

# Estadística en Analítica

2023-2

Pablo A. Saldarriaga  
[psaldar2@eafit.edu.co](mailto:psaldar2@eafit.edu.co)

UNIVERSIDAD  
**EAFIT**

# ¿Dudas del Taller 1?

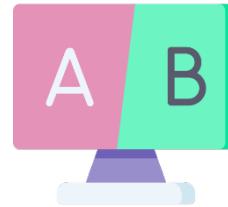
# Métodos de Envoltura para selección de variables



*Selección de un modelo base*



*Definición de métrica de desempeño*



*Partición del conjunto de datos*



*Entrenar el modelo con varias combinaciones de características*

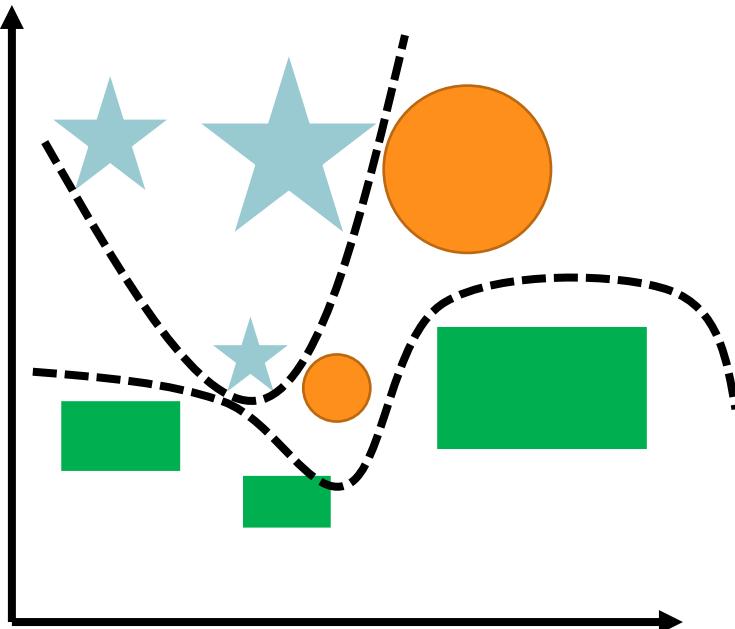


*Selección de las mejores características*

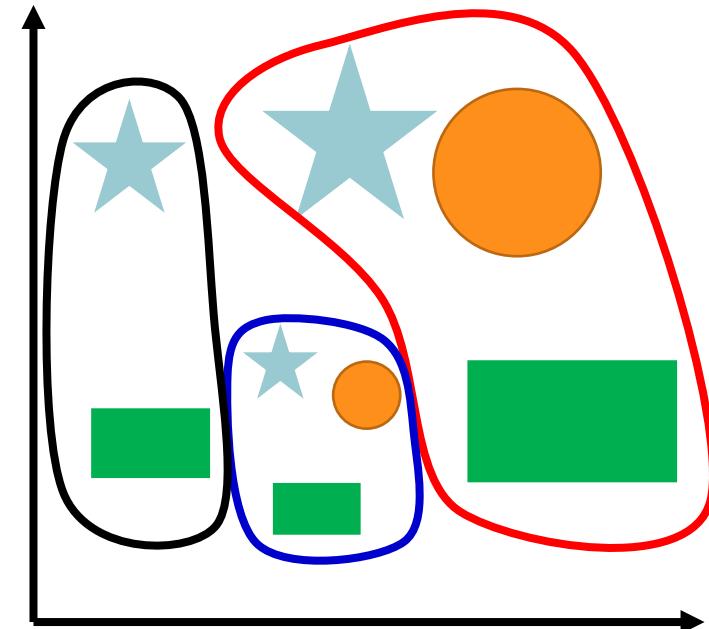
# Técnicas supervisadas: Clasificación

# Problema de Clasificación

*Supervisado*



*No Supervisado*



# Problema de Clasificación

*En la clasificación supervisada, la variable respuesta  $Y$  es cualitativa (Que toma valores en un conjunto  $C$ ). En este problema se busca:*

- ✓ Construir un clasificador  $C(X)$  que asigne una etiqueta de clase de un conjunto  $C$  a un valor futuro que no tiene categoría proveniente de la observación  $X$
- ✓ Medir la incertidumbre en cada clasificación
- ✓ Entender la importancia de los predictores en la respuesta que se tiene
- ✓ En varias ocasiones estamos interesados en estimar la probabilidad de que cada  $X$  pertenezca a cada una de las categorías de  $C$

# Aplicaciones de Modelos de Clasificación

*¿El score crediticio y la tenencia de una vivienda pueden ayudar a predecir si un cliente pagará un crédito?*

✓ **Variables predictoras :**

- ✓ Score crediticio (300 - 850)
- ✓ Tipo de Vivienda: Familiar, propia, rentada

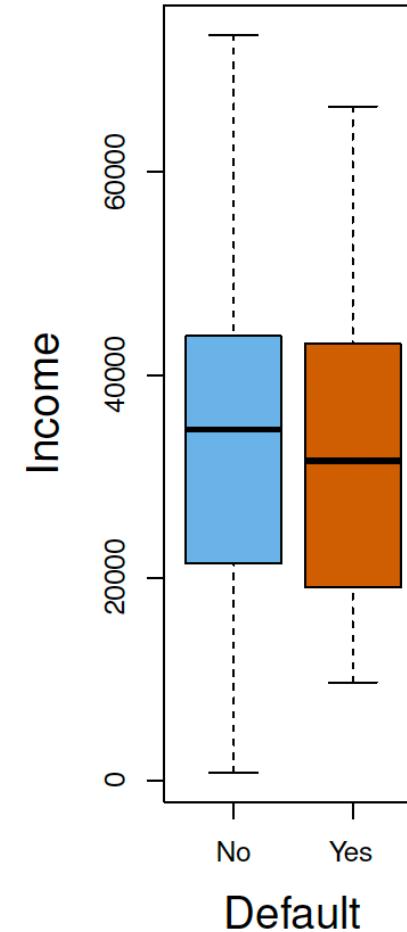
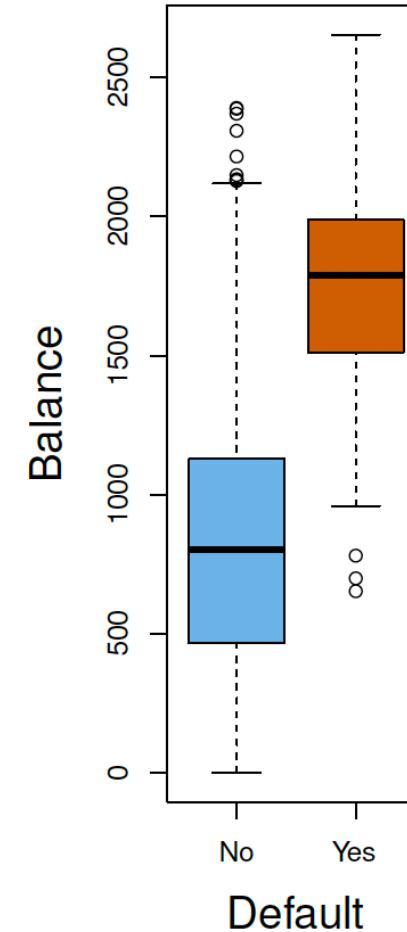
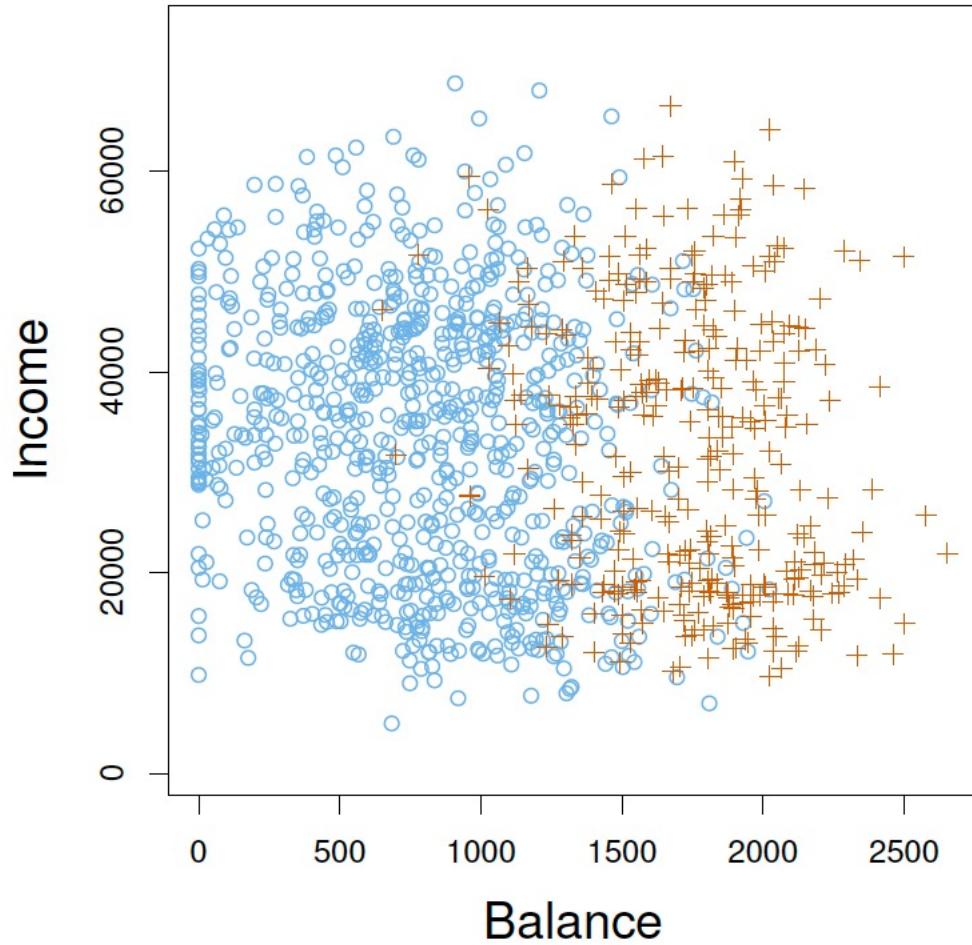


✓ **Variable respuesta:**

- ✓ No paga el crédito (default): 1/0



# Aplicaciones de Modelos de Clasificación

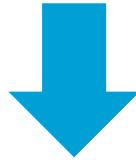


# Aplicaciones de Modelos de Clasificación

*¿El alcohol y cigarillo están asociados a problemas cardíacos?*

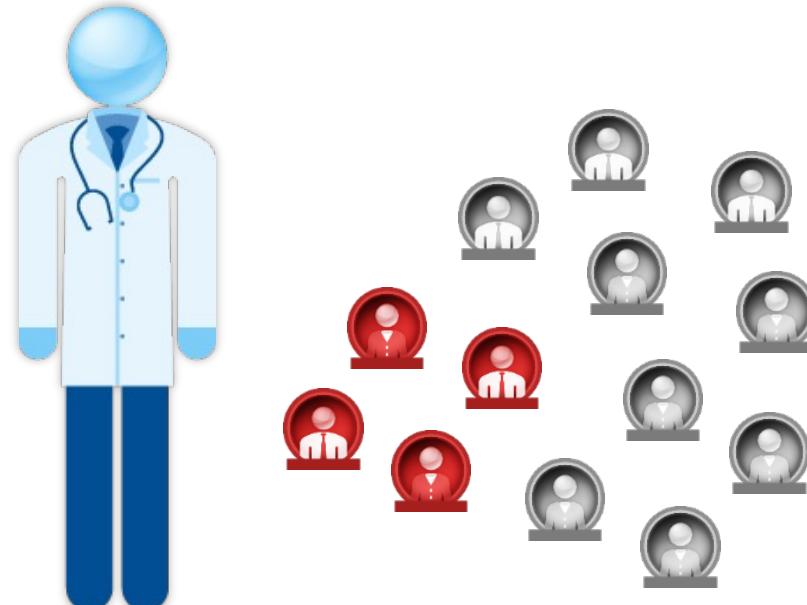
✓ **Variables predictoras :**

- ✓ Alcohol: ounces per day
- ✓ Smoking: cigarettes per day



✓ **Variable respuesta:**

- ✓ Problemas Cardíacos: 1/0



# Aplicaciones de Modelos de Clasificación

*¿Un cliente realiza una compra basado en comportamientos históricos?*

✓ **Variables predictoras:**

- ✓ Compras en los últimos 90 días
- ✓ Grupo de edad
- ✓ Género



✓ **Variable respuesta:**

- ✓ Compró en una campaña: Si / No



# Sesgo VS Varianza

Suponga que tenemos un modelo ajustado  $f(x)$ , y considere una observación  $(x_0, y_0)$  a ser probada en el modelo. Si el modelo real es  $Y = f(X) + \varepsilon$  (con  $f(x) = E[Y|X]$ ), entonces:

