



# Regresión Logística

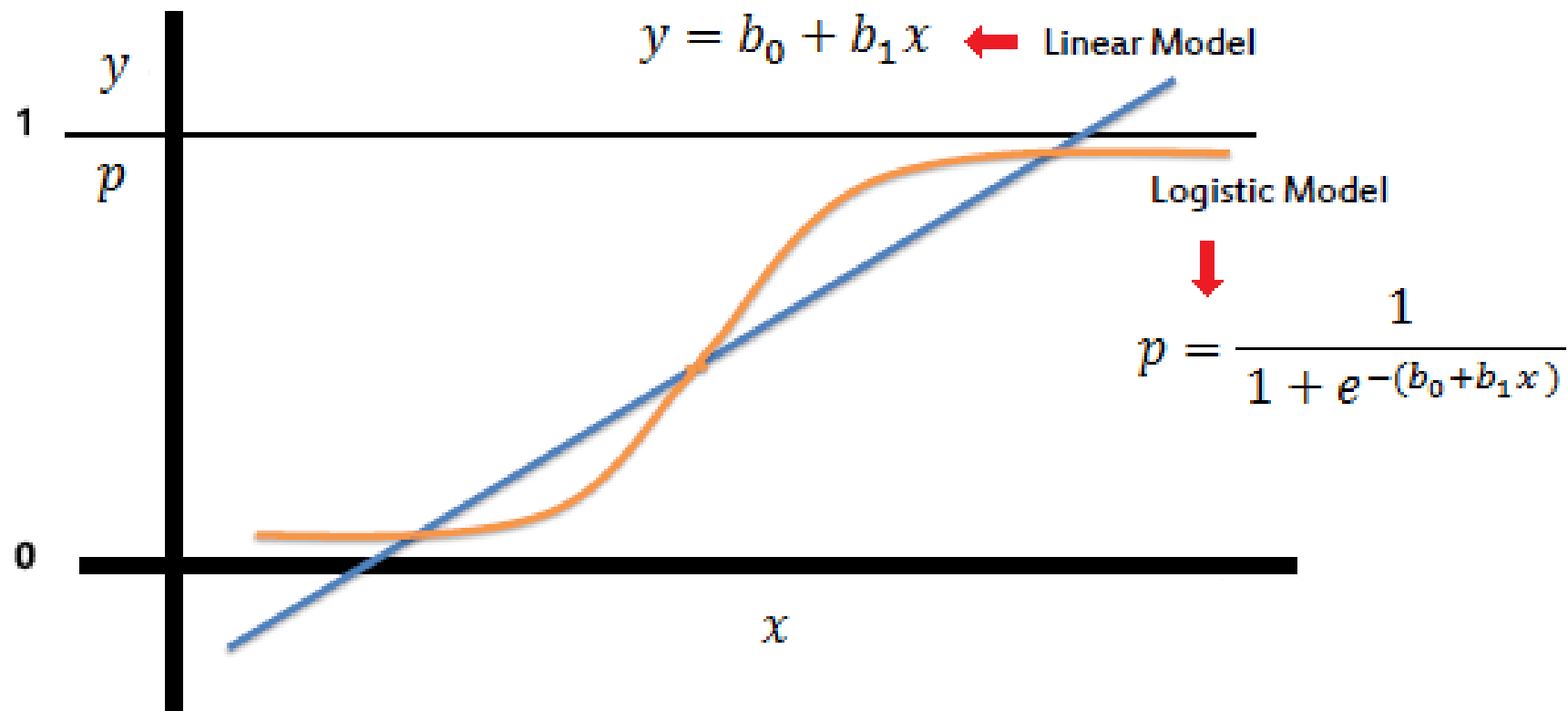


# Regresión Logística

A partir de un conjunto de variables escalares (o dummies) se busca predecir una respuesta tipo binomial

- ¿Qué tan probable es que se acepte una oferta?
- ¿Puedo acercarme a la predicción de una falla?





