

CLASIFICADORES

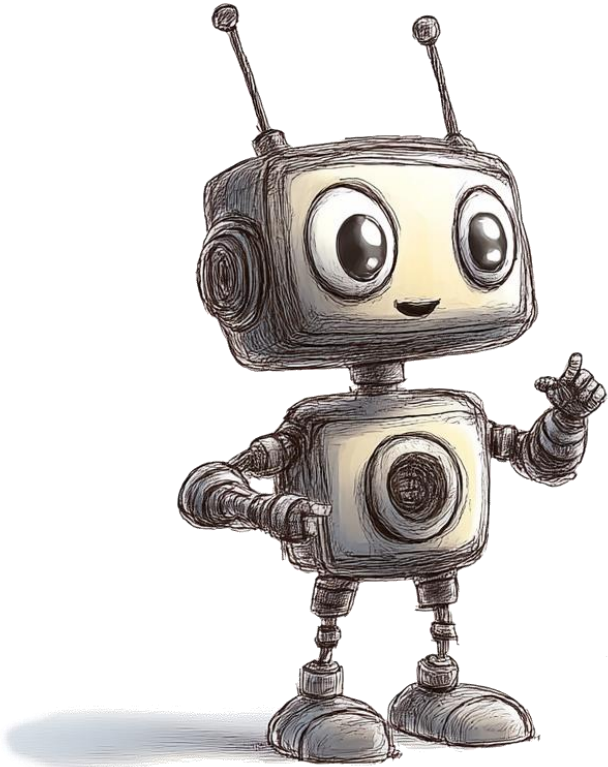


Inteligencia Artificial

CEIA - FIUBA

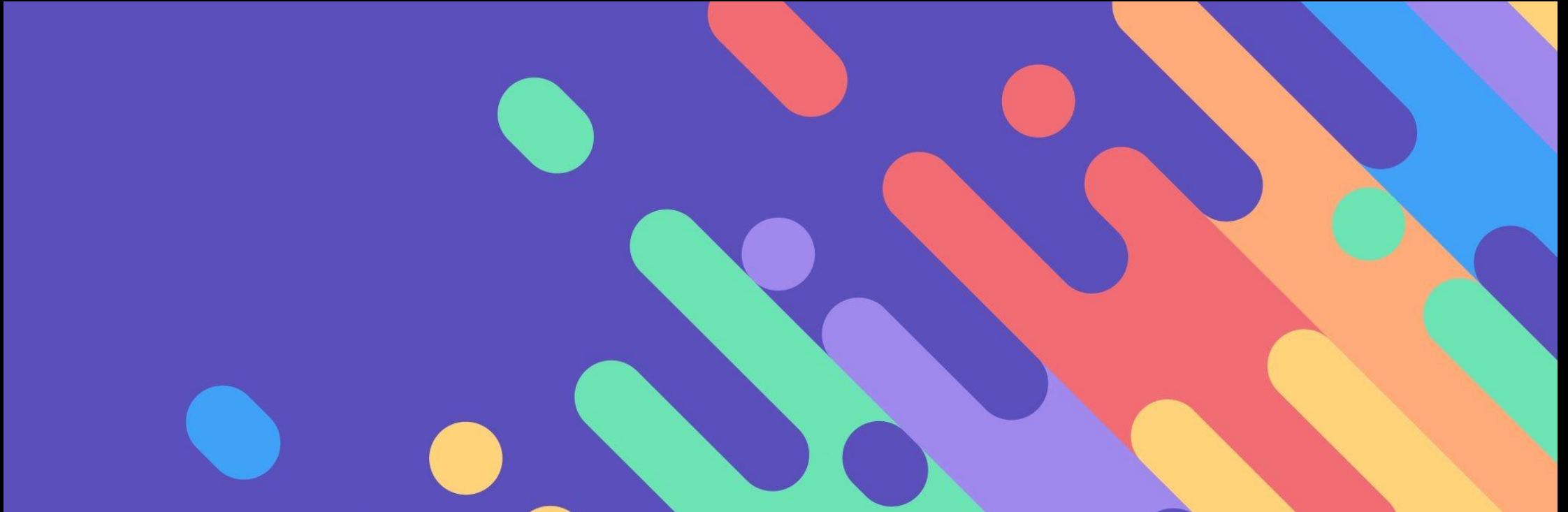
Dr. Ing. Facundo Adrián
Lucianna

CLASIFICACIÓN



Los temas que veremos en este video son:

- Definición de clasificación
- Regresión Logística
 - Definición
 - Ajuste
 - Regresión logística multiclase

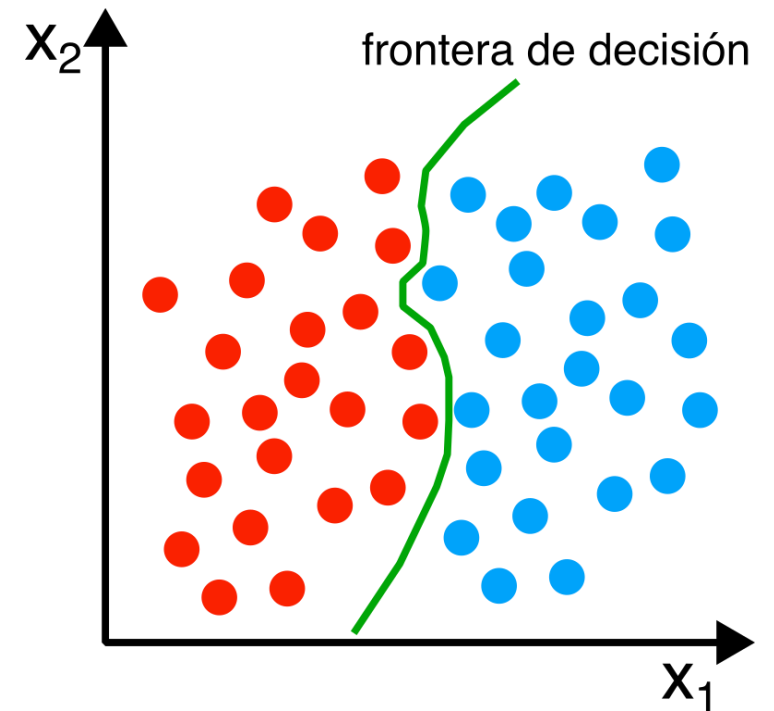


CLASIFICACIÓN

CLASIFICACIÓN

Es más común encontrarnos con problemas de clasificación que con problemas de regresión:

- Una persona llega a una guardia con un conjunto de síntomas, y debe ser asignada a una de tres condiciones médicas.
- Un servicio de banca online debe determinar si una transacción es fraudulenta o no, utilizando información como la dirección IP, historial de transacciones, etc.
- A partir de la secuencia de ADN de varios pacientes con y sin una enfermedad, un genetista debe identificar qué mutaciones de ADN tienen un efecto nocivo relacionado con la enfermedad.



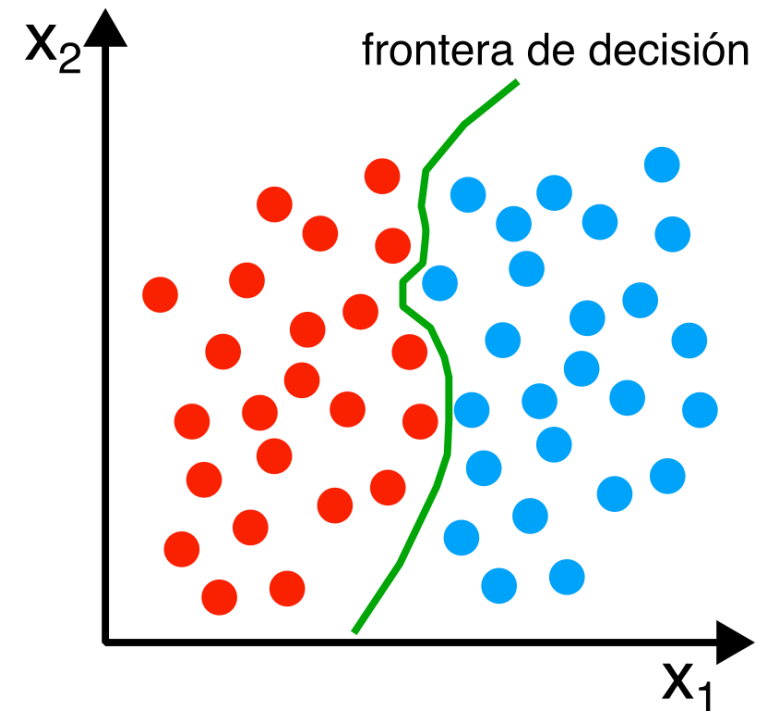
CLASIFICACIÓN

La regresión y la clasificación son problemas muy similares entre sí. En ambos buscamos predecir una variable, pero la diferencia radica en que la regresión predice una variable **numérica** y la **clasificación** una **categorica**.

¿Por qué no usar regresión para predecir respuestas cualitativas?

Tomemos el ejemplo de los pacientes que llegan a la guardia. Supongamos que hay tres diagnósticos:

- **ACV: Accidente cerebrovascular**
- **Sobredosis**
- **Ataques epilépticos**



CLASIFICACIÓN

Realizamos la siguiente codificación:

- **ACV**: 1
- **Sobredosis**: 2
- **Ataques epilépticos**: 3

Aplicamos un modelo de regresión lineal para predecir en base a los atributos del paciente.

El problema con esto es que la codificación implica un orden en los resultados, poniendo a "**sobredosis**" entre "**ACV**" y "**ataques epilépticos**", y además, la distancia entre "**ACV**" y "**sobredosis**" es la misma que entre "**sobredosis**" y "**ataques epilépticos**".