- | La varianza se refiere a que tanto puede cambiar la función a estimar  $f$  si cambiamos el conjunto de datos

$$E(y_0 - \hat{f}(x_0))^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\varepsilon).$$

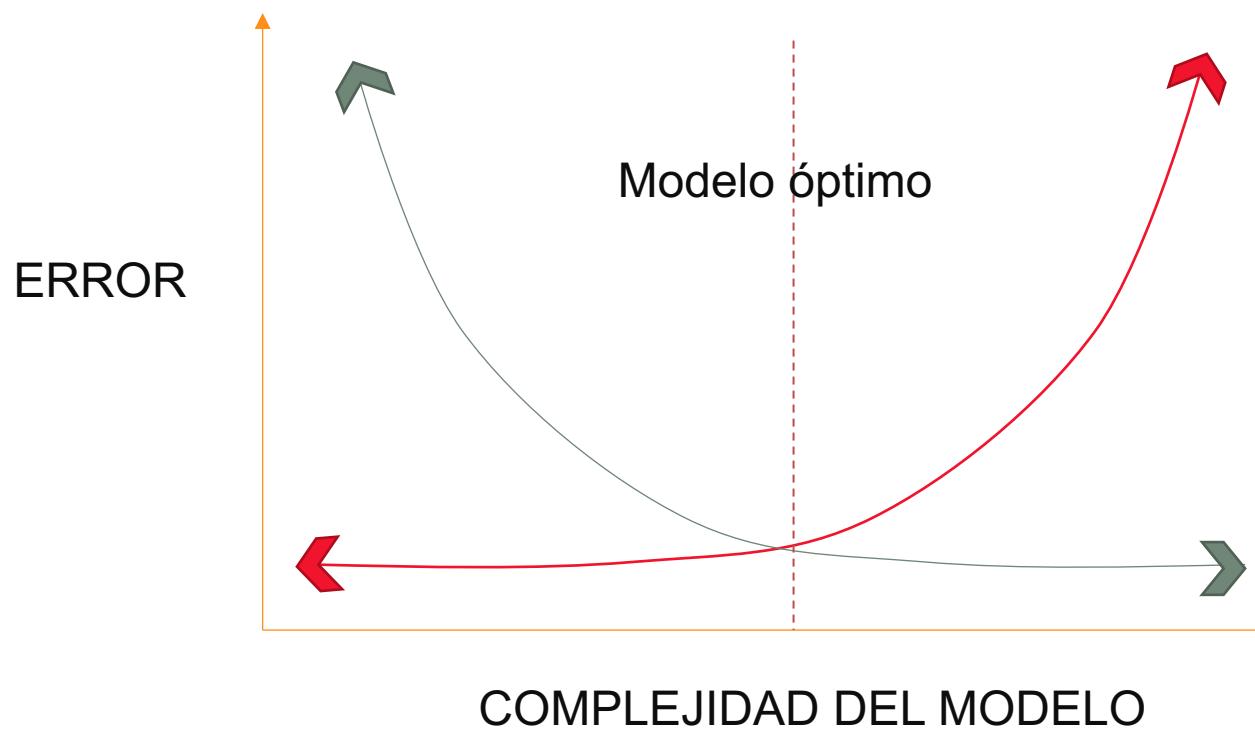
- | El sesgo se refiere al error que es introducido por un modelo al considerar un modelo más simple

Comúnmente la **flexibilidad** de  $f$  aumenta si su varianza aumenta y disminuye el sesgo. Por lo que se debe seleccionar un punto de **trade-off entre la varianza y sesgo** de un modelo

# Sesgo VS Varianza

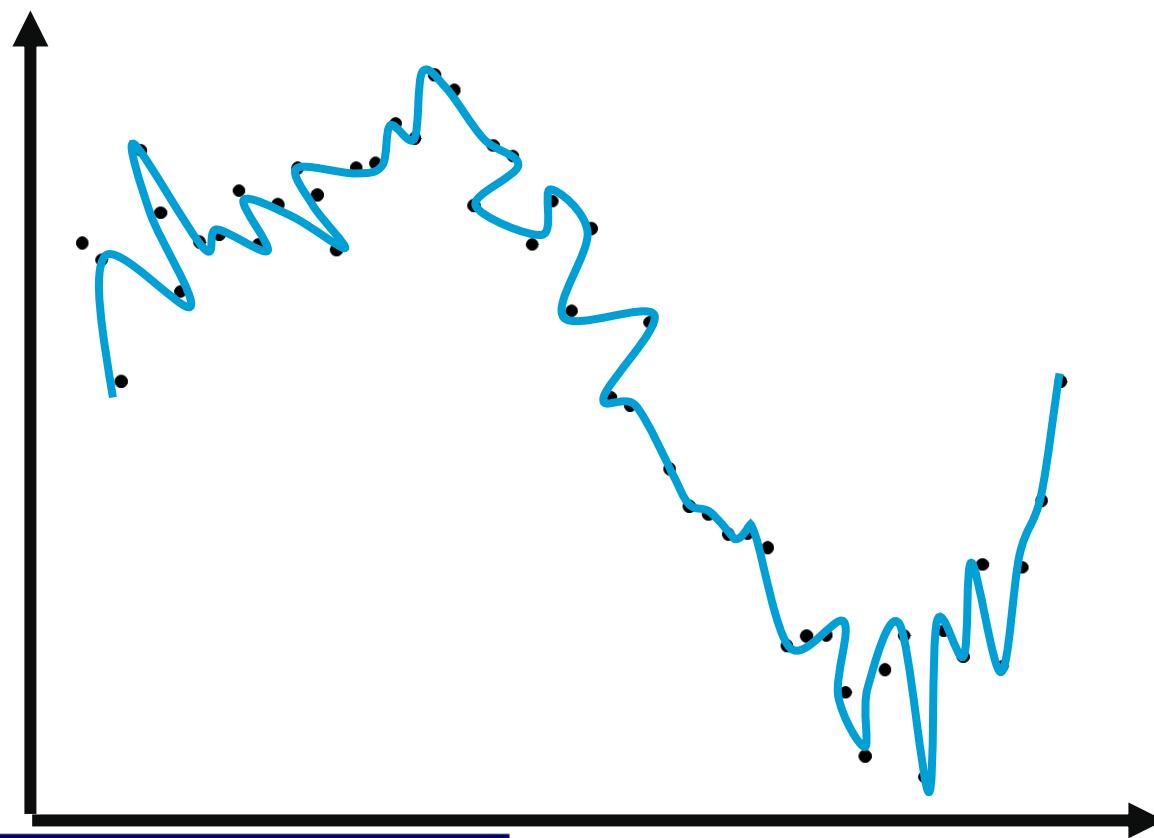
**Sesgo:** Tendencia del modelo a aprender lo incorrecto e ignorar la información de los datos.

**Varianza:** La sensibilidad de un modelo a pequeños cambios en el conjunto de entrenamiento.



# Overfitting

Sucede cuando el modelo se ajusta muy bien a los datos de entrenamiento pero **no** generaliza bien.



# Overfitting

El modelo es demasiado complejo para la cantidad/calidad de datos.

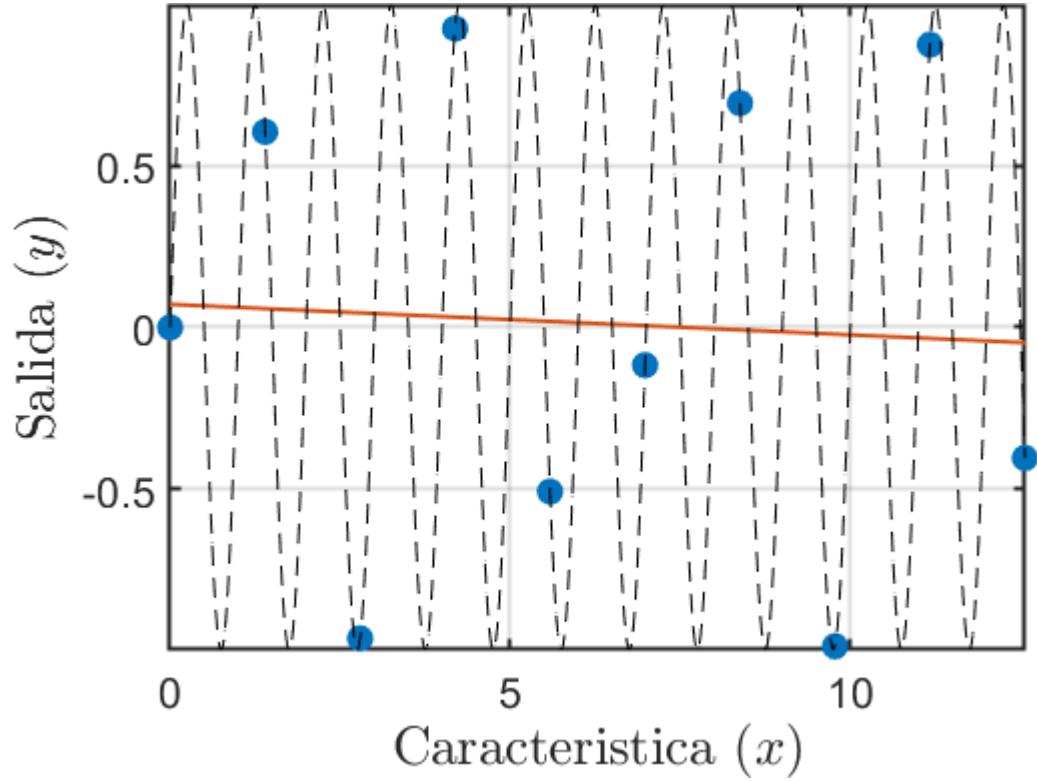
- Conjunto de datos de entrenamiento es muy ruidoso.
- Muy pocos datos de entrenamiento.

**Posibles soluciones:**

- Simplificar el modelo.
- Incluir más datos.
- Reducir el ruido de los datos.
- Añadir restricciones al modelo (**Regularización**).

# Underfitting

Sucede cuando el modelo **no** se ajusta bien a los datos de entrenamiento y es demasiado general.



## Posibles soluciones:

- Elegir un modelo más complejo.
- Mejorar (incluir más) características.
- Reducir las restricciones del modelo.

