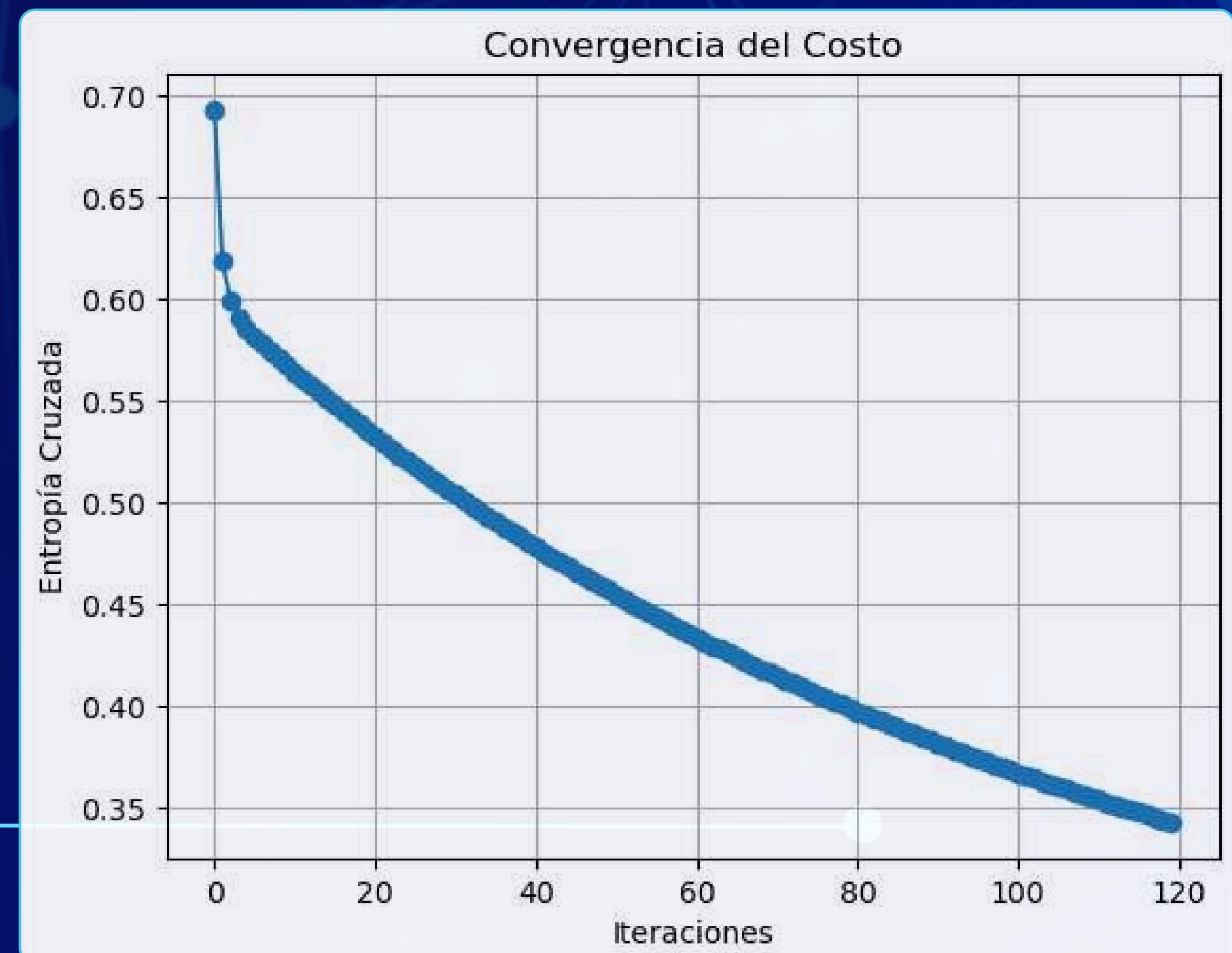
Resultados después de Iteración 3:

$$\beta_0=0.0236, \beta_1=0.1769, J=0.5992$$

# Regresión Logística: Ejemplo Paso a Paso

## Función de Costo

- Visualiza cómo el costo varía va disminuyendo en cada iteración
- En las primeras iteraciones, la entropía cruzada disminuye rápidamente, indicando que el modelo está aprendiendo de manera eficiente y mejorando sus predicciones.
- Después de muchas iteraciones (en este caso, alrededor de 120), el costo se estabiliza alrededor de un valor cercano a 0.35. Esto ocurre porque el modelo ha encontrado una configuración de parámetros que minimiza la entropía cruzada en el conjunto de datos de entrenamiento.