CLASIFICACIÓN

Pero podríamos haber elegido la siguiente codificación:

- Ataques epilépticos: 1
- ACV: 2
- Sobredosis: 3

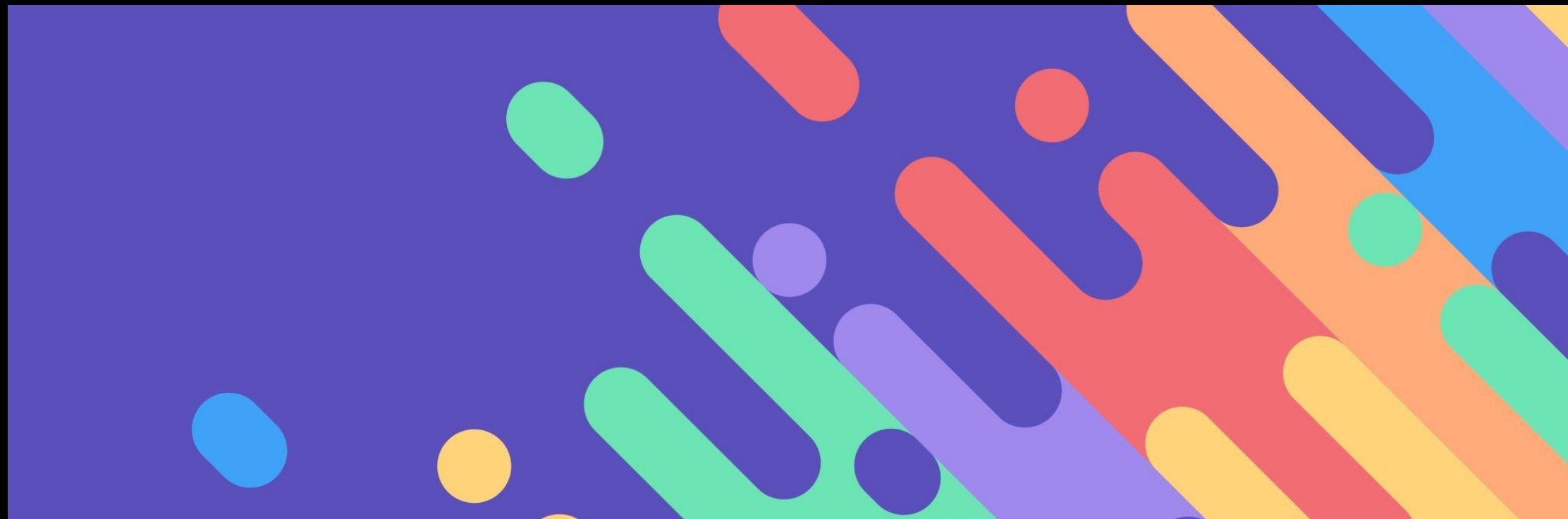
Esto nos da una relación totalmente diferente. Cada una de estas codificaciones produciría modelos lineales distintos que, en última instancia, conducirían a diferentes conjuntos de coeficientes sobre las observaciones de prueba.

Si el target es una **variable categórica ordinal**, el orden tiene sentido, y en este caso se encuentra en un área gris la elección entre modelos de clasificación y regresión.

CLASIFICACIÓN

En el caso de respuestas booleanas, por ejemplo, si una persona tiene **ACV** (igual a 1) o no (igual a 0), podemos mostrar que un modelo de regresión lineal es, de hecho, una **estimación de la probabilidad** de tener **ACV** dado un conjunto de entradas:

$$P(ACV = 1|X) = b + W^T X$$



REGRESIÓN LOGÍSTICA

REGRESIÓN LOGÍSTICA

Lo que buscamos modelar en la regresión logística no es el label y , sino la probabilidad de que y pertenezca a una clase en particular:

$$P(y = k|X)$$

En una clasificación multiclase, k puede ser 0, 1, 2, ... (también podría ser cualquier cosa, como "perro", "gato", "cebra").

En el caso de clasificación binaria:

$$P(y = 0|X)$$

$$P(y = 1|X)$$

REGRESIÓN LOGÍSTICA

En el caso de dos clases:

$$P(y = 1|X) = 1 - P(y = 0|X)$$

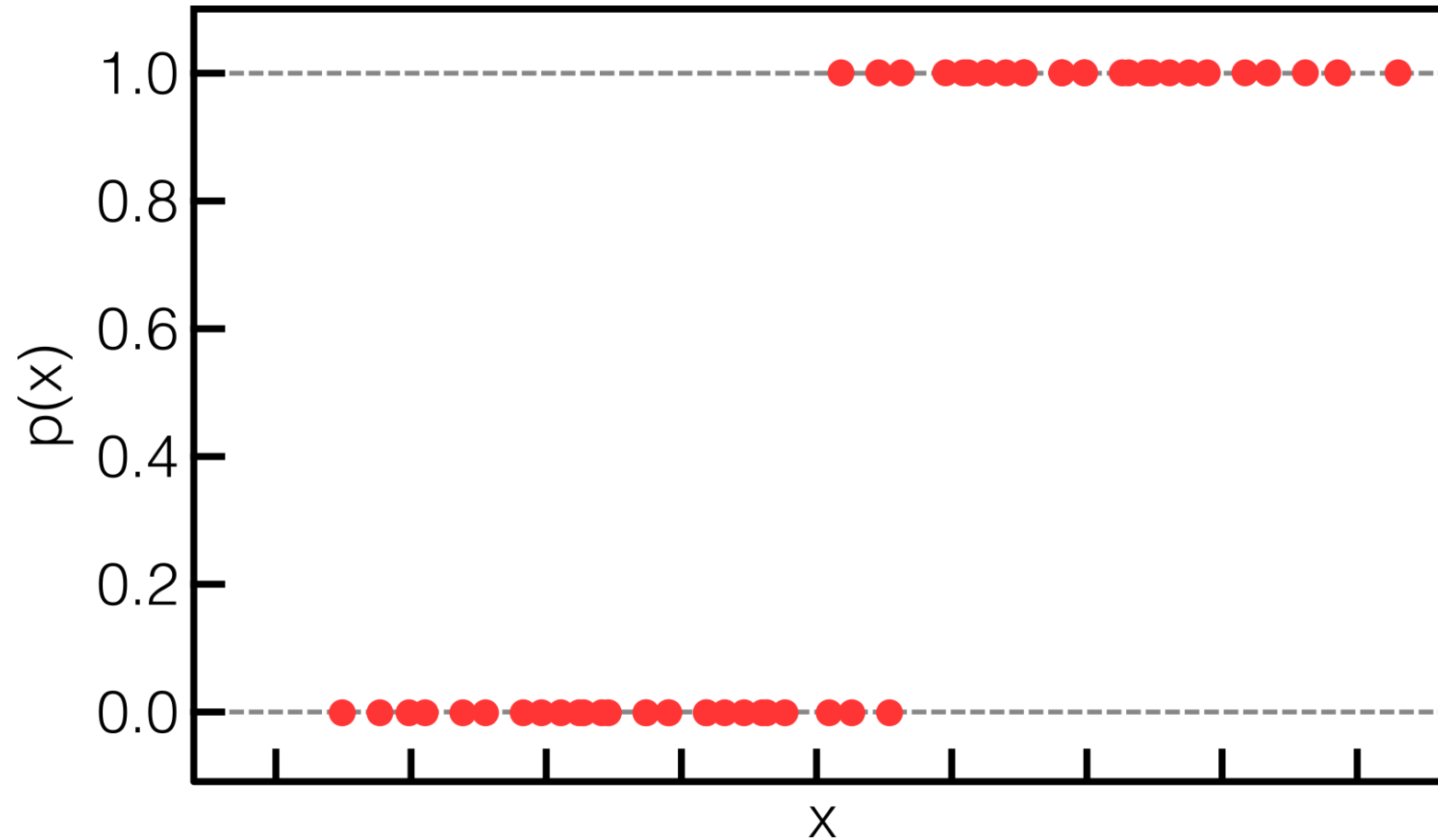
Por lo tanto, podemos simplificar la notación:

$$P(y = 1|X) = p(X)$$

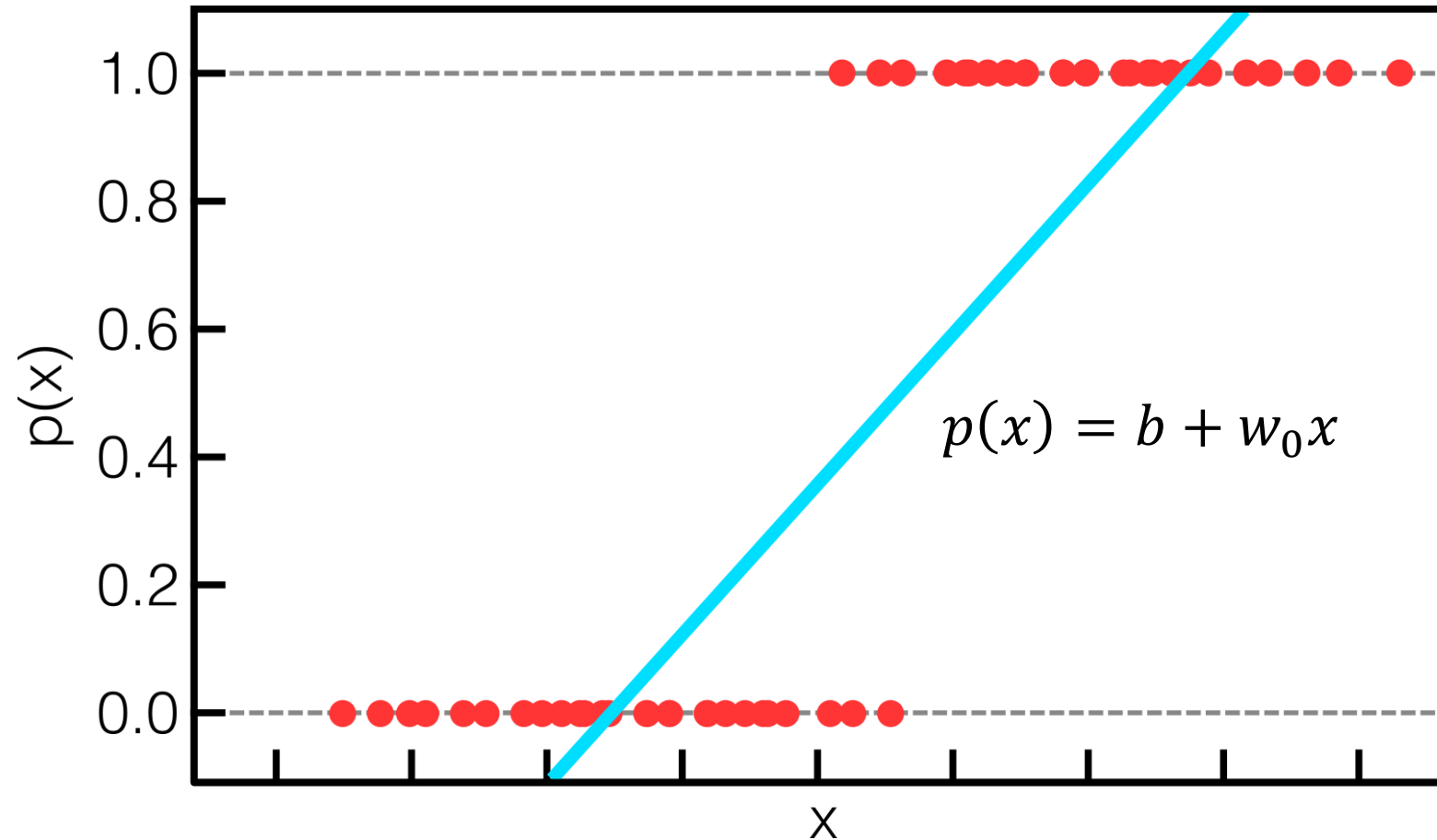
Las probabilidades son valores que van entre 0 y 1.