## Regresión lineal múltiple

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

## Modelo de Regresión Logística

$$Prob(Y = 1) = p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_1 X_1 + \cdots + \beta_1 X_1)}}$$

$$\text{Log} \left( \frac{p}{1 - p} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

# Odds

- Odd (ejemplo 1): Las apuestas están 4:1 -> Odd es 4
- Odd (ejemplo 2): Probabilidad de compra: 0.6

$$Odd = \frac{0.6}{0.4} = 1.5$$

# Odd Ratio

El odd ratio es una medida de asociación entre dos variables que indica la fortaleza de la relación entre dos variables

Ejemplo: qué tan probable es votar al partido liberal teniendo en cuenta si las personas son creyentes o no.

# Odd Ratio

No cree en Dios

Prob(SI vota Liberal) = 0.42  
Prob(NO vota Liberal) = 0.58

$$ODD = \frac{0.42}{0.58} = 0.72$$

$$ODD \text{ RATIO} = \frac{0.72}{0.33} = 2.17$$

	Creer en Dios	No creer en Dios	Total
Sí vota al partido liberal	100	250	350
No vota al partido liberal	300	350	650
Total	400	600	1000

Sí cree en Dios

Prob(SI vota Liberal) = 0.25  
Prob(NO vota Liberal) = 0.75

$$ODD = \frac{0.25}{0.75} = 0.33$$

Para una persona no creyente los *odds* de sí votar al partido liberal son 2,17 veces más grandes que los *odds* de una persona que sí cree en Dios, de sí votar al partido liberal.

# Odd Ratio

## ¿Cómo se interpretan los odd ratio?

- Los *odd ratio* oscilan entre 0 e infinito.
- Cuando el *odd ratio* es 1 indica ausencia de asociación entre las variables.
- Los **valores menores de 1** señalan una **asociación negativa** entre las variables y los **valores mayores de 1** indican **asociación positiva** entre las variables.
- Cuanto más se aleje el *odd ratio* de 1, más fuerte es la relación



# Odd Ratio

Aumento porcentual del Odd



Si Odd ratio  $> 1$ : Se incrementa la probabilidad de lo que se definió como objetivo

Si Odd ratio  $< 1$ : Se disminuye la probabilidad de lo que se definió como objetivo



# Odd Ratio

## Ejemplo 1

Odd ratio=2, con Odd de 1.5 significa que la probabilidad aumentó de 60% a 75%.

$$Odd1 = \frac{0.6}{0.4} = 1.5$$

$$Odd2 = \frac{0.75}{0.25} = 3$$

# Odd Ratio

## Ejemplo 2

Odd ratio=2.25, pero Odd inicial de 0.11 significa que la probabilidad aumentó de 10% a 20%.

$$Odd1 = \frac{0.1}{0.9} = 0.11$$

$$Odd2 = \frac{0.2}{0.8} = 0.25$$

# Odd Ratio

## Ejemplo 3

Odd ratio = 0.9 con Odd inicial de 0.25 entonces la probabilidad disminuyó de 20% a 18.37%.

$$Odd\ 1 = \frac{0.2}{0.8} = 0.25$$

$$Odd\ 2 = \frac{0.1837}{0.8163} = 0.225$$

**Luego, el Odd ratio debe ser interpretado sobre la probabilidad base**

# Odd Ratio

- Los *odd ratios* son usados en modelos de **regresión logística** para comparar la influencia de las variables explicativas (o independientes) sobre la variable dependiente.
- Al realizar regresiones logísticas, los *odd ratios* se denominan **exponencial de  $b$**  y se expresan así:  **$\text{Exp}(b)$** .

# Odd Ratio

El odd ratio es una medida de asociación entre dos variables que indica la fortaleza de la relación entre dos variables

Ejemplo: qué tan probable es votar al partido liberal teniendo en cuenta si las personas creen en Dios o no.

Variable dependiente:

**Votar por el partido liberal: 1 (SI), 0 (NO)**

Variables independientes:

- Creer en Dios: 1(SI), 0 (NO)
- Edad: # años
- Nivel de Ingresos: Escala de 1 a 10.