# Regresión Logística: Ejemplo Paso a Paso

Iteración #100

Calcular nuevas  $p_i$

$$p_i = \frac{1}{1 + e^{-( -1.4266 + 0.5097X_i)}}$$

Resultados después de Iteración 100:

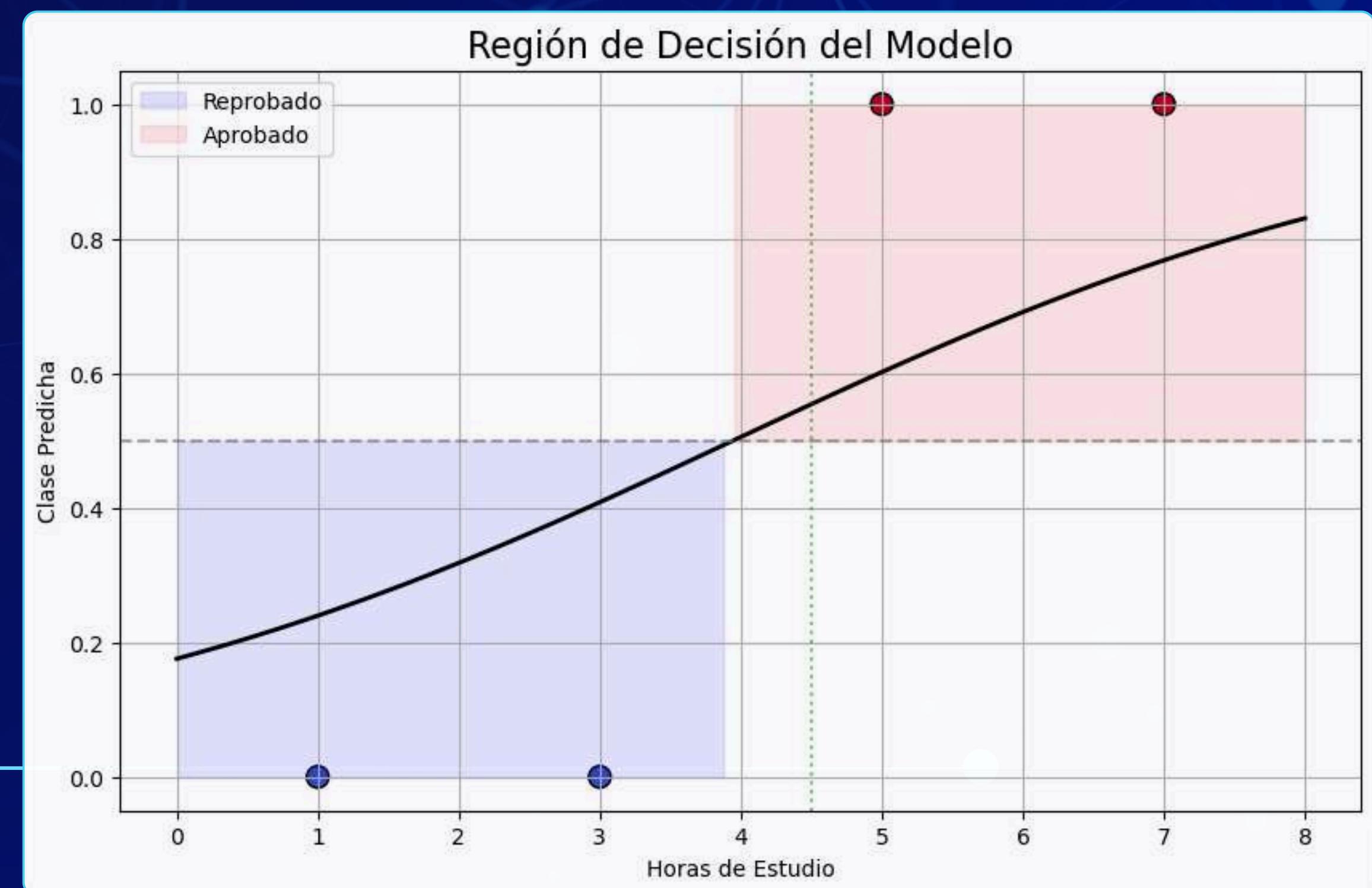
$$\beta_0 = -1.4381, \beta_1 = 0.5123, J = 0.3688$$

Horas de Estudio (X)	$p_i$	Aprobado (y)
1	0.2855...	0
3	0.5256...	0
5	0.7543...	1
7	0.8948...	1

# Regresión Logística: Ejemplo Paso a Paso

## Función de Costo

- Muestra cómo el modelo divide el espacio en zonas de clasificación.
- El punto de cruce ( $\approx 4.5$  horas) es donde  $P(y=1)=0.5$ .