Para simplificar aún más, consideremos el caso de un solo atributo: $p(x)$

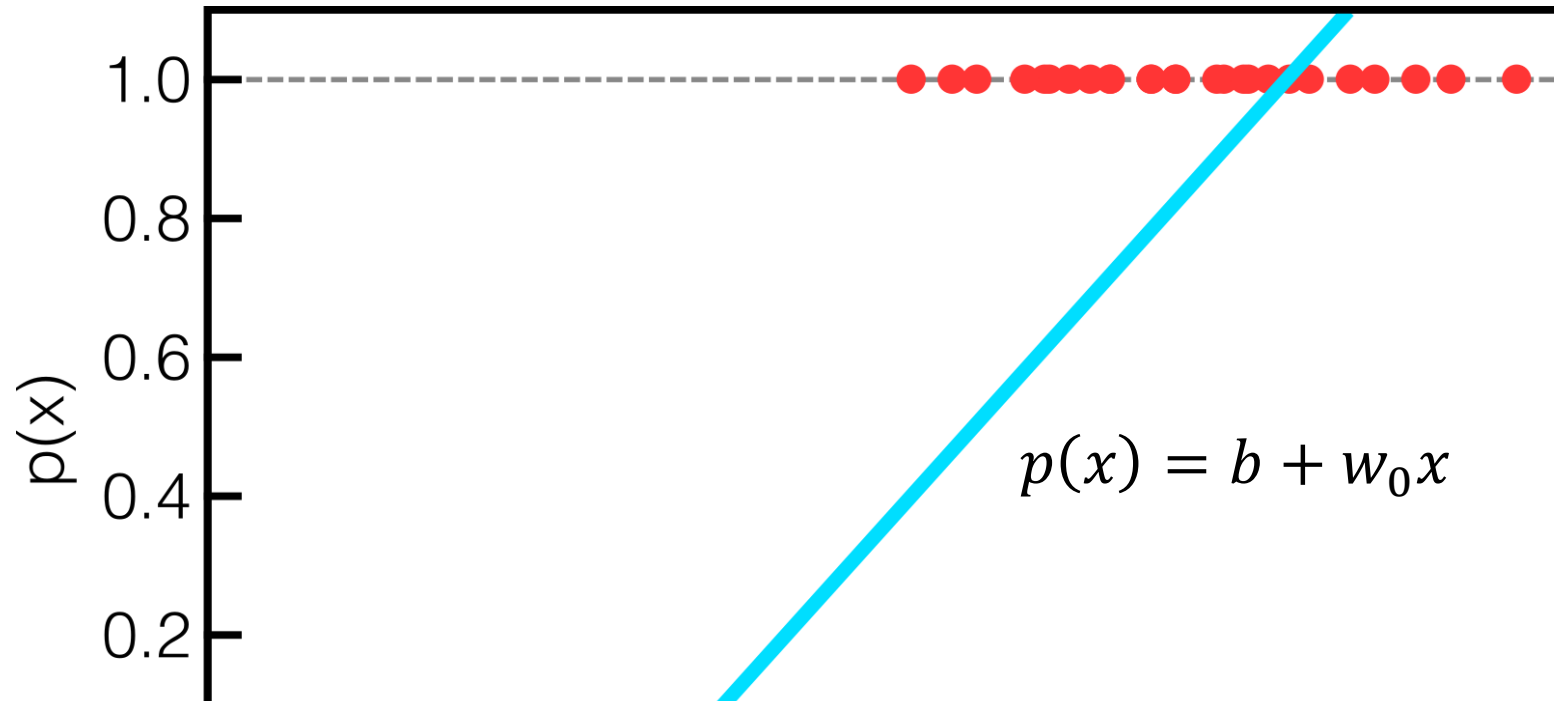
REGRESIÓN LOGÍSTICA



REGRESIÓN LOGÍSTICA



REGRESIÓN LOGÍSTICA



En la gráfica se observa el problema de predecir usando **regresión lineal**. Dada la naturaleza de la función, hay valores donde se obtienen $p(x) < 0$ o $p(x) > 1$. Esto ocurrirá con cualquier regresión que produzca valores fuera del rango entre 0 y 1.

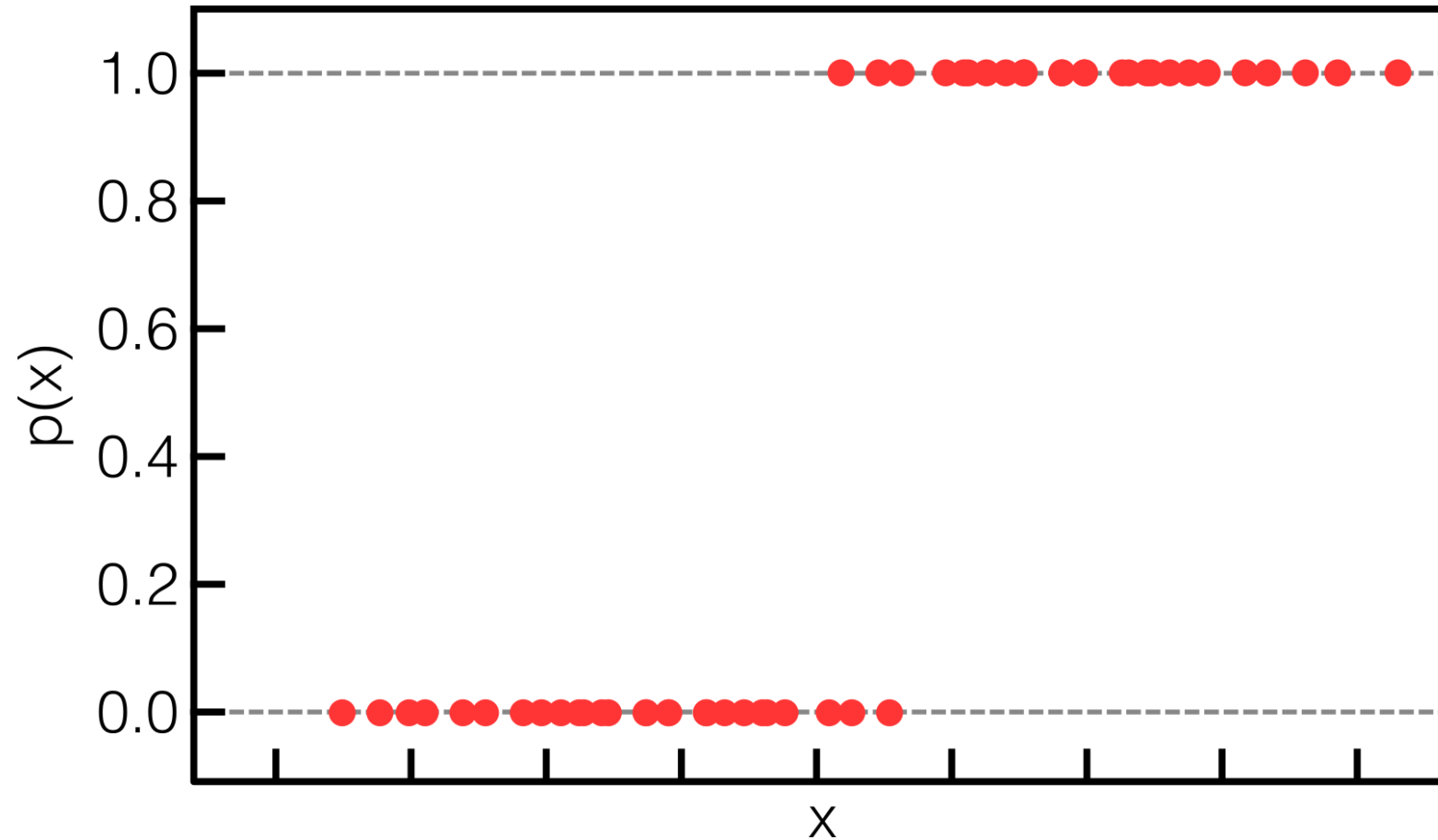
REGRESIÓN LOGÍSTICA

Para evitar esto, podemos modelar la probabilidad utilizando una función que nos asegure que **siempre tendremos valores entre 0 y 1**.