# Interpretación de efectos en la regresión logística

## Resultados del procesamiento del modelo de RL

	Coeficiente b	Sig.	Exp(b)
Creer en Dios	-0,550	0,000	0,577
Edad	0,010	0,001	1,010
Ingresos	0,415	0,000	1,508
Constante	1,214	0,000	3,366

Variable dependiente: Votar al partido liberal

- Los exponenciales de b, **Exp(b)**, son *odd ratios* y pueden compararse entre sí para saber qué variable es más explicativa de la variable dependiente o está asociada de manera más fuerte.
- Cuando el **Exp(b)** es mayor de 1 señala que un aumento de la variable independiente, aumenta los *odds* que ocurra el evento (es decir, la variable dependiente).
- Cuando el **Exp(b)** es menor de 1 indica que un aumento de la variable independiente, reduce los *odds* que ocurra el evento (variable dependiente).

# Interpretación de efectos en la regresión logística

- Cuando el *odd ratio* es menor de 1 es conveniente calcular su inversa para poder comparar más fácilmente todos los  $\text{Exp}(b)$ .
- El aumento de una unidad de nivel de ingresos, y si el resto de variables se mantuvieran constantes, aumenta los odds de votar al partido liberal en 1,508 veces más que si no se aumentara esa unidad del nivel de ingresos.
- Si se aumenta la edad en una unidad, y todos los valores de las otras variables del modelo permanecen constantes, los odds de votar al partido liberal aumentan 1,01 veces más que los odds de si no se aumentara la edad en esa unidad.
- Los no creyentes tienen más odds de votar al partido liberal en 1,73 que los creyentes



# Calidad – Prueba de hipótesis para los parámetros

Esta prueba examina cada una de las variables independientes e indica si tienen o no significado en el modelo.

$H_0$ : La variable no tiene significado o no está asociada ( $\beta = 0$ )

$H_1$ : La variable tiene significado y está asociada ( $\beta \neq 0$ )

# Supuestos del modelo de Regresión Logística

**Linealidad del logaritmo del odd ratio:** los parámetros y su interpretación carecen de sentido si en realidad los datos no proceden de un modelo lineal de la función logarítmica