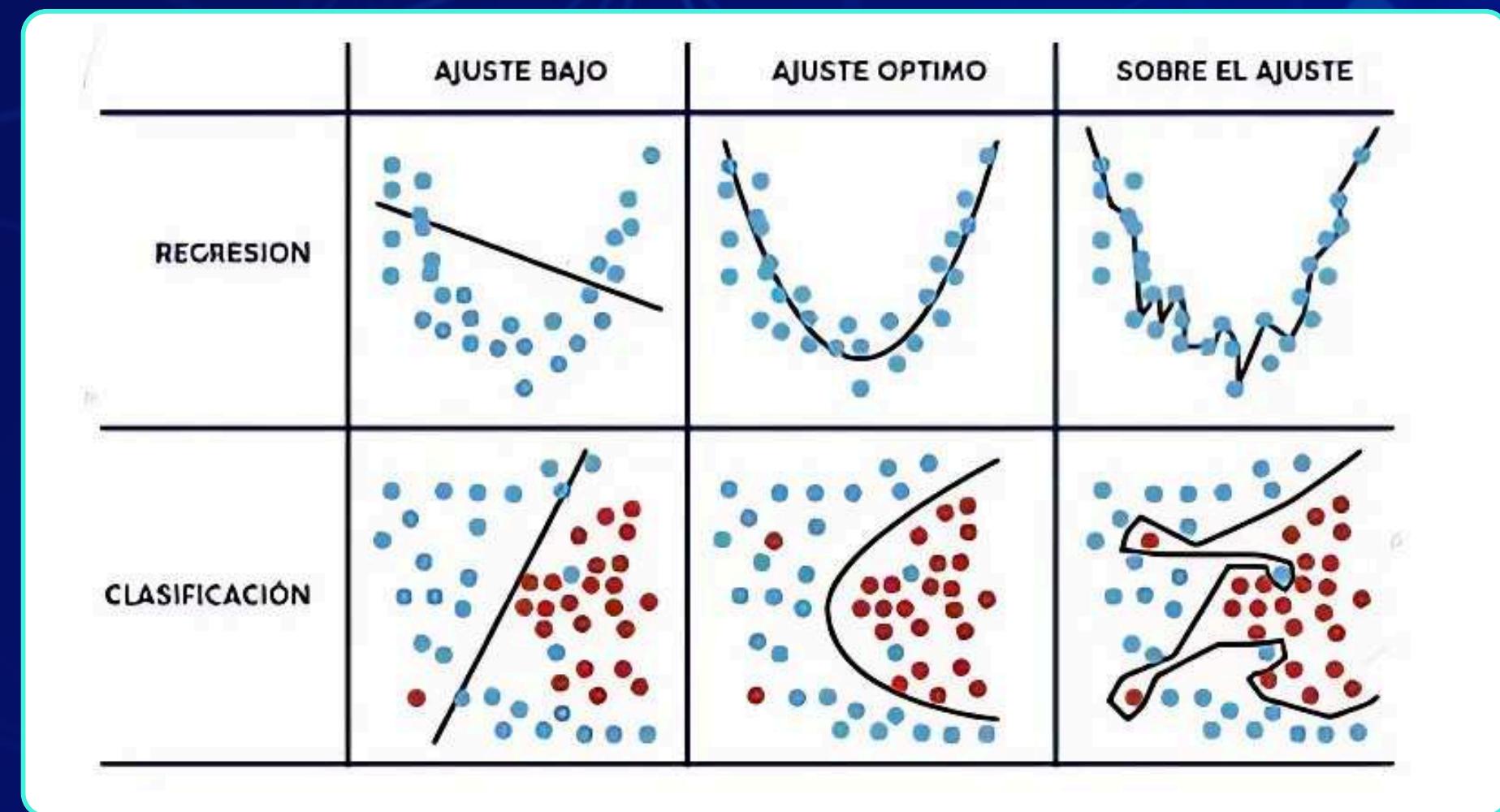




## Over-fitting vs Under-fitting

# Sobreajuste y subajuste

- El subajuste ocurre cuando el modelo presenta un sesgo muy alto y no puede capturar los patrones complejos de los datos. Esto genera mayores errores de entrenamiento y validación, ya que el modelo no es lo suficientemente complejo como para clasificar los datos subyacentes.
- El sobreajuste es lo opuesto, ya que el modelo es demasiado complejo (o superior) y captura incluso el ruido de los datos.





Evaluación

# Matriz de Confusión





# Entropía Cruzada

Estas métricas evalúan cómo bien el modelo predice las probabilidades de pertenecer a cada clase.

### Entropía Cruzada Binaria (Log Loss)

La entropía cruzada binaria mide la discrepancia entre las probabilidades predichas ( $P(y_i=1)$ ) y las etiquetas reales ( $y_i$ ).

- Menor es mejor: Cuanto más cercano sea el valor de  $L$  a cero, mejor ajusta el modelo.
- Ideal para probabilidades: Es la métrica preferida cuando el interés está en evaluar la calidad de las probabilidades predichas.



# Curva ROC y AUC

La curva ROC es una herramienta gráfica que mide la capacidad de un modelo para discriminar entre clases. Representa la relación entre:

- True Positive Rate (TPR) : La sensibilidad o recall. Es la proporción de casos positivos correctamente identificados.
- False Positive Rate (FPR) : La tasa de falsos positivos. Es la proporción de casos negativos incorrectamente identificados como positivos.

La curva ROC traza los valores de TPR contra FPR para diferentes umbrales de clasificación. Un buen modelo tendrá una curva que se acerca a la esquina superior izquierda del gráfico, donde:  
•  $TPR = 1$  : Todos los casos positivos están correctamente identificados.  
•  $FPR = 0$  : Ningún caso negativo está incorrectamente identificado.



# Curva ROC y AUC

- El AUC es el área bajo la curva ROC. Mide la capacidad del modelo para discriminar entre clases. Un AUC cercano a 1 indica un excelente modelo, mientras que un valor cercano a 0.5 indica que el modelo no discrimina mejor que el azar.