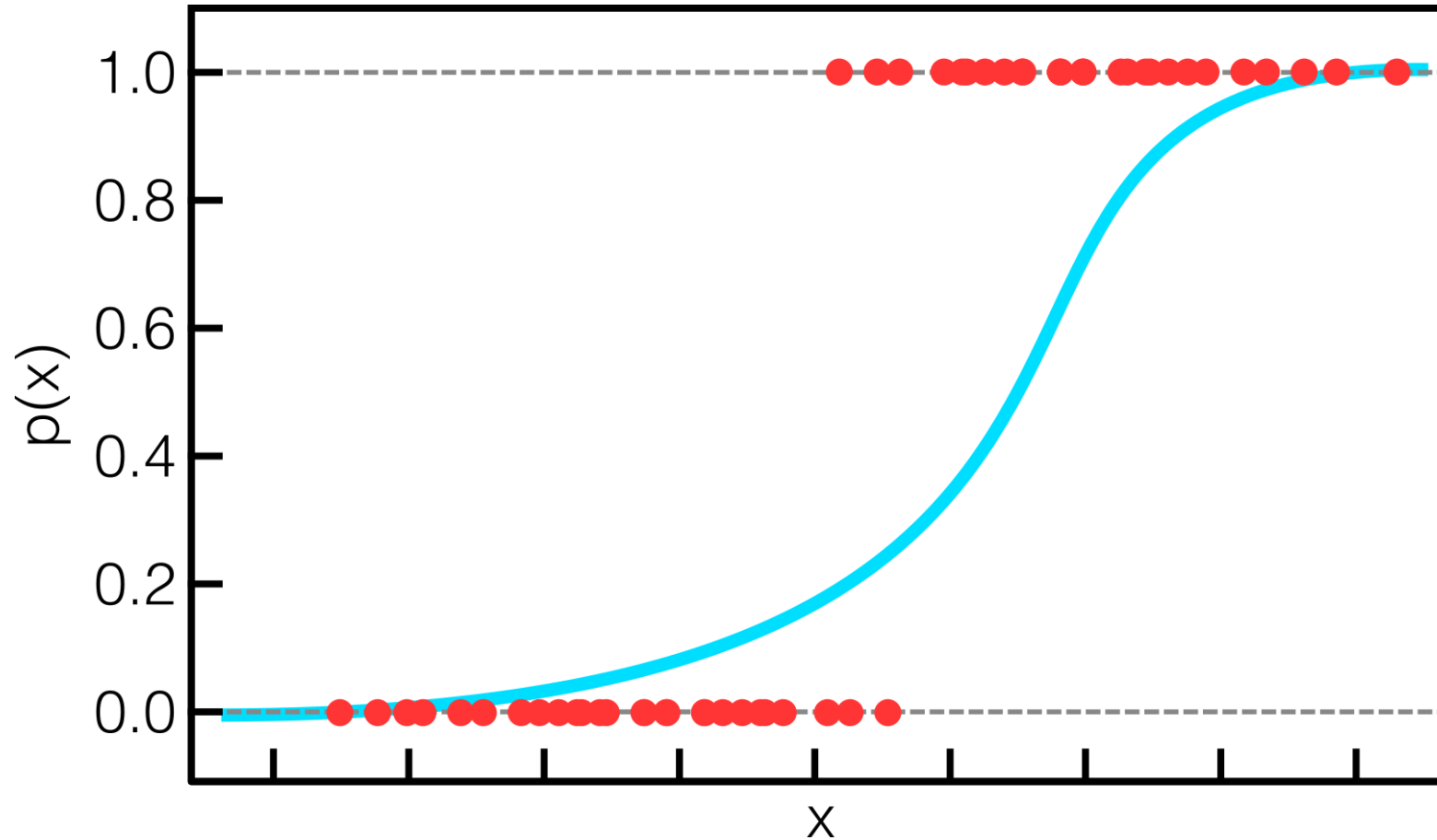
En la regresión logística, esto se resuelve utilizando una **función sigmoide**:

$$p(x) = \frac{e^{b+w_0x}}{1 + e^{b+w_0x}} = \frac{1}{1 + e^{-(b+w_0x)}}$$

REGRESIÓN LOGÍSTICA



REGRESIÓN LOGÍSTICA



REGRESIÓN LOGÍSTICA

Si manipulamos la ecuación $p(x) = \frac{e^{b+w_0x}}{1+e^{b+w_0x}}$, llegamos a:

$$\frac{p(x)}{1 - p(x)} = e^{b+w_0x}$$

Este término es conocido como **chance** (o en inglés **odds**), que es la proporción entre dos probabilidades complementarias. Estos valores pueden variar desde 0 hasta infinito.

Para entenderlo, en una semana la probabilidad de que sea sábado es 1/7, pero la "chance" es 1/6, es decir, 1 a 6 de que sea sábado.

REGRESIÓN LOGÍSTICA

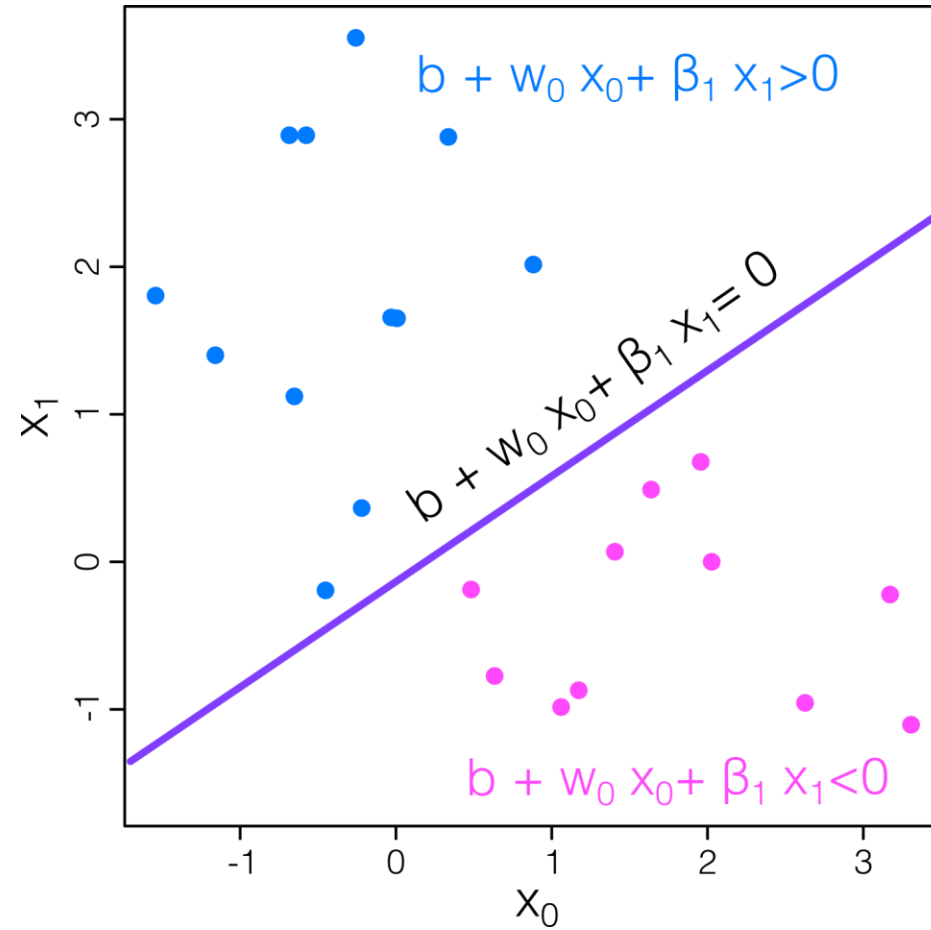
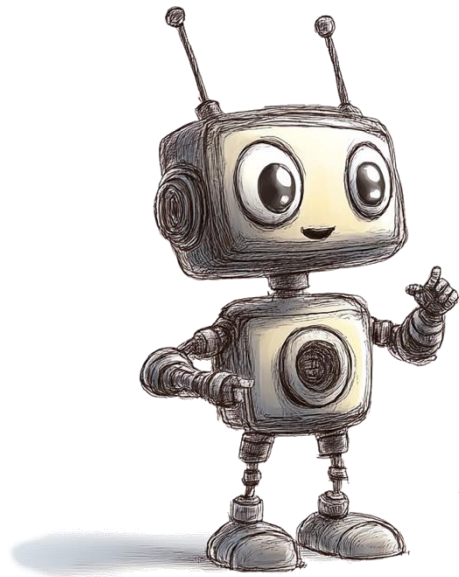
Si aplicamos el logaritmo de ambos lados:

$$\text{logit}(p) = \ln \left(\frac{p(x)}{1 - p(x)} \right) = b + w_0 x$$

Obtenemos la función **logit**.

La función logit se utiliza para transformar variables de entrada en un rango que puede interpretarse como probabilidades. En la regresión logística, esta es una relación lineal.

REGRESIÓN LOGÍSTICA



Clasificador lineal

REGRESIÓN LOGÍSTICA - AJUSTE

Para encontrar los coeficientes (b y w_0), es decir, entrenar el modelo, lo hacemos mediante **máxima verosimilitud**.

Tratamos de encontrar b y w_0 tales que las estimaciones sean lo más cercanas a 1 para la clase positiva y lo más cercanas a 0 para la clase negativa.

REGRESIÓN LOGÍSTICA - AJUSTE

Matemáticamente, la función de verosimilitud es:

$$l(b, w_0) = \prod_{i=1}^N p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

Similar a la regresión lineal, es mejor minimizar la función **log-verosimilitud** multiplicada por -1:

$$J(b, w_0) = - \sum_{i=1}^N y_i \ln(p(x_i)) + (1 - y_i) \ln(1 - p(x_i))$$

REGRESIÓN LOGÍSTICA - AJUSTE

Para encontrar el mínimo, aplicamos el gradiente:

$$\frac{\partial J}{\partial b} = 0$$
$$\frac{\partial J}{\partial w_0} = 0$$

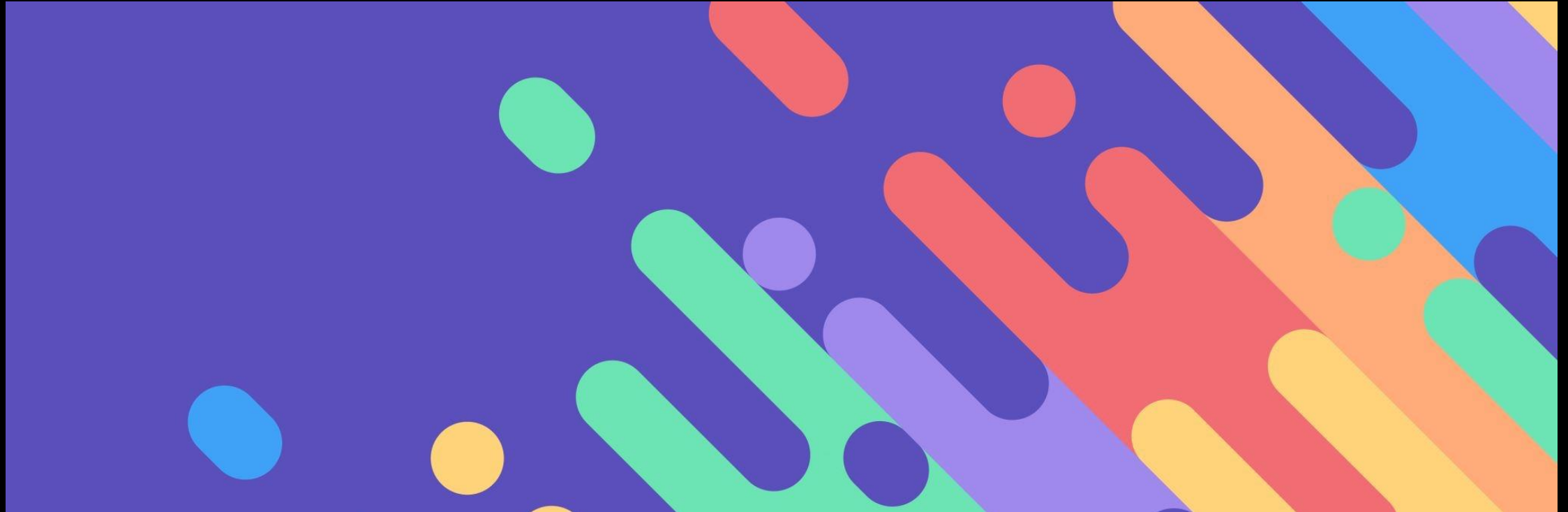
Para resolver esto, se necesitan aplicar métodos numéricos o usar gradiente descendente.