# Validación Cruzada

## Error en entrenamiento

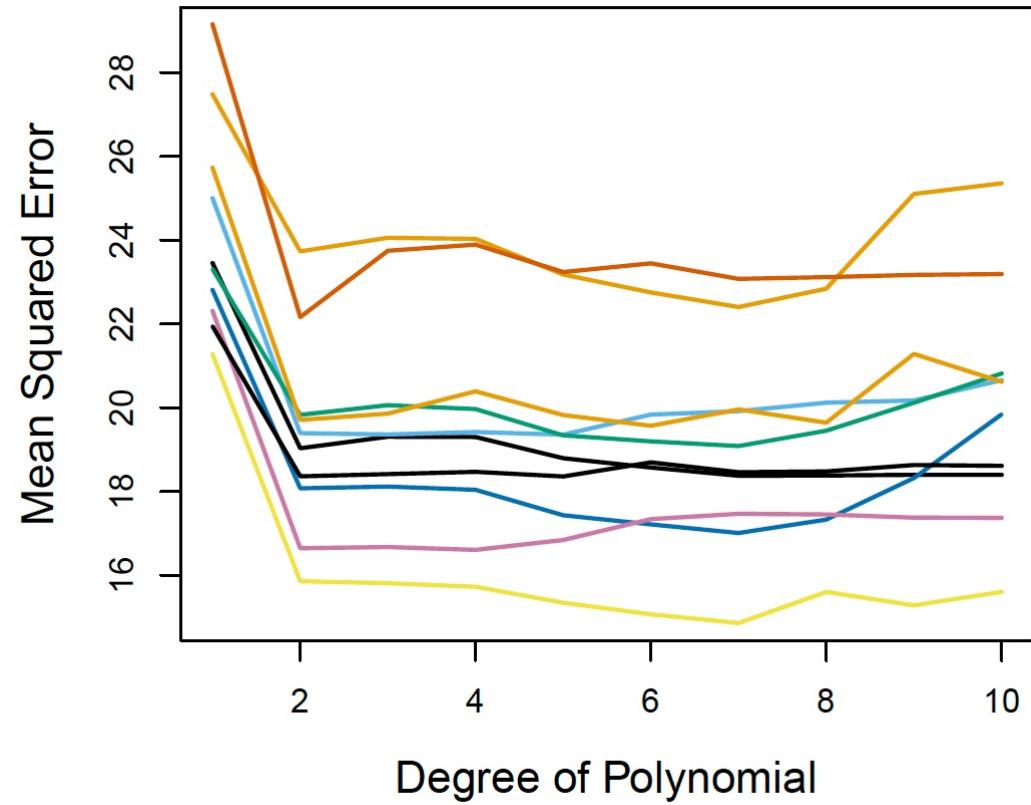
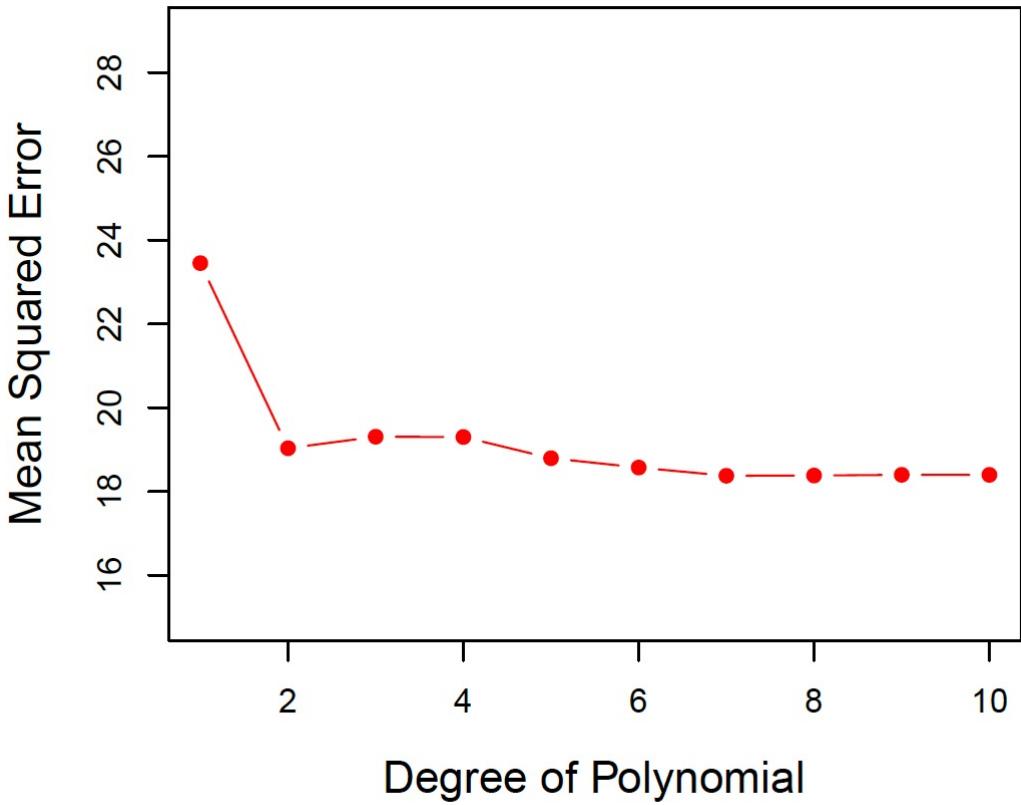
Es el error que se puede calcular contrastando la aplicación del método estadístico a los datos de entrenamiento, y comparando con los valores reales

## Error en test

Es el error promedio que resulta de utilizar aprendizaje estadístico al momento de responder a una observación nueva

*Existen  
aproximaciones para  
medir este error*

# Validación Cruzada



# Validación Cruzada

*Leave one out Cross-validation*

1. Definir un conjunto de datos de entrenamiento
2. Seleccionar un registro que no es usado en el entrenamiento
3. Ajustar el modelo con los  $n - 1$  registros
4. Obtener el error en el registro que no fue usado en el entrenamiento
5. Repetir estos pasos hasta que haya sacado todos los registros una vez

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n MSE_i$$

*Esta estimación aplica para cualquier error*

# Validación Cruzada

*K-Folds Cross-validation*

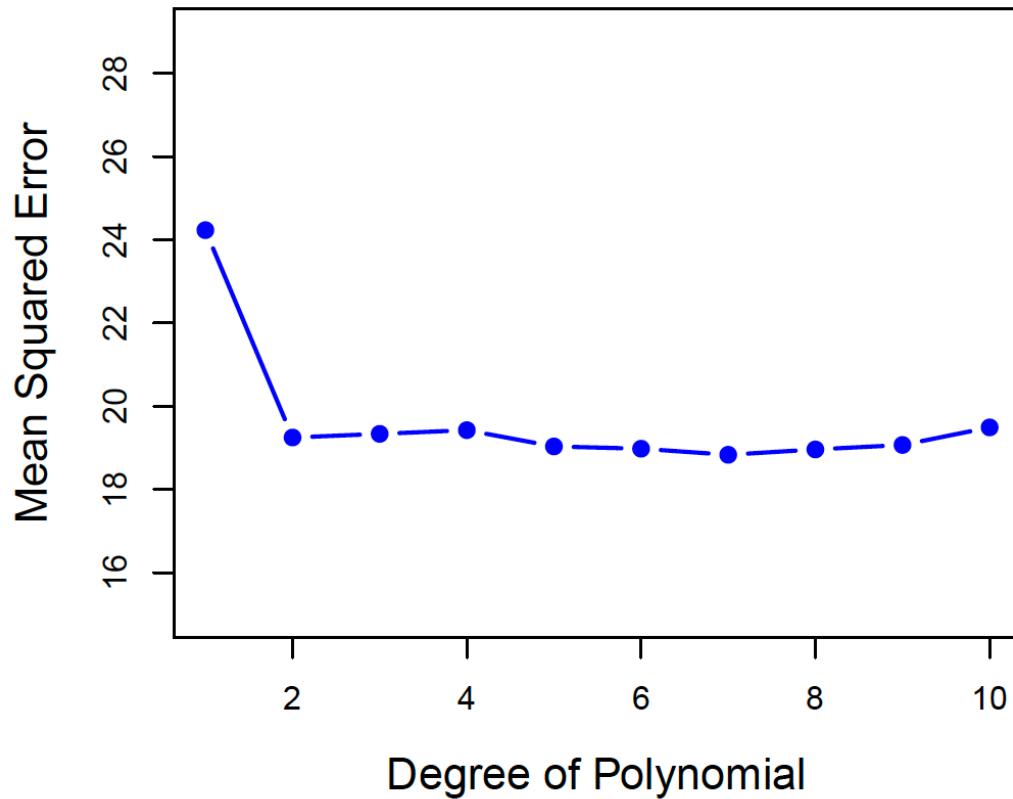
1. Se partitiona aleatoriamente el conjunto de datos en  $k$  subconjuntos
2. Cada uno de los subconjuntos obtenidos se utilizará de test set para evaluar el modelo con el resto de subconjuntos
3. Se obtiene el promedio de las evaluaciones realizadas para obtener el resultado final

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k MSE_i$$

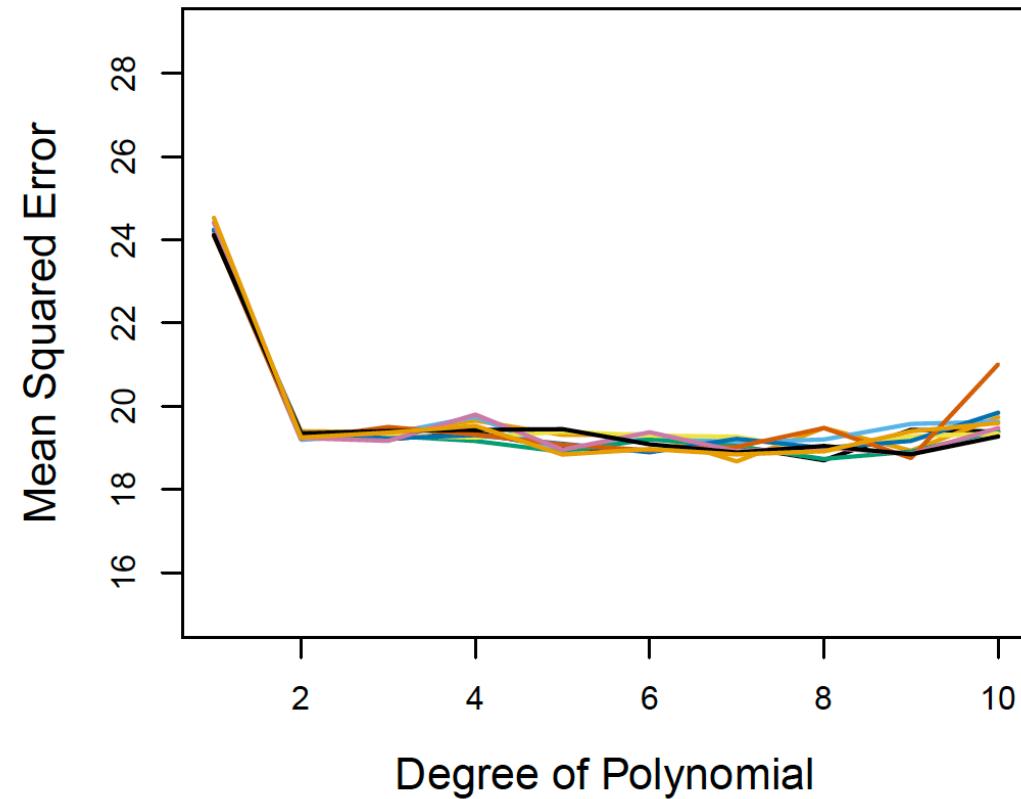
*Esta estimación aplica para cualquier error*