**Independencia de variables explicativas:** Las variables son linealmente independientes

# Métricas de evaluación en Clasificación



# Métricas de Clasificación



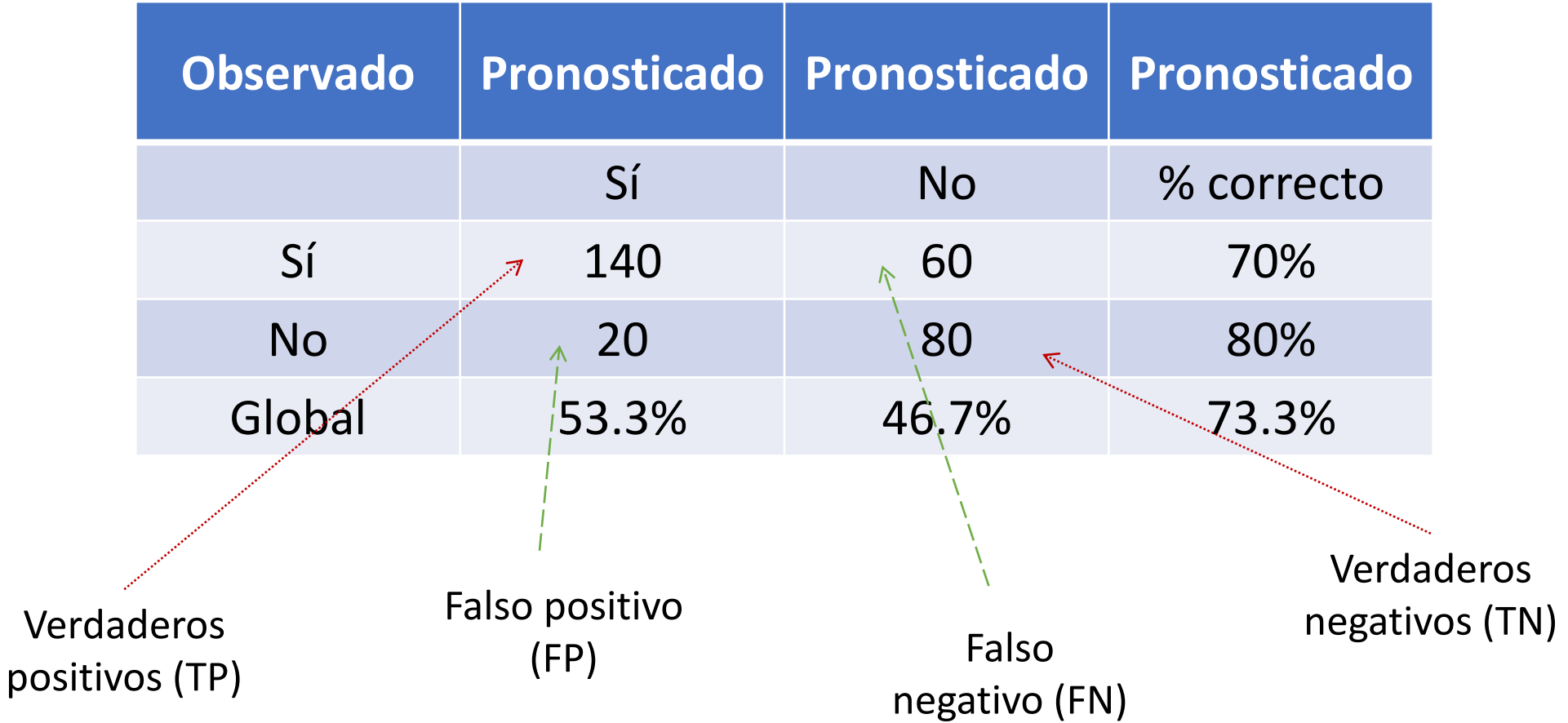
# Matriz de Confusión

Observado	Pronosticado		Pronosticado
	Sí	No	% correcto
Sí	140	60	70%
No	20	80	80%
Global	53.3%	46.7%	73.3%

Fallo

Acierto

# Matriz de Confusión



# Precisión y Exhaustividad

- **Precisión / Especificidad (precision):** Verdaderos positivos de todas las predicciones positivas.
- **Exhaustividad / Sensibilidad (recall):** Verdaderos Positivos de todos los que son realmente positivos.

# Precisión y Exhaustividad

Observado	Pronosticado	Pronosticado	Pronosticado
	Sí	No	% correcto
Sí	140	60	70%
No	20	80	80%
Global	53.3%	46.7%	73.3%

Verdaderos positivos (TP)

Falso positivo (FP)

Falso negativo (FN)

Verdaderos negativos (TN)

$$\text{Precisión} = \frac{TP}{TP + FP} = \frac{140}{140 + 20} = 87.5\%$$

$$\text{Exhaustividad} = \frac{TP}{TP + FN} = \frac{140}{140 + 60} = 70\%$$



# F-Score

Media Armónica de precisión y exhaustividad

$$F - Score = 2 * \frac{precisión * exhaustividad}{precisión + exhaustividad} = 77.7\%$$

# Métricas para comparar modelos

## Matriz de Confusión - Precisión

### Limitaciones de la precisión ("accuracy") :

---

Supongamos un problema con 2 clases:

- 9990 ejemplos de la clase 1
- 10 ejemplos de la clase 2

Si el modelo de clasificación siempre dice que los ejemplos son de la clase 1, su precisión es

$$9990/10000 = 99.9\%$$

Totalmente engañosa, ya que nunca detectaremos ningún ejemplo de la clase 2.

# Métricas para comparar modelos

## Curva ROC

- Desarrolladas en los años 50 para analizar señales con ruido: caracterizar el compromiso entre aciertos y falsas alarmas.
- Permiten comparar visualmente distintos modelos de clasificación.
- El área que queda bajo la curva es una medida de la precisión (accuracy) del clasificador:
  - Cuanto más cerca estemos de la diagonal (área cercana a 0.5), menos preciso será el modelo.
  - Un modelo "perfecto" tendrá área 1.