REGRESIÓN LOGÍSTICA MÚLTIPLE

Al igual que en la regresión lineal, podemos tener **más de una variable**:

$$\text{logit}(p) = \ln \left(\frac{p(\mathbf{X})}{1 - p(\mathbf{X})} \right) = b + \mathbf{W}^T \mathbf{X}$$

$$p(\mathbf{X}) = \frac{e^{b + \mathbf{W}^T \mathbf{X}}}{1 + e^{b + \mathbf{W}^T \mathbf{X}}}$$



REGRESIÓN LOGÍSTICA MULTICLASE

REGRESIÓN LOGÍSTICA MULTICLASE

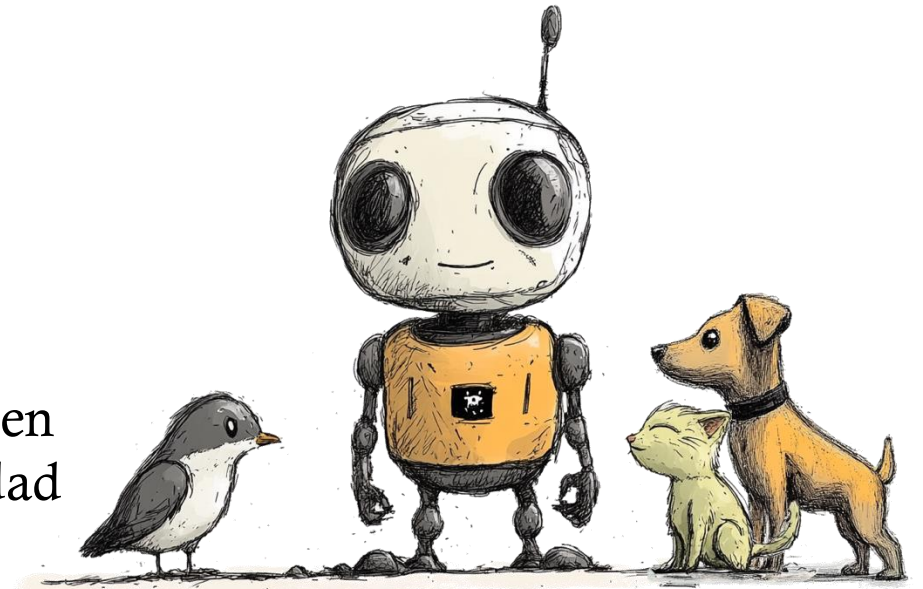
Hasta ahora hemos visto clasificadores binarios, es decir, que pueden predecir dos clases. Sin embargo, es posible extender la regresión logística para que pueda predecir tres o más clases.

Por ejemplo, si queremos clasificar entre tres clases: **perro**, **gato** y **tero**, creamos tres regresiones logísticas individuales

Para una observación particular, obtenemos:

[0.73, 0.55, 0.2]

Notamos que si sumamos los tres valores obtenemos un número mayor a uno ($0.73 + 0.55 + 0.2 = 1.48$), lo cual va en contra de lo que buscamos, que es mantener la probabilidad dentro del rango de 0 a 1.



REGRESIÓN LOGÍSTICA MULTICLASE

Si normalizamos los tres valores con respecto a la suma, recuperamos la propiedad deseada:

$$\left[\frac{0.73}{1.48}, \frac{0.55}{1.48}, \frac{0.2}{1.48} \right]$$

$$[0.49, 0.38, 0.13]$$

Nuestro clasificador combinado nos indica que, para esta observación, la clase más probable es **perro**. En problemas de clasificación multiclase, se elige la salida con el valor más alto. Observe que esta salida tiene una forma de **one-hot encoding**.

$$[1, 0, 0]$$

REGRESIÓN LOGÍSTICA MULTICLASE

Este proceso es lo que conocemos como **regresión logística multiclase**:

$$P(y = k|X) = \frac{e^{b_k + W_k^T X}}{\sum_k e^{b_{(k)} + W_{(k)}^T X}}$$

Se puede chequear que esta fórmula vuelve a la formula de la **regresión logística** si tenemos 2 clases, y se hace:

- $b = b_1 - b_0$
- $W = W_1 - W_0$

REGRESIÓN LOGÍSTICA MULTI-CLASE

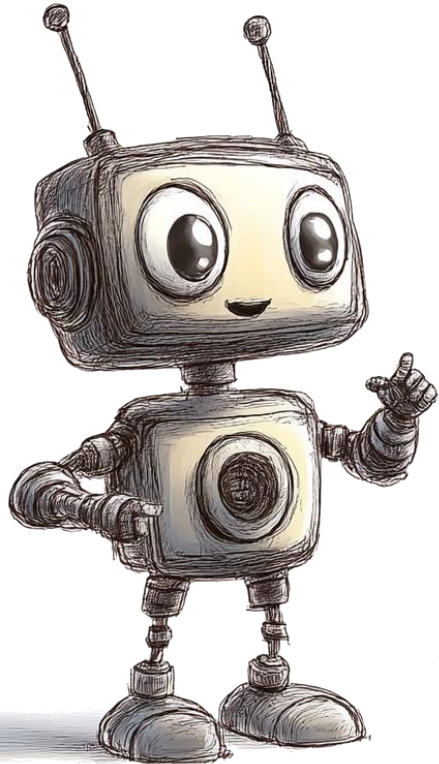
De hecho, no importa cuántas clases haya; siempre podemos elegir una clase y hacer que todos sus parámetros sean cero, **sin perder generalidad**. Esto es posible porque la probabilidad de una clase está formada por el complemento de las otras

Por convención, se elige generalmente la primera clase:

$$P(y = 0|\mathbf{X}) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{b_{(k)} + \mathbf{W}_{(k)}^T \mathbf{X}}}$$

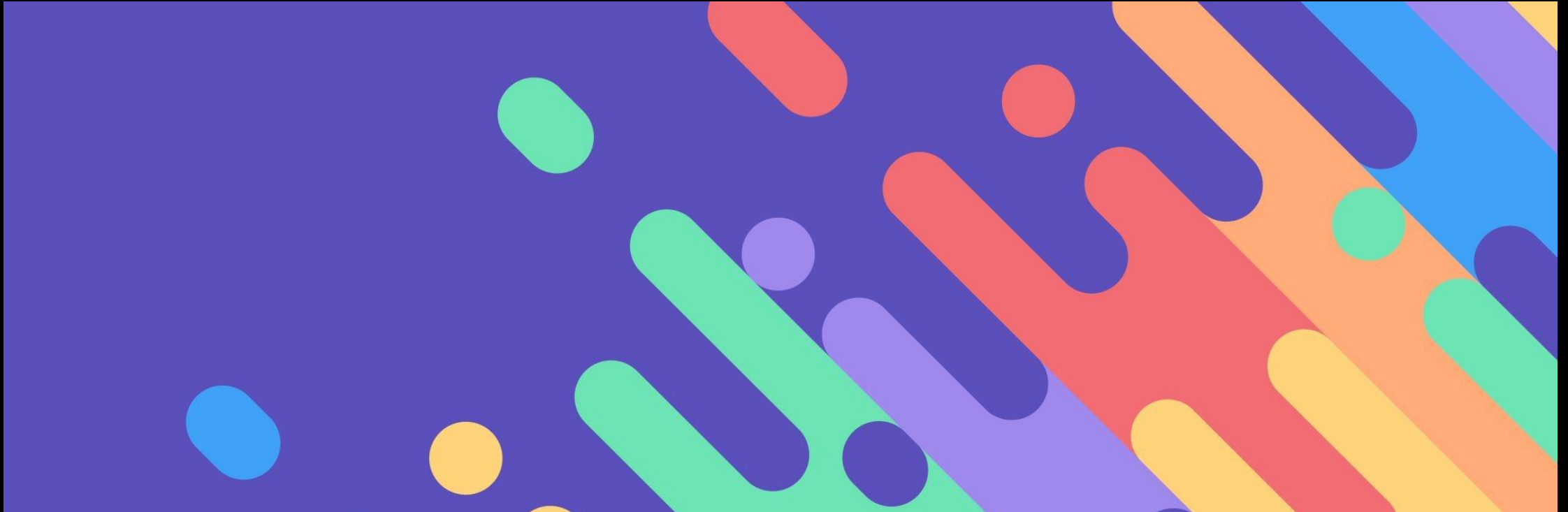
$$P(y = k|\mathbf{X}) = \frac{e^{b_k + \mathbf{W}_k^T \mathbf{X}}}{1 + \sum_{k=1}^{K-1} e^{b_{(k)} + \mathbf{W}_{(k)}^T \mathbf{X}}}$$

CLASIFICACIÓN



Los temas que veremos en este video son:

- Teorema de Bayes
- Clasificador Bayesiano ingenuo



CLASIFICADOR BAYESIANO INGENUO

CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

Este teorema es uno de los más importantes en probabilidad y uno que, hasta el día de hoy, genera divisiones filosóficas debido a sus implicancias.

Describe la probabilidad de un evento basándose en el conocimiento previo de condiciones que pueden estar relacionadas con el evento.