# Validación Cruzada

LOOCV

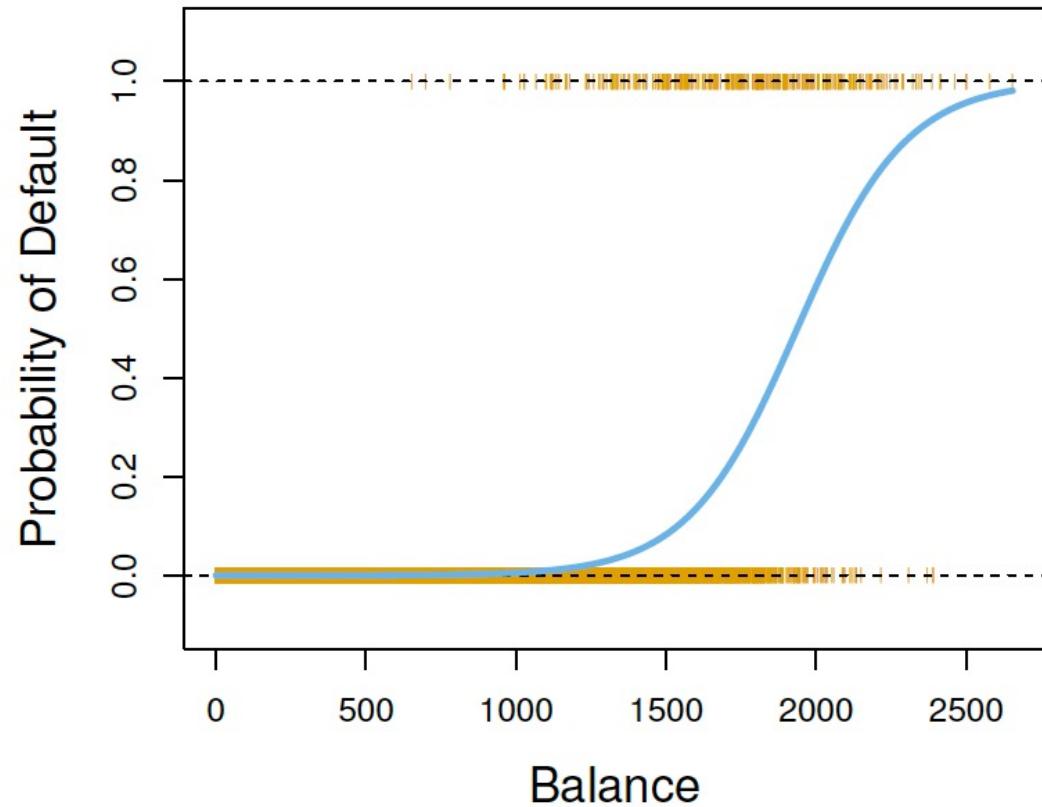
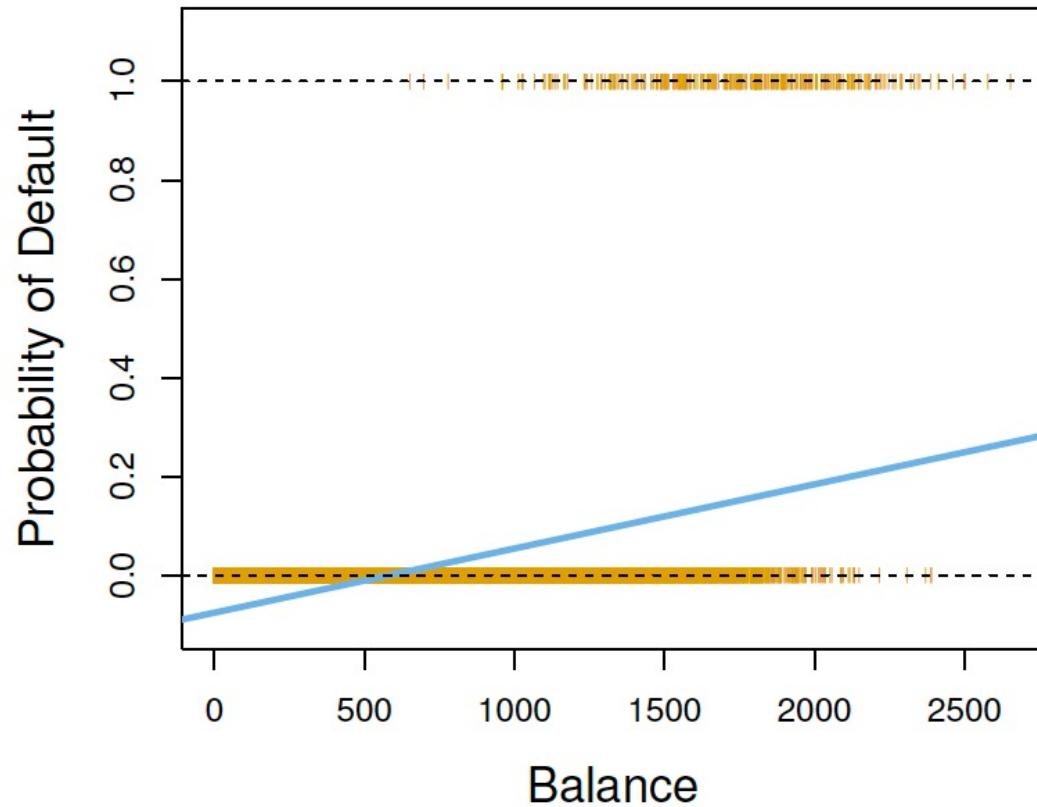


10-fold CV



# Técnicas supervisadas: Clasificación

# ¿Por qué un clasificador y no un regresor?



# Regresión logística

Técnica que se usa para clasificar los datos en categorías/clases. Los modelos de regresión logística son de la forma:

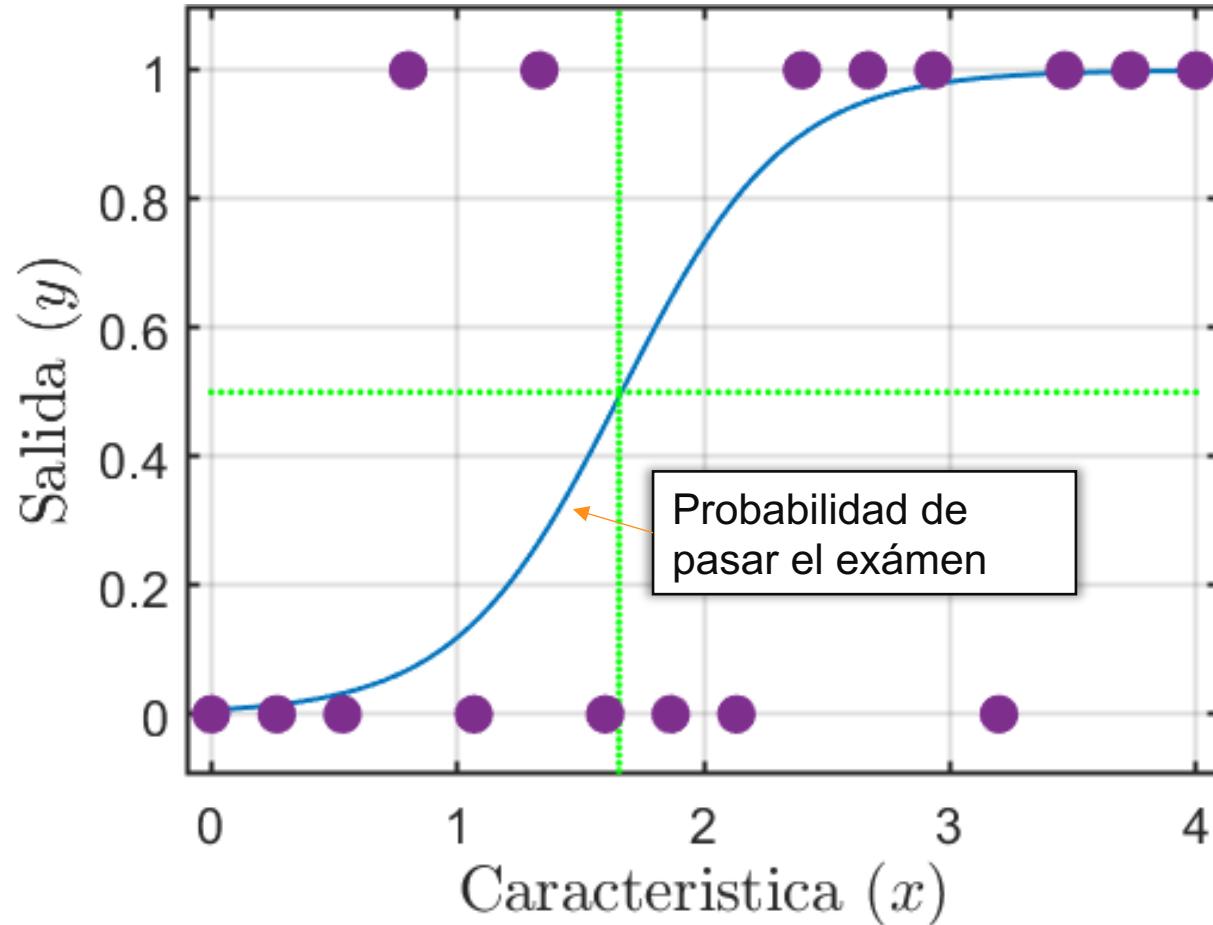
$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}$$

Busca minimizar:

$$\varepsilon = -\frac{1}{n} \sum_{i=1}^n y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

$$\log \left( \frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

# Regresión logística



Un grupo de 16 estudiantes estudiaron para un examen un tiempo entre 0 a 4 horas.

¿Cómo afecta el número de horas que un estudiante dedica a preparar un examen a la probabilidad de aprobarlo?

$$y = \frac{1}{1 + e^{-(0.05+0.6x)}}$$

# Regresión logística

*Predicción de default para un crédito*

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student [Yes]	-0.6468	0.2362	-2.74	0.0062

# Regresión logística

<b>Dep. Variable:</b>	Class	<b>No. Observations:</b>	699			
<b>Model:</b>	Logit	<b>Df Residuals:</b>	690			
<b>Method:</b>	MLE	<b>Df Model:</b>	8			
<b>Date:</b>	Thu, 02 Nov 2023	<b>Pseudo R-squ.:</b>	0.8438			
<b>Time:</b>	15:21:13	<b>Log-Likelihood:</b>	-70.333			
<b>converged:</b>	True	<b>LL-Null:</b>	-450.26			
<b>Covariance Type:</b>	nonrobust	<b>LLR p-value:</b>	9.174e-159			
	coef	std err	z	P> z	[0.025	0.975]
<b>const</b>	-9.9456	1.032	-9.634	0.000	-11.969	-7.922
<b>Clump_thickness</b>	0.5776	0.119	4.852	0.000	0.344	0.811
<b>Uniformity_of_cell_size</b>	-0.0116	0.176	-0.066	0.948	-0.356	0.333
<b>Uniformity_of_cell_shape</b>	0.5679	0.191	2.969	0.003	0.193	0.943
<b>Marginal_adhesion</b>	0.3137	0.100	3.125	0.002	0.117	0.510
<b>Single_epithelial_cell_size</b>	0.1306	0.141	0.929	0.353	-0.145	0.406
<b>Bland_chromatin</b>	0.5800	0.146	3.984	0.000	0.295	0.865
<b>Normal_nucleoli</b>	0.1232	0.099	1.248	0.212	-0.070	0.317
<b>Mitoses</b>	0.6079	0.324	1.875	0.061	-0.027	1.243