# El espacio ROC

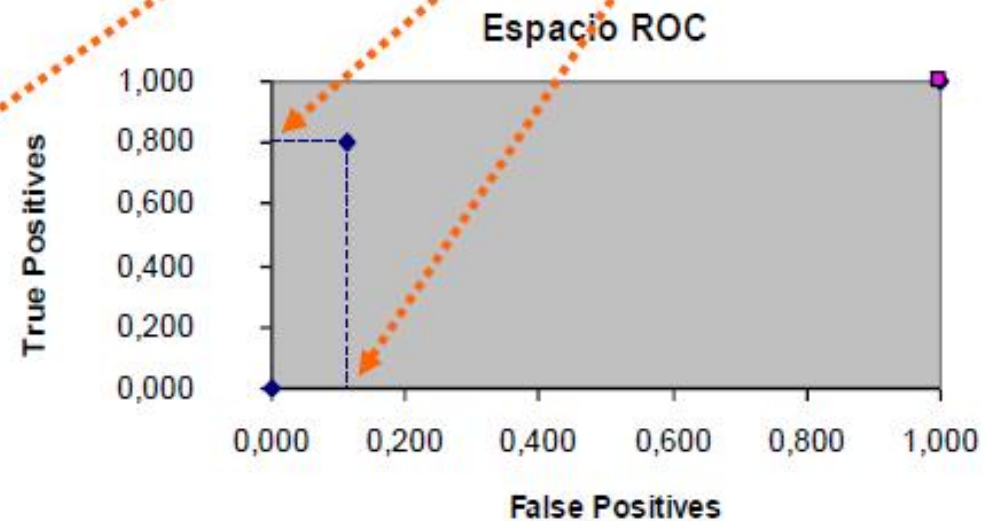
- Se normaliza la matriz de confusión por columnas: TPR, FNR TNR, FPR.

		Real	
		abrir	cerrar
Pred	ABRIR	400	12000
	CERRAR	100	87500

$$\begin{aligned} \text{TPR} &= 400 / 500 = 80\% \\ \text{FNR} &= 100 / 500 = 20\% \\ \text{TNR} &= 87500 / 99500 = 87,9\% \\ \text{FPR} &= 12000 / 99500 = 12,1\% \end{aligned}$$