Por ejemplo, si se sabe que el riesgo de desarrollar problemas de salud aumenta con la edad, el teorema de Bayes permite evaluar con mayor precisión el riesgo para un individuo de una edad conocida, condicionando la probabilidad en relación con su edad, en lugar de asumir que el individuo es representativo de la población general.

CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)}$$

$$P(H|E) = \frac{P(H)P(E|H)}{P(H)P(E|H) + P(!H)P(E|!H)}$$

Este nuevo valor es lo que se conoce como la probabilidad a posteriori, es decir, la creencia sobre la hipótesis luego de observar la evidencia.

CLASIFICADOR BAYESIANO INGENUO

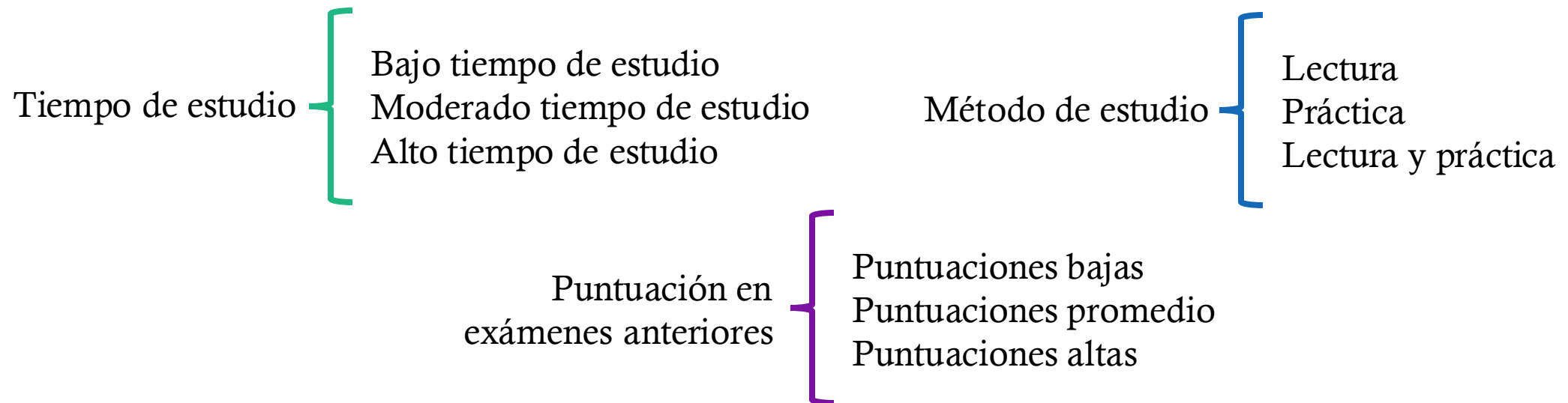
Una de las aplicaciones de este teorema es el **clasificador bayesiano ingenuo**.

Este clasificador utiliza la probabilidad de observar atributos, dado un resultado, para estimar la probabilidad de observar un resultado y , dado un conjunto de atributos.

CLASIFICADOR BAYESIANO INGENUO

Para entender su funcionamiento, usemos un caso de ejemplo:

Queremos construir un clasificador **que prediga si un estudiante va a aprobar un examen**, dado:



CLASIFICADOR BAYESIANO INGENUO

Tiempo de estudio	Método de estudio	Puntuación	Resultado
Bajo	Lectura	Bajo	Desaprobó
Bajo	Práctica	Alta	Aprobó
Moderado	Lectura y Práctica	Promedio	Aprobó
Alto	Lectura y Práctica	Alta	Aprobó
Alto	Lectura	Alta	Desaprobó
Bajo	Lectura y Práctica	Baja	Desaprobó
Alto	Práctica	Alta	Aprobó
Moderado	Lectura	Alta	Aprobó
Moderado	Lectura y Práctica	Promedio	Aprobó
Moderado	Práctica	Bajo	Desaprobó

CLASIFICADOR BAYESIANO INGENUO

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Tiempo de estudio	Bajo	1	2
	Moderado	3	1
	Alto	2	1

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Método de estudio	Lectura	1	2
	Practica	2	1
	L y P	3	1

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Puntuación	Bajo	0	3
	Promedio	2	0
	Alto	4	1

CLASIFICADOR BAYESIANO INGENUO

$P(E)$

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Tiempo de estudio	Bajo	1	2
	Moderado	3	1
	Alto	2	1

$P(H)$

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Método de estudio	Lectura	1	2
	Practica	2	1
	L y P	3	1

$P(E)$

$P(E)$

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Puntuación	Bajo	0	3
	Promedio	2	0
	Alto	4	1

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)}$$

CLASIFICADOR BAYESIANO INGENUO

$P(H)$

Aplicamos el teorema para cada atributo asumiendo que son independientes entre sí (por lo que multiplicamos las probabilidades). Este es el supuesto *ingenuo* que hacemos.

$P(E)$

estudio	Moderado	3	1
	Alto	2	1

$P(E)$

estudio	Practica	2	1
	L y P	3	1

$P(E)$

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Puntuación	Bajo	0	3
	Promedio	2	0
	Alto	4	1

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)}$$

CLASIFICADOR BAYESIANO INGENUO

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Tiempo de estudio	Bajo	1	2	3
	Moderado	3	1	4
	Alto	2	1	3
		6	4	10

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Método de estudio	Lectura	1	2	3
	Practica	2	1	3
	L y P	3	1	4
		6	4	10

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Puntuación	Bajo	0	3	3
	Promedio	2	0	2
	Alto	4	1	5
		6	4	10

CLASIFICADOR BAYESIANO INGENUO

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Tiempo de estudio	Bajo	1/6	1/2	3/10
	Moderado	1/2	1/4	2/5
	Alto	1/3	1/4	3/10
		3/5	2/5	10

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Método de estudio	Lectura	1/6	1/2	3/10
	Practica	1/3	1/4	3/10
	L y P	1/2	1/4	2/5
		3/5	2/5	10

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Puntuación	Bajo	0	3/4	3/10
	Promedio	1/3	0	1/5
	Alto	2/3	1/4	1/2
		3/5	2/5	10

CLASIFICADOR BAYESIANO INGENUO

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Tiempo de estudio	Bajo	1/6	1/2	3/10
	Moderado	1/2	1/4	2/5
	Alto	1/3	1/4	3/10
		3/5	2/5	10

$$P(H) = P(\text{Aprobo}) = 3/5 = 0.6$$

$$P(!H) = P(\text{Desaprobo}) = 2/5 = 0.4$$

$$P(E) = P(\text{Bajo tiempo de estudio}) = 3/10 = 0.3$$

$$P(E|H) = P(\text{Bajo}|\text{Aprobo}) = 1/6 = 0.17$$

$$P(E|!H) = P(\text{Bajo}|\text{Desaprobo}) = 1/2 = 0.5$$

Apliquemos el teorema de Bayes:

$$P(H|E) = P(\text{Aprobo}|\text{Bajo}) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(\text{Bajo}|\text{Aprobo})P(\text{Aprobo})}{P(\text{Bajo tiempo de estudio})} = \frac{0.17 * 0.6}{0.3} = 0.34$$

$$P(!H|E) = P(\text{Desaprobo}|\text{Bajo}) = \frac{P(E|!H)P(!H)}{P(E)} = \frac{P(\text{Bajo}|\text{Desaprobo})P(\text{Desaprobo})}{P(\text{Bajo tiempo de estudio})} = \frac{0.5 * 0.4}{0.3} = 0.67$$

CLASIFICADOR BAYESIANO INGENUO

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Tiempo de estudio	Bajo	1/6	1/2	3/10
	Moderado	1/2	1/4	2/5
	Alto	1/3	1/4	3/10

$$P(H) = P(\text{Aprobo}) = 3/5 = 0.6$$

$$P(!H) = P(\text{Desaprobo}) = 2/5 = 0.4$$

$$P(E) = P(\text{Bajo tiempo de estudio}) = 3/10 = 0.3$$

$$P(E|H) = P(\text{Bajo}|\text{Aprobo}) = 1/6 = 0.17$$

$$P(E|!H) = P(\text{Bajo}|\text{Desaprobo}) = 1/2 = 0.5$$

Con esto, podemos concluir que es más probable que un alumno desaprobe si estudia poco.

Apliquemos el teorema de Bayes:

$$P(H|E) = P(\text{Aprobo}|\text{Bajo}) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(\text{Bajo}|\text{Aprobo})P(\text{Aprobo})}{P(\text{Bajo tiempo de estudio})} = \frac{0.17 * 0.6}{0.3} = 0.34$$

$$P(!H|E) = P(\text{Desaprobo}|\text{Bajo}) = \frac{P(E|!H)P(!H)}{P(E)} = \frac{P(\text{Bajo}|\text{Desaprobo})P(\text{Desaprobo})}{P(\text{Bajo tiempo de estudio})} = \frac{0.5 * 0.4}{0.3} = 0.67$$

CLASIFICADOR BAYESIANO INGENUO

Se asume que los tres atributos son independientes entre sí:

$P(E)$ equivale a *Estudio = Alto*, *Método = L y P*, *Puntuación = Alta*

$$P(H) = P(\text{Aprueba})$$

$$P(H|E) = \frac{P(\text{Alto}|\text{Aprueba})P(\text{LyP}|\text{Aprueba})P(\text{Alta}|\text{Aprueba})P(\text{Aprueba})}{P(\text{Alto})P(\text{LyP})P(\text{Alto})}$$

$$P(H|E) = \frac{0.333 * 0.5 * 0.667 * 0.6}{0.3 * 0.4 * 0.5} = 1.11$$

CLASIFICADOR BAYESIANO INGENUO

Se asume que los tres atributos son independientes entre sí:

$P(E)$ equivale a *Estudio = Alto*, *Método = L y P*, *Puntuación = Alta*

$$P(H) = P(\text{Aprueba})$$

$$P(H|E) = \frac{\overbrace{P(\text{Alto}|\text{Aprueba})P(\text{LyP}|\text{Aprueba})P(\text{Alta}|\text{Aprueba})}^{P(E|H)} \overbrace{P(\text{Aprueba})}^{P(H)}}{\underbrace{P(\text{Alto})P(\text{LyP})P(\text{Alta})}_{P(E)}}$$
$$P(H|E) = \frac{0.333 * 0.5 * 0.667 * 0.6}{0.3 * 0.4 * 0.5} = 1.11$$