# Regresión logística – Regularización

$$\mathcal{E} = -\frac{1}{n} \sum_{i=1}^n y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) + \lambda \sum_{j=1}^p \text{penalty}(\beta_j)$$

$$\text{penalty}(\beta_j) = |\beta_j|$$

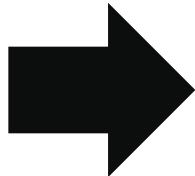
*Regularización L1*

$$\text{penalty}(\beta_j) = \beta_j^2$$

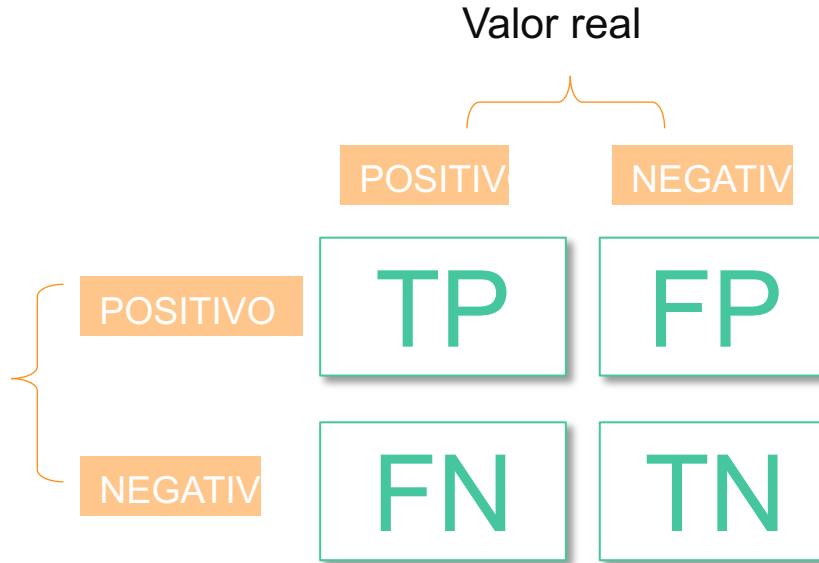
*Regularización L2*

# Métricas de desempeño: Modelos de Clasificación

**Matriz de Confusión**



Resultado de la predicción



$$\frac{TP + TN}{TP + FP + FN + TN}$$

**Exactitud**

$$\frac{TP}{TP + FP}$$

**Precisión**

$$\frac{TP}{TP + FN}$$

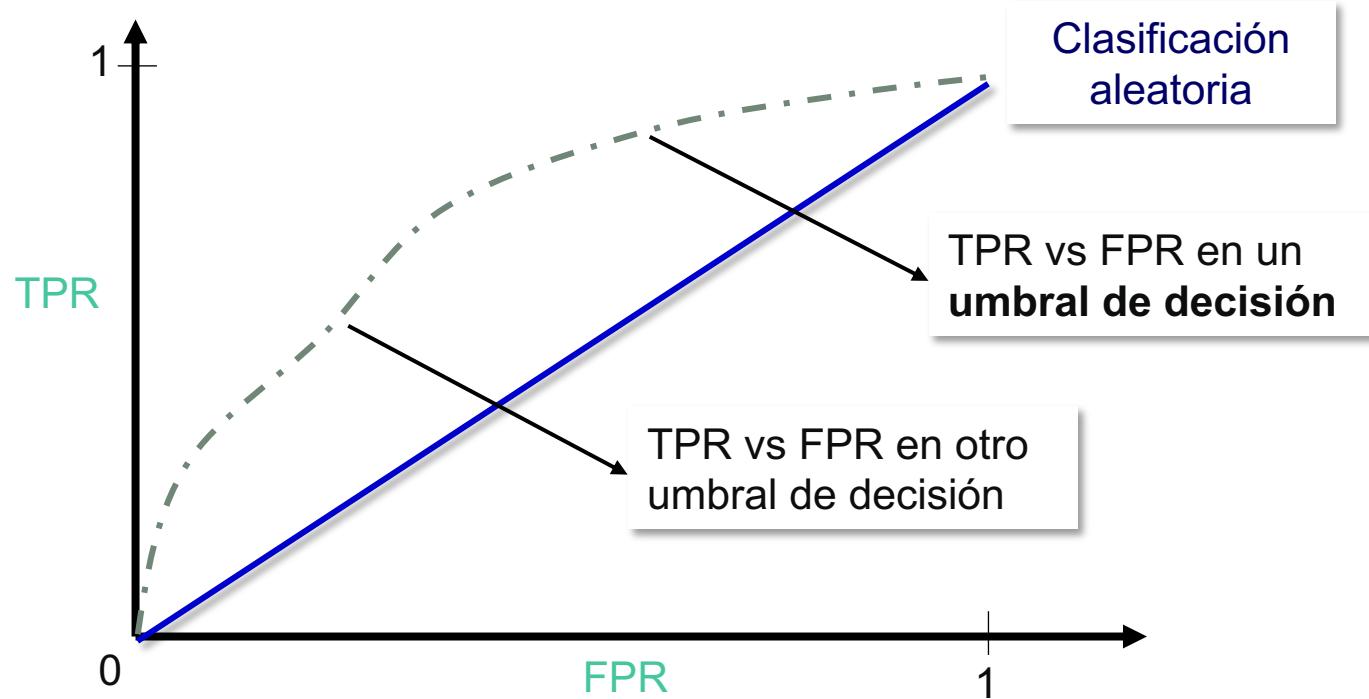
**Recall**

$$2 * \frac{Precision * recall}{Precision + recall}$$

**F-Score**

# Métricas de desempeño: Modelos de Clasificación

Curva ROC (Curva de característica operativa del receptor)  
Clasificadores probabilísticos



Tasa de verdaderos positivos  
(Recall – sensibilidad)

$$\frac{TP}{TP + FN}$$

Tasa de falsos positivos

$$\frac{FP}{TN + FP}$$

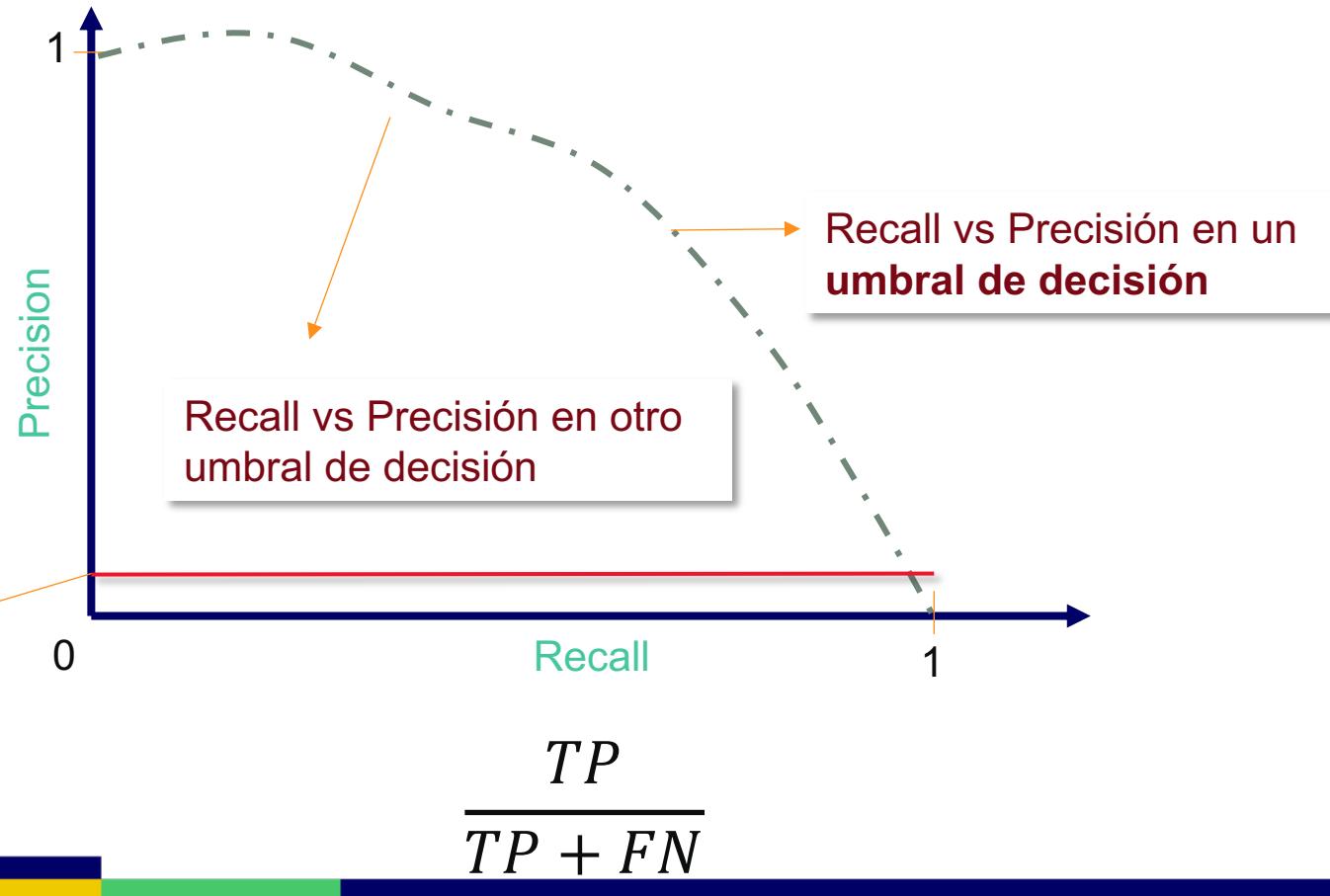
# Métricas de desempeño: Modelos de Clasificación

## Curva Precisión - Recall

Bases de datos desbalanceadas.

$$\frac{TP}{TP + FP}$$

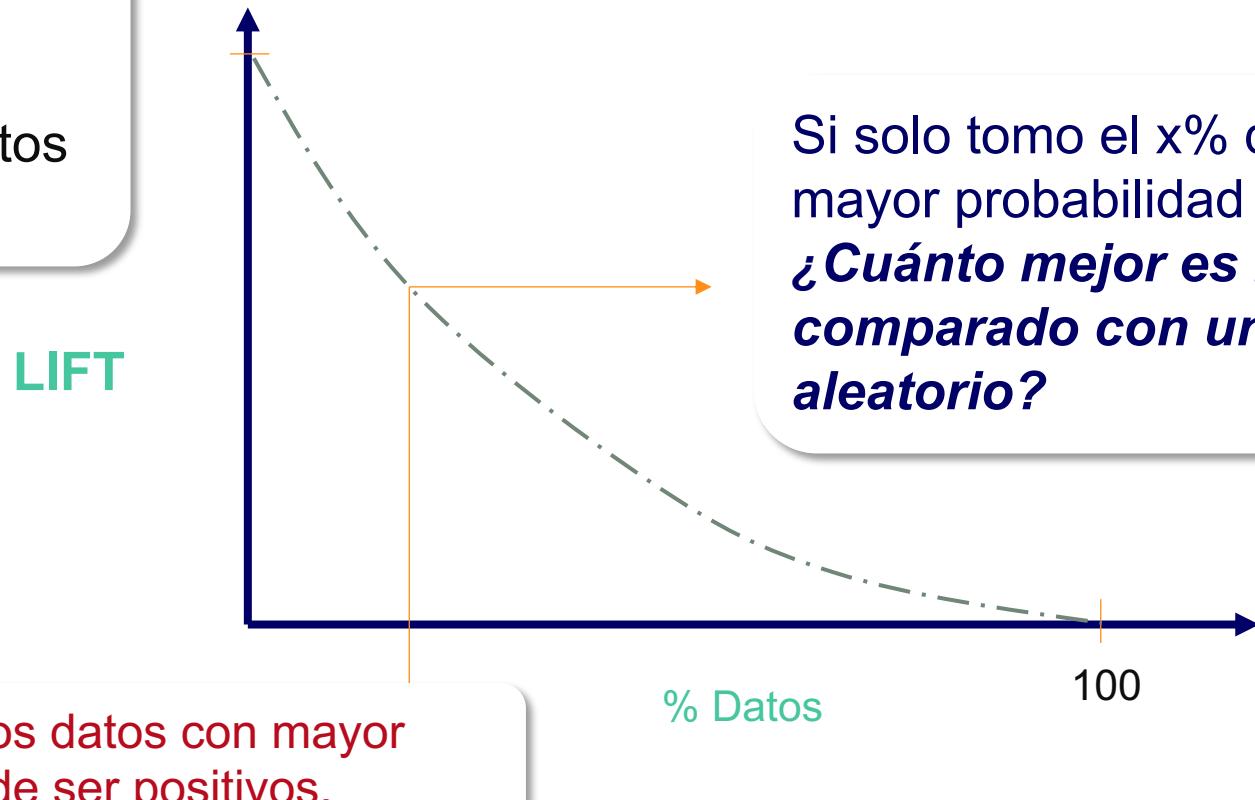
Clasificador aleatorio (Tasa real de positivos)