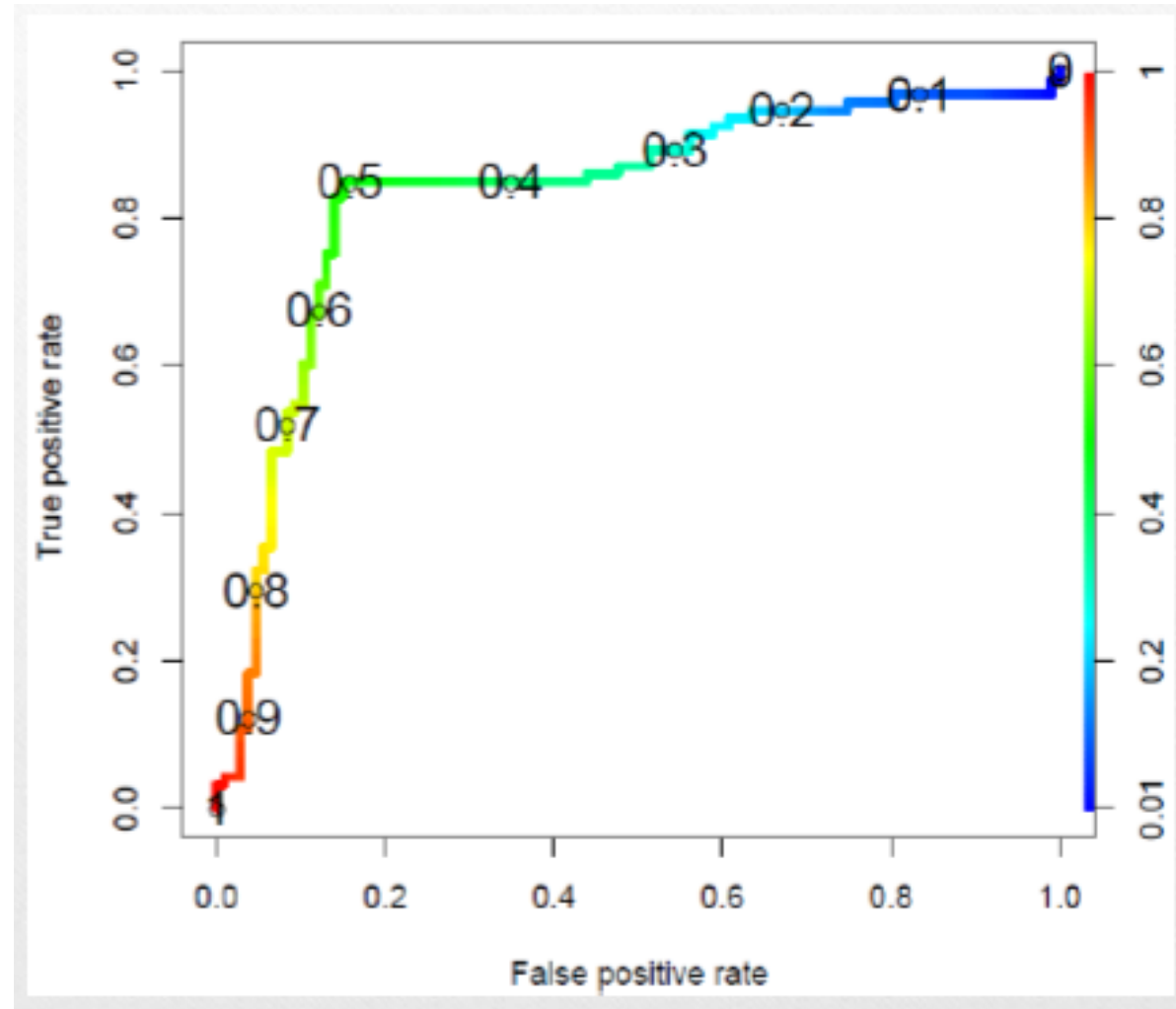
		Real	
		abrir	cerrar
Pred	ABRIR	0,8	0,121
	CERRAR	0,2	0,879





# Curva ROC

- **Discretización:**  
Seleccionando un  
valor de corte:  
 $f(x) \geq c \rightarrow \text{clase} = 1$   
 $f(x) < c \rightarrow \text{clase} = -1$



# KPI: Ganancia Neta

Es demostrable que la ganancia neta de un modelo de predicción binaria está dada por:

$$B * E * (\$I - \frac{\$C}{P})$$

B = Número total de posibles éxitos

E = Exhaustividad

P = Precisión

\$I = Ingreso por el éxito

\$C = Costo (o ganancia perdida) por el fracaso

# KPI: Ganancia Neta

De la anterior ecuación se puede deducir que para obtener una ganancia neta es necesario y suficiente que:

$$P > \frac{\$C}{\$I}$$

- Aumentar la exhaustividad (con la precisión constante) implica ampliar el volumen de éxitos.
- Aumentar la precisión (con exhaustividad constante) implica lograr una ganancia marginal mayor por reducción de costos.
- La evaluación de costos/beneficios con frecuencia no tiene en cuenta el tiempo (Customer lifetime value) o un posible crecimiento en portafolio de productos.

Otro modelos: <https://towardsdatascience.com/model-performance-cost-functions-for-classification-models-a7b1b00ba60>

# Regresión Logística y temas previos

Para mejorar un proceso de regresión logística se puede pensar en:

- **Feature Selection:** otros métodos para seleccionar las variables.
- **Feature Engineering:** ¿Hay posibilidad de relaciones no lineales? ¿Interacciones? ¿Otras variables?
- **Regularización:** Efectivamente se pueden aplicar Redes elásticas, Lasso o Ridge.