CLASIFICADOR BAYESIANO INGENUO

Veamos la hipótesis de que el alumno *desaprueba*:

$P(E)$ equivale a *Estudio = Alto*, *Método = L y P*, *Puntuación = Alta*

$$P(H) = P(\text{Desaprueba})$$

$$P(H|E) = \frac{P(\text{Alto}|\text{Aprueba})P(\text{LyP}|\text{Aprueba})P(\text{Alta}|\text{Aprueba})P(\text{Aprueba})}{P(\text{Alto})P(\text{LyP})P(\text{Alto})}$$

$$P(H|E) = \frac{0.25 * 0.25 * 0.25 * 0.4}{0.3 * 0.4 * 0.5} = 0.10$$

CLASIFICADOR BAYESIANO INGENUO

Normalizamos:

$$Suma = P(H|E) + P(!H|E) = 1.11 + 0.10 = 1.21$$

$$P(H|E) = \frac{1.11}{1.21} = 0.92$$

$$P(!H|E) = \frac{0.1}{1.21} = 0.08$$

CLASIFICADOR BAYESIANO INGENUO

Normalizamos:

$$Suma = P(H|E) + P(!H|E) = 1.11 + 0.10 = 1.21$$

$$P(H|E) = \frac{1.11}{1.21} = 0.92$$

$$P(!H|E) = \frac{0.1}{1.21} = 0.08$$

Dado que, si un alumno le dedica muchas horas, estudia practicando y leyendo, y tiene buenas notas, es muy probable que **apruebe el examen**.

CLASIFICADOR BAYESIANO INGENUO

Normalizamos:

$P(H)$

$P(H)$

$P(H)$

En el contexto del clasificador Naive Bayes, las probabilidades condicionales pueden sumar más de uno debido a la independencia asumida entre las características, lo que puede resultar en una **sobreestimación**.

Aunque una probabilidad no puede ser mayor que uno, en este contexto, este valor no invalida el resultado. Sin embargo, para interpretarlas como una distribución de probabilidad válida, se requiere **normalizar**.

Dado que, si un alumno le dedica muchas horas, estudia practicando y leyendo, y tiene buenas notas, es muy probable que **apruebe el examen**.

CLASIFICADOR BAYESIANO INGENUO

El **denominador es siempre el mismo**

Si solo nos interesa clasificar, solo se calcula el numerador:

- $P(H|E) = 0.333 * 0.5 * 0.667 * 0.6 = 0.06663$
- $P(!H|E) = 0.25 * 0.25 * 0.25 * 0.4 = 0.00625$

CLASIFICADOR BAYESIANO INGENUO

A veces, si no tenemos valores en alguna combinación, ya sea por un *dataset pequeño o falta de datos*, esto puede afectar el resultado.

Por ejemplo, si queremos ver si el alumno aprueba si *estudio mucho tiempo*, *realizo practica y lecturas*, y *venia con puntuación baja en exámenes anteriores* :

- $P(H|E) = 0.333 * 0.5 * \mathbf{0} * 0.6 = \mathbf{0}$

Podemos mitigar este problema, **sumando un valor** a cada uno de los valores en la tabla de frecuencias.

CLASIFICADOR BAYESIANO INGENUO

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Tiempo de estudio	Bajo	1+1=2	2+1=3
	Moderado	3+1=4	1+1=2
	Alto	2+1=3	1+1=2

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Método de estudio	Lectura	1+1=2	2+1=3
	Practica	2+1=3	1+1=2
	L y P	3+1=4	1+1=2

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Puntuación	Bajo	0+1=1	3+1=4
	Promedio	2+1=3	0+1=1
	Alto	4+1=5	1+1=2

CLASIFICADOR BAYESIANO INGENUO

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Tiempo de estudio	Bajo	2/9	3/7	5/16
	Moderado	4/9	2/7	6/16
	Alto	3/9	2/7	5/16
		9/16	7/16	16

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Método de estudio	Lectura	2/9	3/7	5/16
	Practica	3/9	2/7	5/16
	L y P	4/9	2/7	6/16
		9/16	7/16	16

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Puntuación	Bajo	1/9	4/7	5/16
	Promedio	3/9	1/7	4/16
	Alto	5/9	2/7	6/16
		9/16	7/16	16

CLASIFICADOR BAYESIANO INGENUO

Ahora podemos calcular:

Si queremos ver si el alumno aprueba si *estudio mucho tiempo*, *realizo practica y lecturas*, y *venia con puntuación baja en exámenes anteriores*:

- $P(H|E) = 0.333 * 0.444 * 0.111 * 0.5625 = 0.0092$
- $P(!H|E) = 0.286 * 0.286 * 0.286 * 0.4375 = 0.0102$

CLASIFICADOR BAYESIANO INGENUO

Ahora podemos calcular:

Si queremos ver si el alumno aprueba si *estudio mucho tiempo*, *realizo practica y lecturas*, y *venia con puntuación baja en exámenes anteriores*:

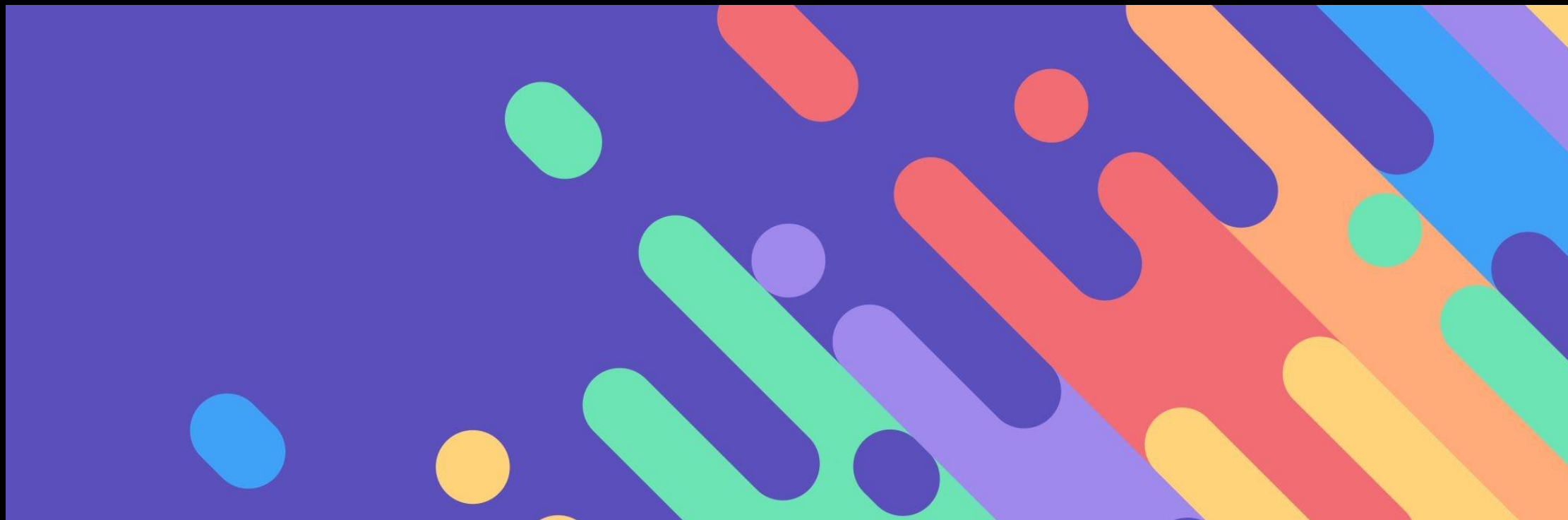
- $P(H|E) = 0.333 * 0.444 * 0.111 * 0.5625 = 0.0092$
- $P(!H|E) = 0.286 * 0.286 * 0.286 * 0.4375 = 0.0102$

Este número que agregamos es un hiperparámetro llamado α (alfa)

CLASIFICADOR BAYESIANO INGENUO

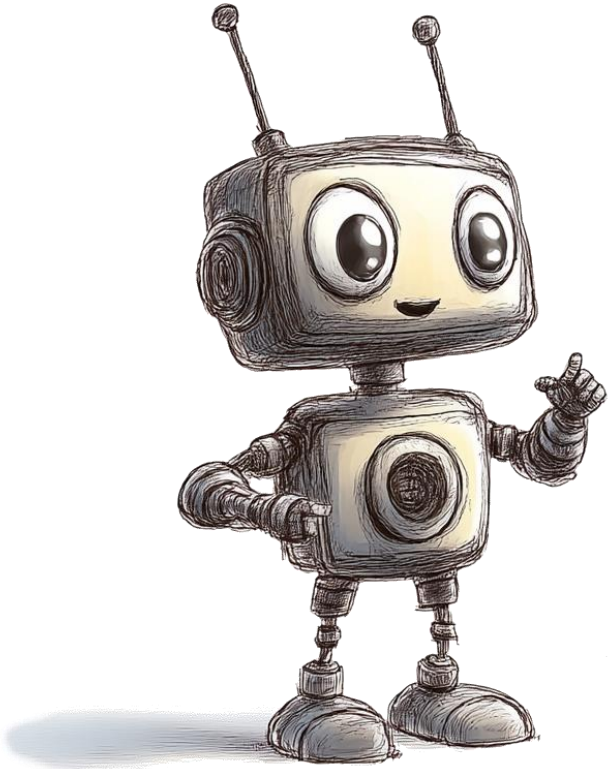
El clasificador bayesiano ingenuo funciona para **variables categóricas**. Para **variables numéricas**, podemos tratarlas de dos maneras

- Discretizarlas en contenedores o rangos.
- Asumir que siguen una distribución y usar esa distribución para calcular la probabilidad.