# Métricas de desempeño: Modelos de Clasificación

## Curva Lift

Se usa en bases de datos desbalanceadas



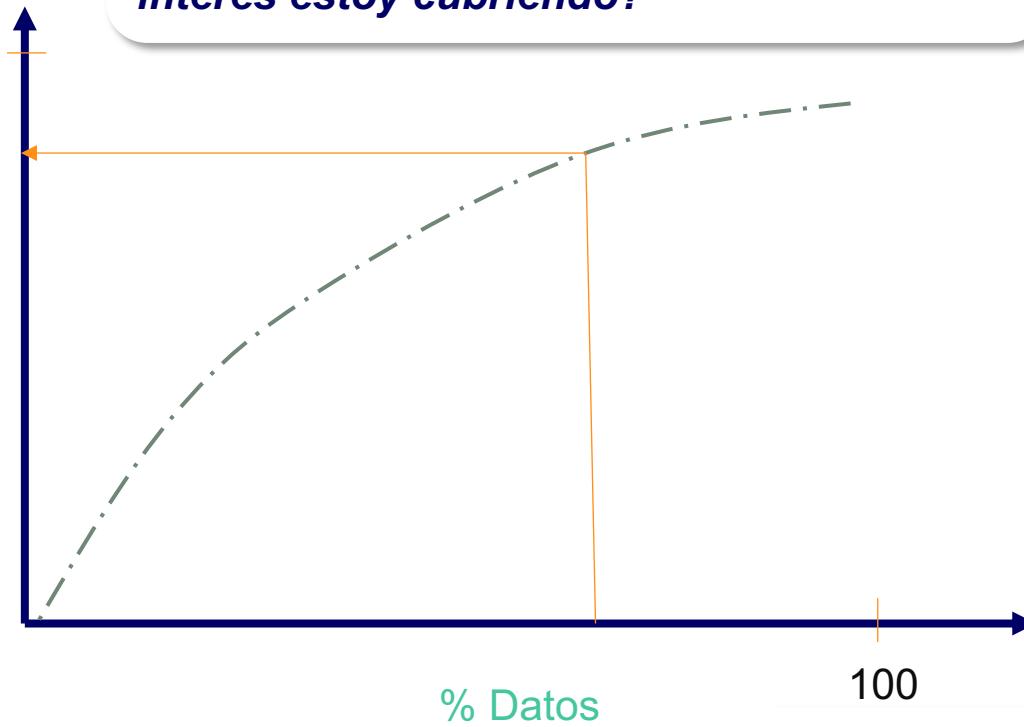
Si solo tomo el  $x\%$  de los datos con mayor probabilidad de ocurrencia.  
***¿Cuánto mejor es mi modelo comparado con un clasificador aleatorio?***

# Métricas de desempeño: Modelos de Clasificación

## Curva Ganancia acumulada

Se usa en bases de datos desbalanceadas

Cumulative gain



Si solo tomo el x% de los datos con mayor probabilidad de ocurrencia.

*¿Qué % de elementos de la clase de interés estoy cubriendo?*

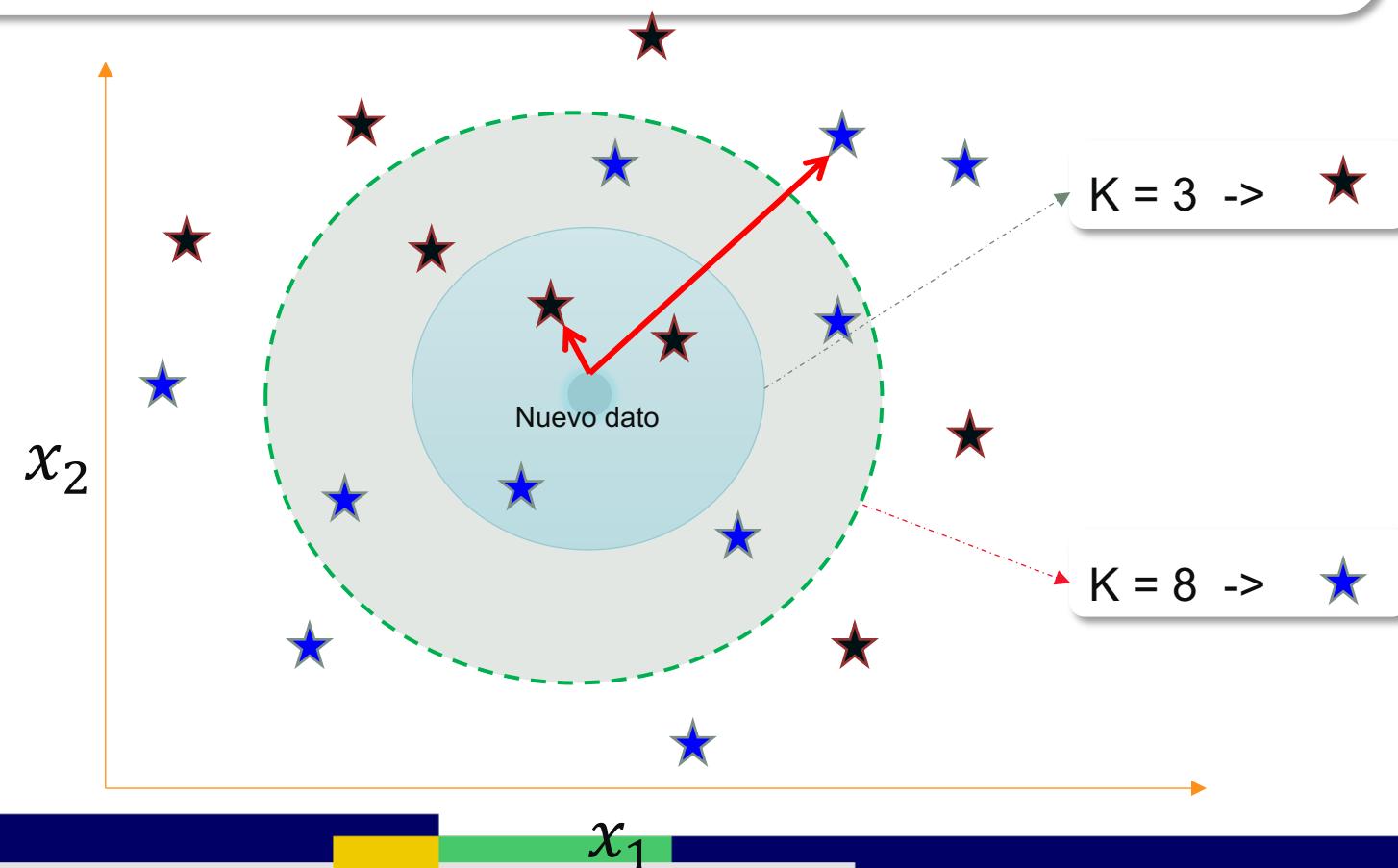
Ordenamos los datos con mayor probabilidad de ser positivos.

# K-nearest neighbors (KNN)

La respuesta es obtenida **localmente** computando la **distancia** de un nuevo dato al conjunto de datos de entrenamiento.

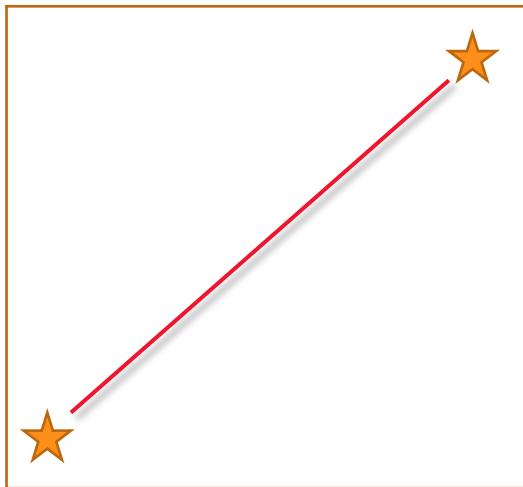
No requiere entrenamiento previo.

**Mejora:** Asignar pesos dependiendo de la distancia ( $\frac{1}{d}$ )

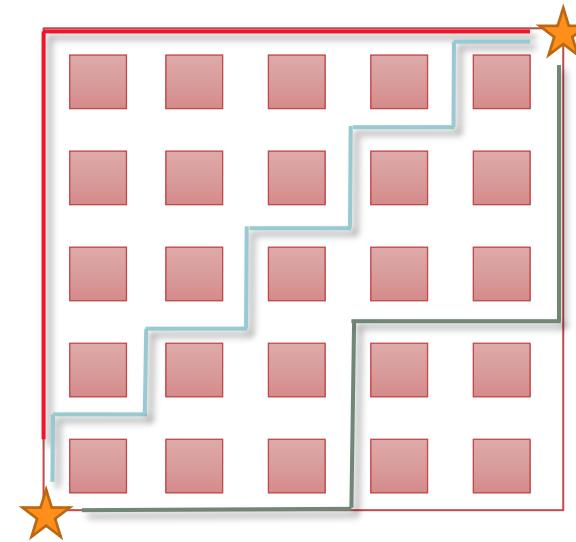


# ¿Cómo medimos?

“¿Cuál es la distancia más corta entre dos puntos?”



$$d_2(p, q) = \|p - q\|_2 = \sqrt{\sum_i (p_i - q_i)^2}$$



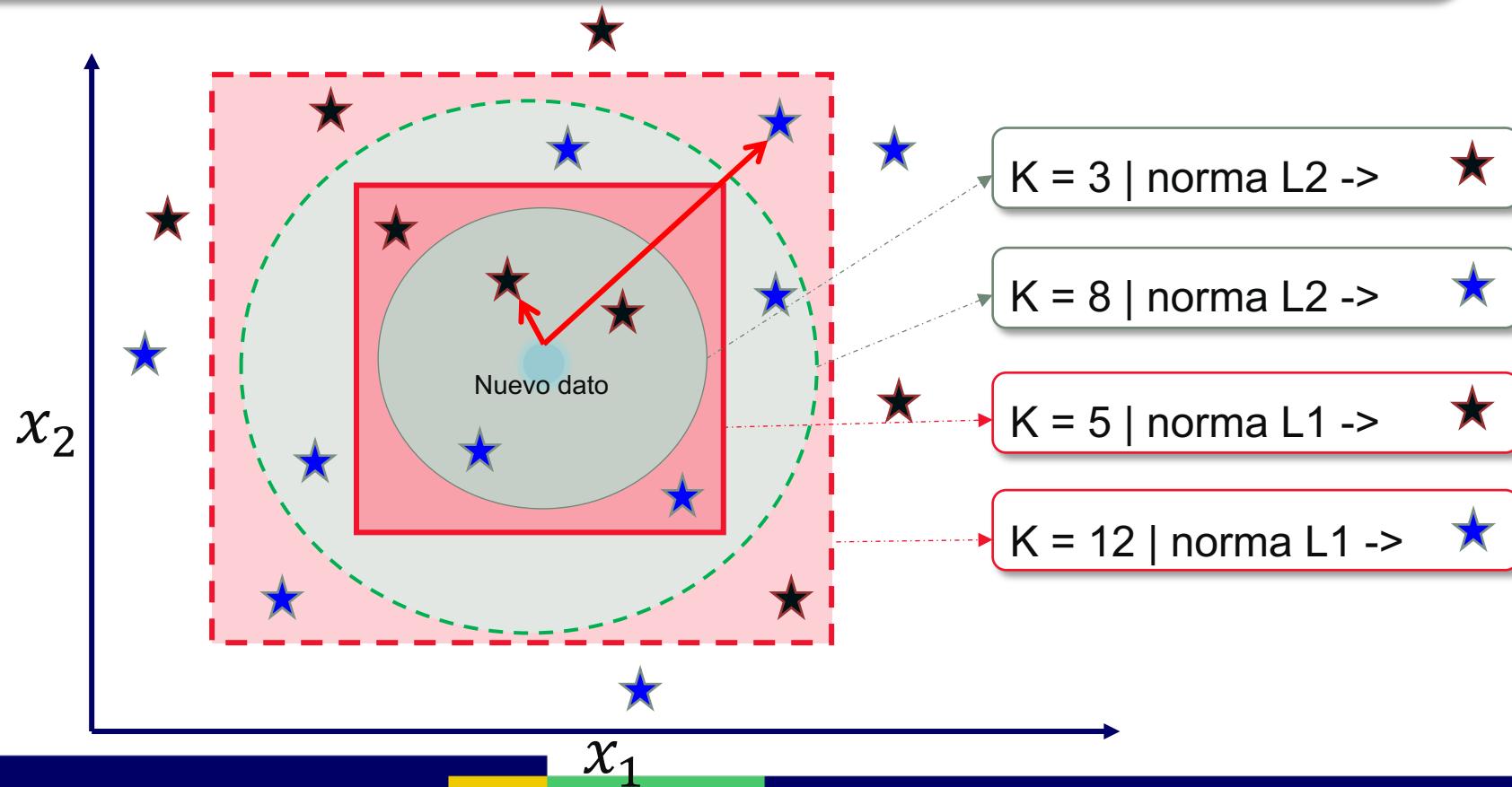
$$d_1(p, q) = \|p - q\|_1 = \sum_i |p_i - q_i|$$

# K-nearest neighbors (KNN)

La respuesta es obtenida **localmente** computando la **distancia** de un nuevo dato al conjunto de datos de entrenamiento.

No requiere entrenamiento previo.

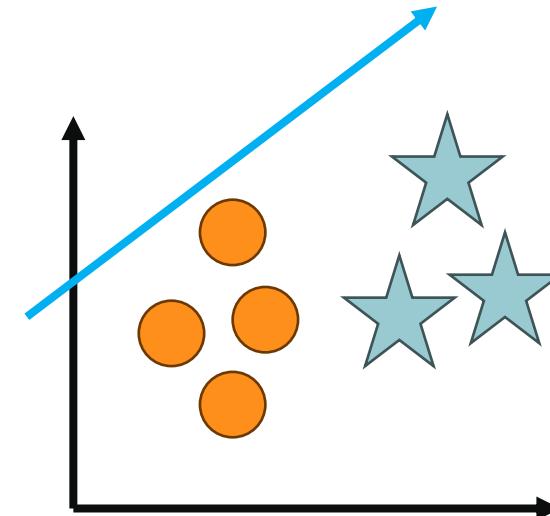
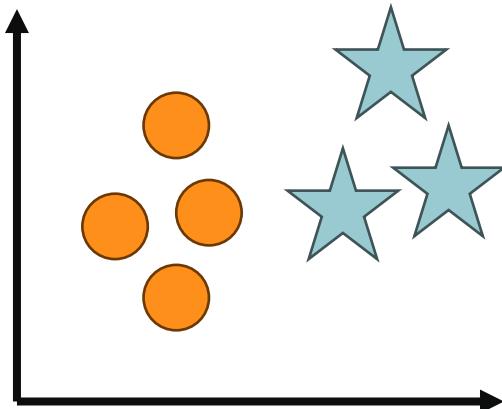
**Mejora:** Asignar pesos dependiendo de la distancia ( $\frac{1}{d}$ )



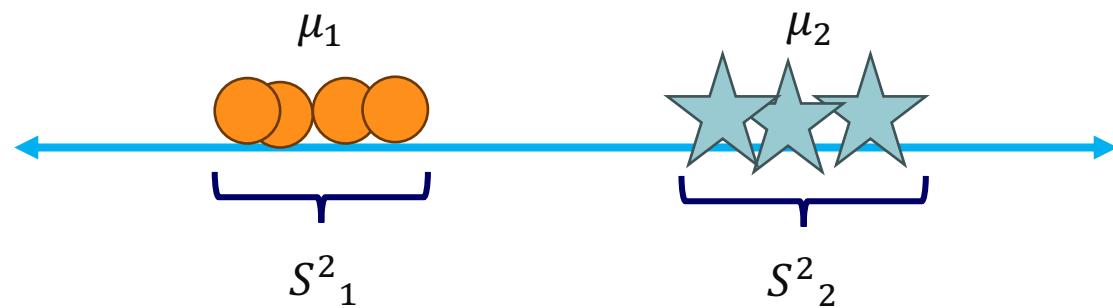
# Análisis discriminante Lineal

*Busca maximizar la separabilidad entre categorías conocidas*

*¿Cómo proyectó mis datos?*



# Análisis discriminante Lineal



$$\frac{(\mu_1 - \mu_2)^2}{S^2_1 + S^2_2}$$

# Análisis discriminante Lineal

Asumimos la función de densidad para los datos de cada clase:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

Utilizando el teorema de bayes, podemos estimar la probabilidad de que se obtenga la clase k dado un punto x como:

$$P(Y = k|x) = p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

# Análisis discriminante Lineal

Considerando:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$$

$$f_k(x) = \Pr(X = x|Y = k)$$

$$\pi_k = \Pr(Y = k)$$

# Análisis discriminante Lineal

Para clasificar un punto  $X$ , necesitamos ver cual es dicha  $p_k(x)$  es la mas grande, podemos considerar la función de score discriminante:

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

**Para estimar los parametros tenemos:**

$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\hat{\sigma}^2 = \frac{1}{n - K} \sum_{k=1}^K \sum_{i: y_i=k} (x_i - \hat{\mu}_k)^2$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i: y_i=k} x_i$$

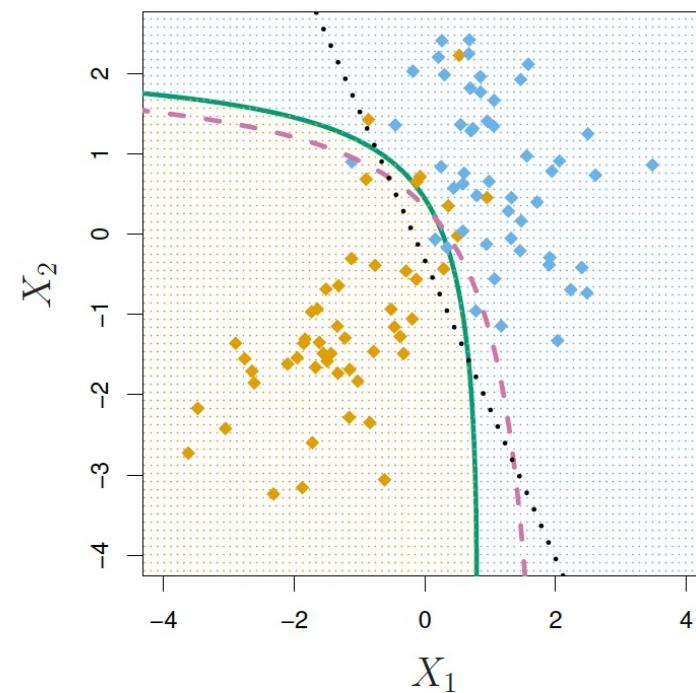
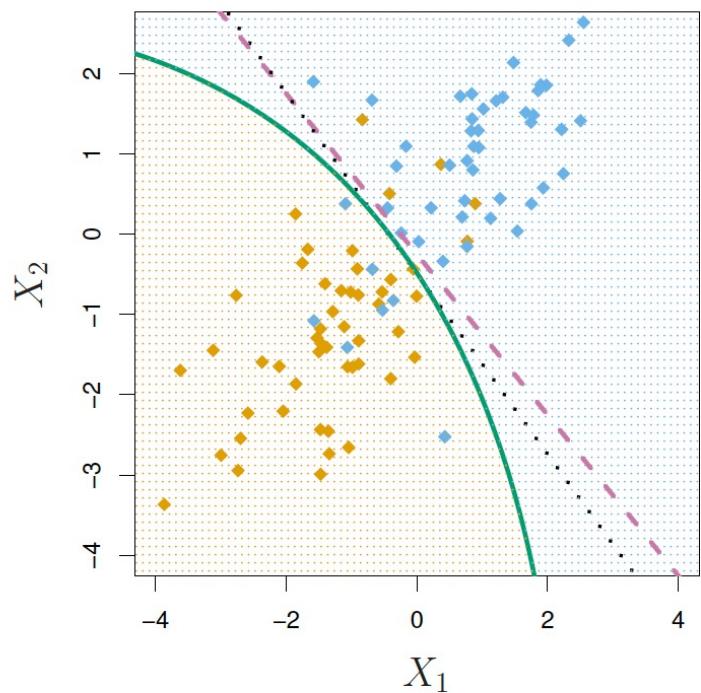
$$= \sum_{k=1}^K \frac{n_k - 1}{n - K} \cdot \hat{\sigma}_k^2$$

# Análisis discriminante Lineal

Una vez se tienen las estimaciones de la función de score discriminante, podemos retornar esto en probabilidades para cada clase como:

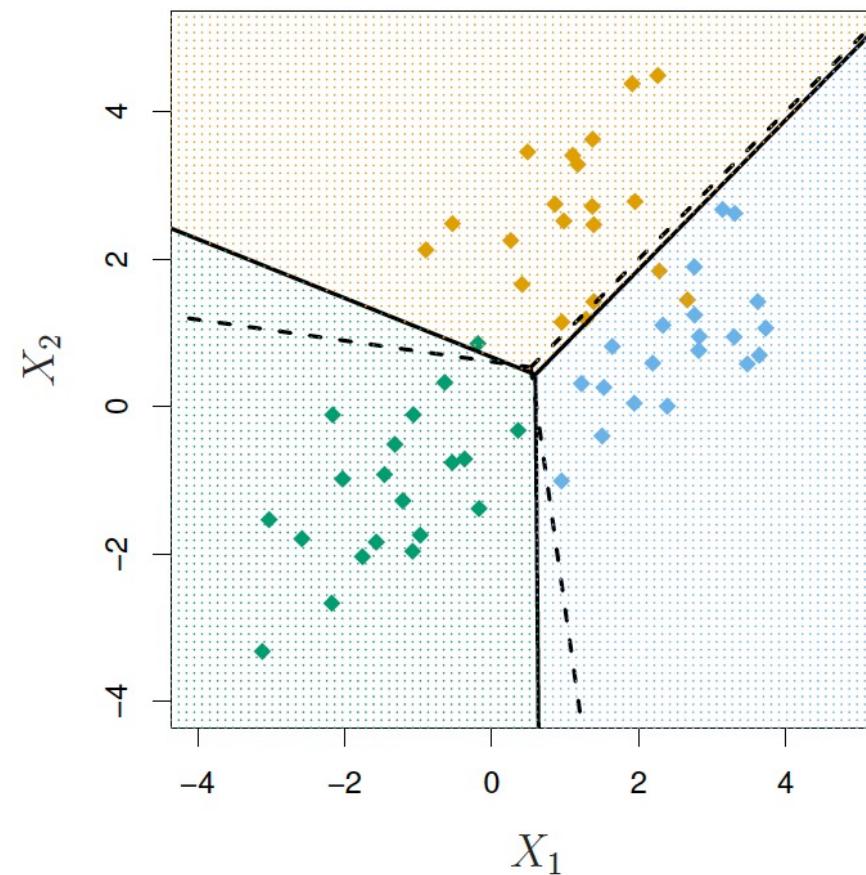
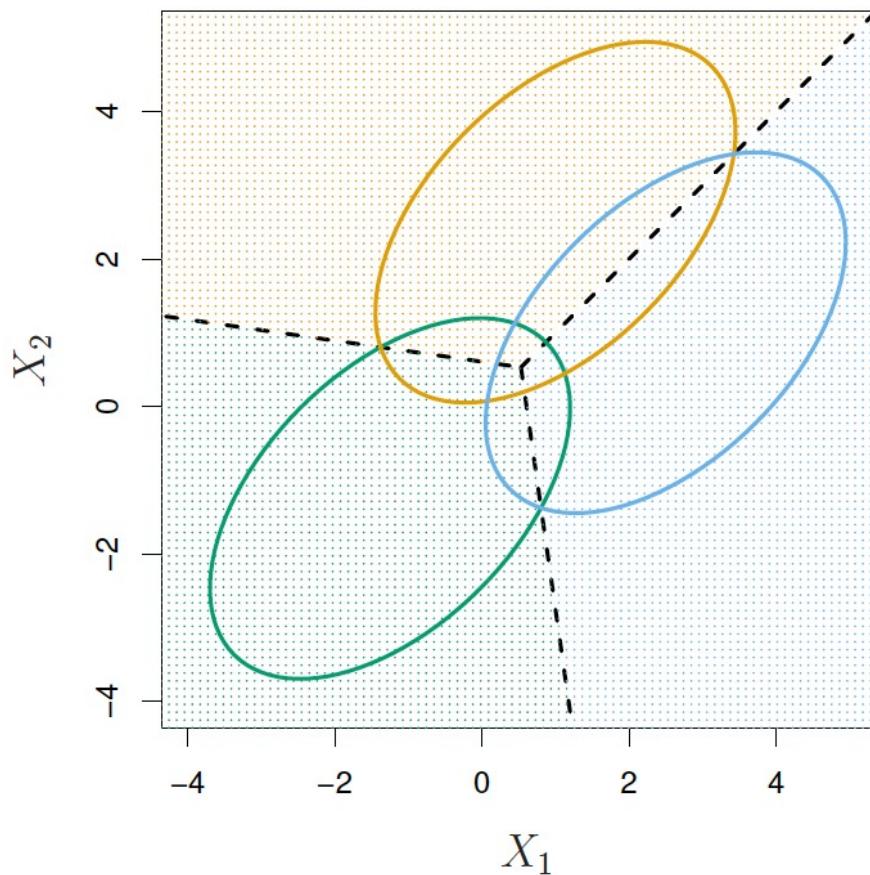
$$\widehat{\Pr}(Y = k | X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^K e^{\hat{\delta}_l(x)}}.$$

# Análisis discriminante cuadrático



$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k - \frac{1}{2} \log |\Sigma_k|$$

# Análisis discriminante Lineal



# Análisis discriminante Lineal

## *¿Por qué LDA?*

- ✓ Cuando las clases estan bien separadas, el modelo de regresión logistica sufre de problemas de estabilidad, mientras que LDA no
- ✓ Si la cantidad de registros es pequeño, y la distribución de los datos es aproximadamente normal para cada clase, entonces se puede demostrar que LDA tiene un mejor desempeño que un modelo de regresión logística
- ✓ Permite projectar los datos en dimensiones bajas que resulta en ayudas en visualización

# KNN VS LR VS LDA VS QDA

*Esperamos que:*

- ✓ KNN tenga un buen desempeño cuando hay líneas de decisión complejas, y  $n$  es suficientemente grande
- ✓ La regresión logística y LDA funcionan bien cuando la línea de decisión linear
  - ✓ LDA extiende mejor su comportamiento a problemas de multi-clase
  - ✓ LDS es más estable durante el proceso de estimación
  - ✓ La regresión logística es más robusta en presencia de outliers
- ✓ QDA es bueno cuando las líneas de decisión tienen un comportamiento cuadrático, y  $n$  es moderadamente grande

# Naïve Bayes

Es un **clasificador** simple basado en el teorema de Bayes:

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

- $P(A|B)$  : Probabilidad de A dado B.
- $P(A)$  : Probabilidad de A.
- $P(B)$  : Probabilidad de B.

Tabla de  
frecuencia

Tabla de  
probabilidades

**Naïve Bayes**  
(prob. conjunta)

Clase de mayor  
probabilidad

# Supuesto ingenuo

$p(C_k|\vec{x}) = p(C_k|x_1, \dots, x_n)$  for  $k = 1, \dots, K$  con  $n$  atributos y  $K$  clases.

$$p(C_k|\vec{x}) = \frac{p(\vec{x}|C_k)p(C_k)}{p(\vec{x})} = \frac{p(x_1, \dots, x_n|C_k)p(C_k)}{p(x_1, \dots, x_n)}$$

Usando la regla de la cadena el numerador quedaría:

$$p(x_1, \dots, x_n|C_k) = p(x_1|x_2, \dots, x_n, C_k)p(x_2|x_3, \dots, x_n, C_k)\dots p(x_{n-1}|x_n, C_k)p(x_n|C_k)$$

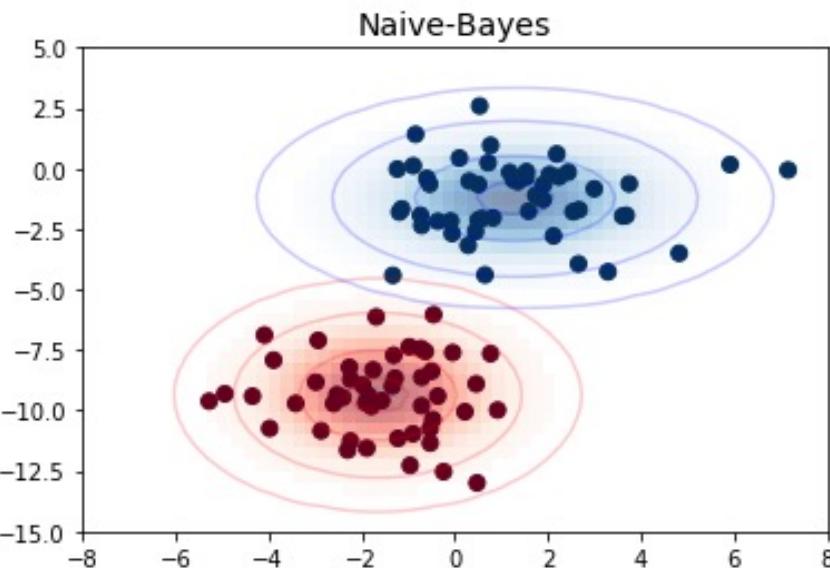
Suponiendo la  
independencia de los  
atributos:

$$p(x_i|x_{i+1}, \dots, x_n|C_k) = p(x_i|C_k) \implies p(x_1, \dots, x_n|C_k) = \prod_{i=1}^n p(x_i|C_k)$$

$$\begin{aligned} p(C_k|x_1, \dots, x_n) &\propto p(C_k, x_1, \dots, x_n) \\ &\propto p(C_k)p(x_1, \dots, x_n|C_k) \\ &\propto p(C_k) p(x_1|C_k) p(x_2|C_k)\dots p(x_n|C_k) \\ &\propto p(C_k) \prod_{i=1}^n p(x_i|C_k). \end{aligned}$$

# Naive-Bayes (Bayes ingenuo)

- ✓ Técnica de *clasificación* basada en el **teorema de Bayes** con el supuesto de independencia en los predictores (ingenuo)
- ✓ Gracias a su supuesto ingenuo, es fácil de construir y funciona rápidamente para muchos datos en múltiples dimensiones
- ✓ Modelo *generativo*, ya que especifica una distribución hipotética que genera los datos.



$$P(\text{Clase}|x) = \frac{P(x|\text{Clase}) * P(\text{Clase})}{P(x)}$$

Datos	Car. 1	Car. 2	Car. 3	Salida
P1	1	2	1	Clase 1
P2	2	2	2	Clase 1
P3	1	1	2	Clase 2
P4	2	1	2	Clase 2

$$P(\text{Clase 1}) = 0.5$$

$$P(\text{Clase 2}) = 0.5$$

Clase 1	Valor 1	Valor 2
Car. 1	1	1
Car. 2	0	2
Car. 3	1	1

Clase 2	Valor 1	Valor 2
Car. 1	1	1
Car. 2	2	0
Car. 3	0	2

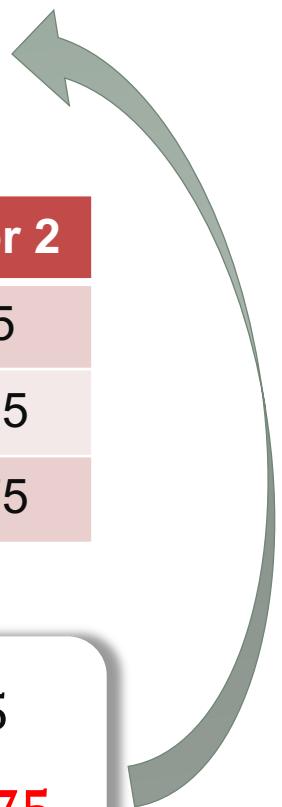
Tabla de frecuencias para cada clase

Clase 1	Valor 1	Valor 2
Car. 1	0.5	0.5
Car. 2	0.25	0.75
Car. 3	0.5	0.5

Clase 2	Valor 1	Valor 2
Car. 1	0.5	0.5
Car. 2	0.75	0.25
Car. 3	0.25	0.75

Tabla de probabilidades para cada clase

Datos	Car. 1	Car. 2	Car. 3	Salida
P1	1	2	1	Clase 1
P2	2	2	2	Clase 1
P3	1	1	2	Clase 2
P4	2	1	2	Clase 2
P5	1	1	1	???



Clase 1	Valor 1	Valor 2
Car. 1	0.5	0.5
Car. 2	0.25	0.75
Car. 3	0.5	0.5

Clase 2	Valor 1	Valor 2
Car. 1	0.5	0.5
Car. 2	0.75	0.25
Car. 3	0.25	0.75

$$P(Clase1|P5) = 0.5 * (0.5 * 0.25 * 0.5) = 0.03125$$

$$P(Clase2|P5) = 0.5 * (0.5 * 0.75 * 0.25) = \mathbf{0.046875}$$

# ¿Cuándo usar Naive Bayes?

**Teniendo en cuenta el supuesto de independencia, suele ser usado como un modelo inicial en la clasificación.**

## Ventajas

- Rápido en el entrenamiento y la predicción
- Predicciones probabilísticas fácilmente interpretables
- Tiene pocos parámetros para modificar (hiper-parámetros)

## Especialmente bueno cuando:

- El supuesto ingenuo se cumple (raro en la práctica)
- Para categorías bien separadas
- Para datos con muchas variables (alta dimensionalidad).

# ¿Cuándo NO usar Naive Bayes?

*Debido a los supuestos ingenuos es importante tomar los resultados con cuidado, en especial las probabilidades.*

## Desventajas

- ✓ Se considera un mal estimador, por lo que las probabilidades resultantes no deben ser consideradas adecuadas
- ✓ El supuesto de independencia es una desventaja para pocos datos con variables relacionadas
- ✓ Si aparece una nueva categoría que no esté en el entrenamiento, se le asigna probabilidad CERO y no podrá hacer predicciones al respecto. Se puede solucionar con suavización, como la estimación de Laplace.

# ¿Por qué es Eficiente?

✓ Se ha notado que incluso cuando las relaciones entre los atributos son claras (violando el supuesto ingenuo), la clasificación funciona relativamente bien.

## ✓ Desempeño

- ✓ Incluso con dependencia, si esta se distribuye equitativamente, sigue siendo el clasificador óptimo;
- ✓ Por lo anterior, la distribución de las dependencias es la que afecta la clasificación y con muchas variables se suelen cancelar los efectos.

## ✓ Velocidad

- ✓ Converge a su exactitud asintótica en el orden de  $\log(n)$  dimensiones, lo que lo hace mucho más veloz que otras técnicas como la regresión logística;
- ✓ También se ha estudiado que para pocos ejemplos supera la regresión logística, a pesar que esta es mejor para muchos datos.

**Inspira  
Crea  
Transforma**