MÉTRICAS DE CLASIFICACIÓN

CLASIFICACIÓN

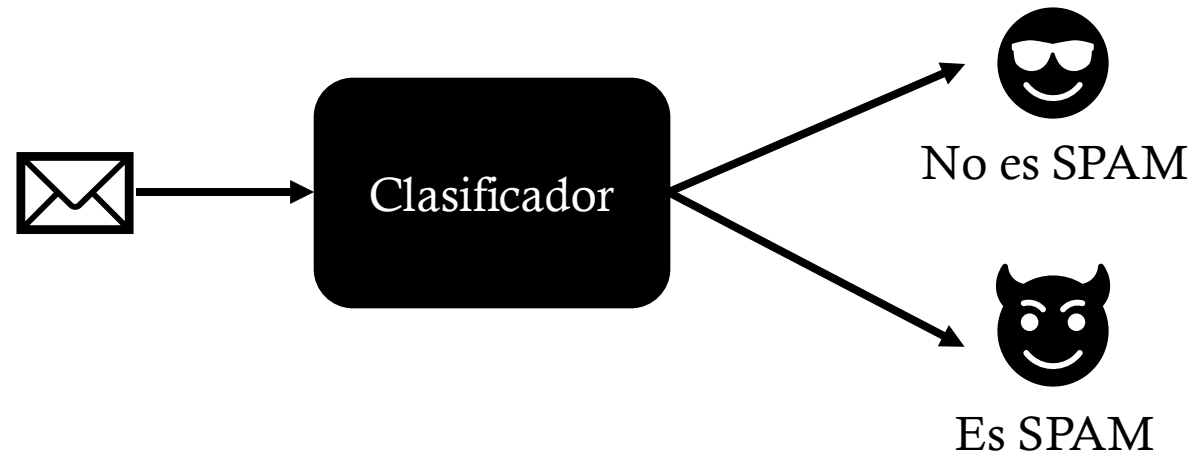
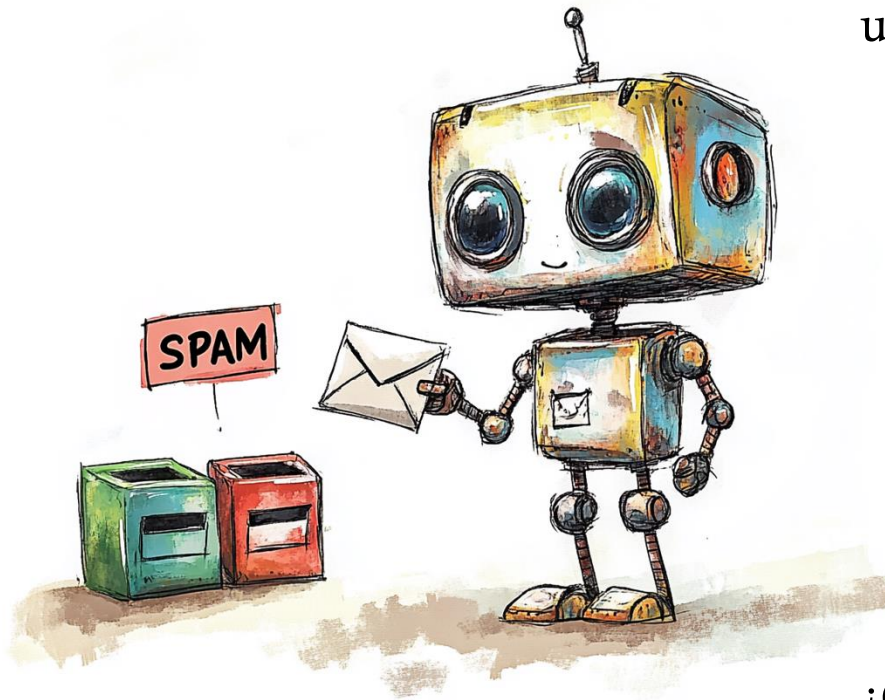


Los temas que veremos en este video son:

- Métricas de clasificación
 - Matriz de confusión
 - Sensibilidad y especificidad
 - Precisión y recuperación (Puntaje F_1)
- Curva ROC
 - Definición
 - Área bajo la curva (AUC)

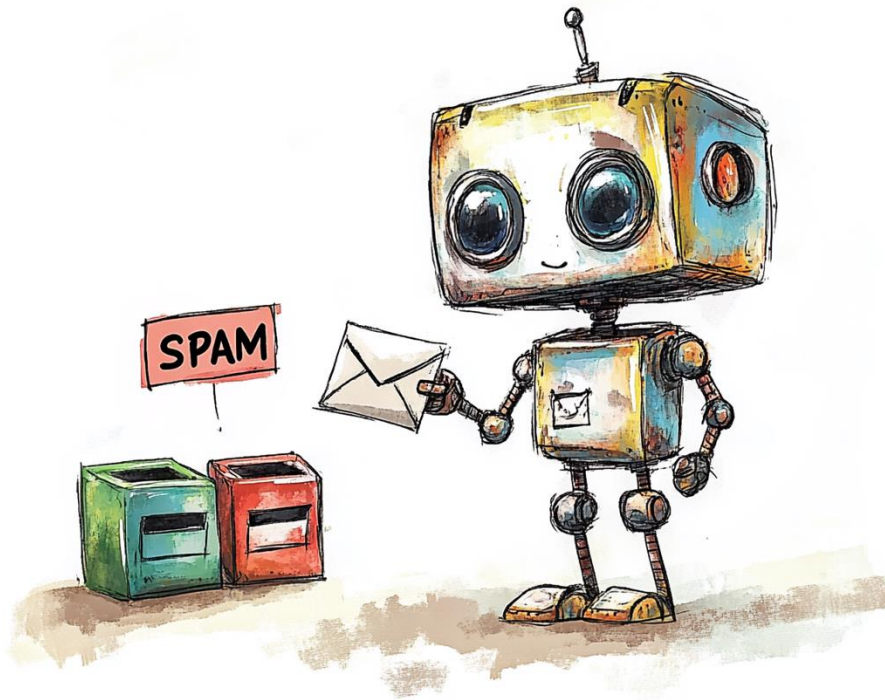
MÉTRICAS DE CLASIFICACIÓN

Supongamos que tenemos un modelo encargado de identificar si un correo es SPAM o no:



¿Cómo medimos la calidad de este clasificador? ¿Cómo sabemos si funciona bien?

MÉTRICAS DE CLASIFICACIÓN



¿Cómo medimos la calidad de este clasificador? ¿Cómo sabemos si funciona bien?





Uno podría pensar intuitivamente en la tasa de aciertos. Pero los correos **SPAM** son mucho menos frecuentes que los no **SPAM**, supongamos que tenemos una relación 1 a 1000.

Entonces, un modelo que clasifica todo como no SPAM tendrá una tasa de aciertos de:

99.9%

Entonces, ¿seguimos considerando la tasa de aciertos como una buena métrica?

MÉTRICAS DE CLASIFICACIÓN

		Valor verdadero	
			
Salida del clasificador		Verdadero positivo (TP)	Falso positivo (FP)
		Falso negativo (FN)	Verdadero negativo (TN)

Esta estructura se llama **matriz de confusión**

MÉTRICAS DE CLASIFICACIÓN

Matriz de confusión

- **Verdadero positivo (TP):** Son las observaciones que clasificamos como 1 y que realmente eran 1.
- **Verdadero negativo (TN):** Son las observaciones que clasificamos como 0 y que realmente eran 0.
- **Falso positivo (FP):** Son las observaciones que clasificamos como 1 y que realmente eran 0. Este error se llama de **tipo I**.
- **Falso negativo (FN):** Son las observaciones que clasificamos como 0 y que realmente eran 1. Este error se llama de **tipo II**.

MÉTRICAS DE CLASIFICACIÓN

Matriz de confusión

- **Sensibilidad (tasa de verdaderos positivos):** Representa la capacidad del clasificador para detectar todos los casos positivos en los datos.

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

- **Especificidad (tasa de verdaderos negativos):** Indica la capacidad del clasificador para identificar correctamente los casos negativos.

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

MÉTRICAS DE CLASIFICACIÓN

Matriz de confusión

Volviendo a nuestro clasificador que dice que todo es no *SPAM*, tendríamos:

$$\text{Sensibilidad} = 0 \quad \text{Especificidad} = 1$$

La **exactitud es la métrica** que vimos, la tasa de aciertos:

$$\text{Exactitud} = \frac{TP + TN}{P + N} = 0.999$$

Pero cuando tenemos un desbalance de clases, conviene calcular la **exactitud balanceada**:

$$\text{Exactitud balanceada} = \frac{TPR + TNR}{2} = 0.5$$

MÉTRICAS DE CLASIFICACIÓN

Matriz de confusión

Volviendo a nuestro clasificador que dice que todo es no *SPAM*, tendríamos:

$$\textit{Sensibilidad} = 0 \quad \textit{Especificidad} = 1$$

La **exactitud es la métrica** que vimos, la tasa de aciertos:

$$\textit{Exactitud} = \frac{TP + TN}{P + N} = 0.999$$

Pero cuando tenemos un desbalance de clases, conviene calcular la **exactitud balanceada**:

$$\textit{Exactitud balanceada} = \frac{TPR + TNR}{2} = 0.5$$

Nos dice que el clasificador está adivinando

MÉTRICAS DE CLASIFICACIÓN

Precisión y recuperación

Otras dos métricas muy importantes son **precisión** y **recuperación**. Estas juegan un papel crucial cuando la clase positiva tiene más importancia que la negativa:

- **Precisión**: Se refiere a la proporción de casos positivos identificados correctamente por el clasificador con respecto a todos los casos que el clasificador etiquetó como positivos.

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

- **Recuperación (Recall)**: Mide la proporción de casos positivos que el clasificador identificó correctamente con respecto a todos los casos positivos reales en los datos. En otras palabras, la recuperación indica la capacidad del clasificador para "*recuperar*" los casos positivos.

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

MÉTRICAS DE CLASIFICACIÓN

Precisión y recuperación

Veamos ejemplos donde una métrica es más importante que la otra:

- **Precisión:** *En nuestro clasificador de SPAM, esta métrica es más importante, ya que queremos que cuando el clasificador diga que es SPAM, realmente esté seguro. No queremos que el usuario pierda correos electrónicos importantes.*
- **Recuperación:** *En un clasificador de imágenes para detectar cáncer, la recuperación es más importante. Es fundamental que el modelo capture la mayor cantidad posible de casos de cáncer para garantizar que los pacientes no pierdan un diagnóstico temprano y, por lo tanto, un tratamiento oportuno. Incluso si esto significa algunos falsos positivos.*

MÉTRICAS DE CLASIFICACIÓN

Precisión y recuperación

A veces nos interesa un balance entre ambas métricas, y para ello podemos usar el **puntaje F_1** :

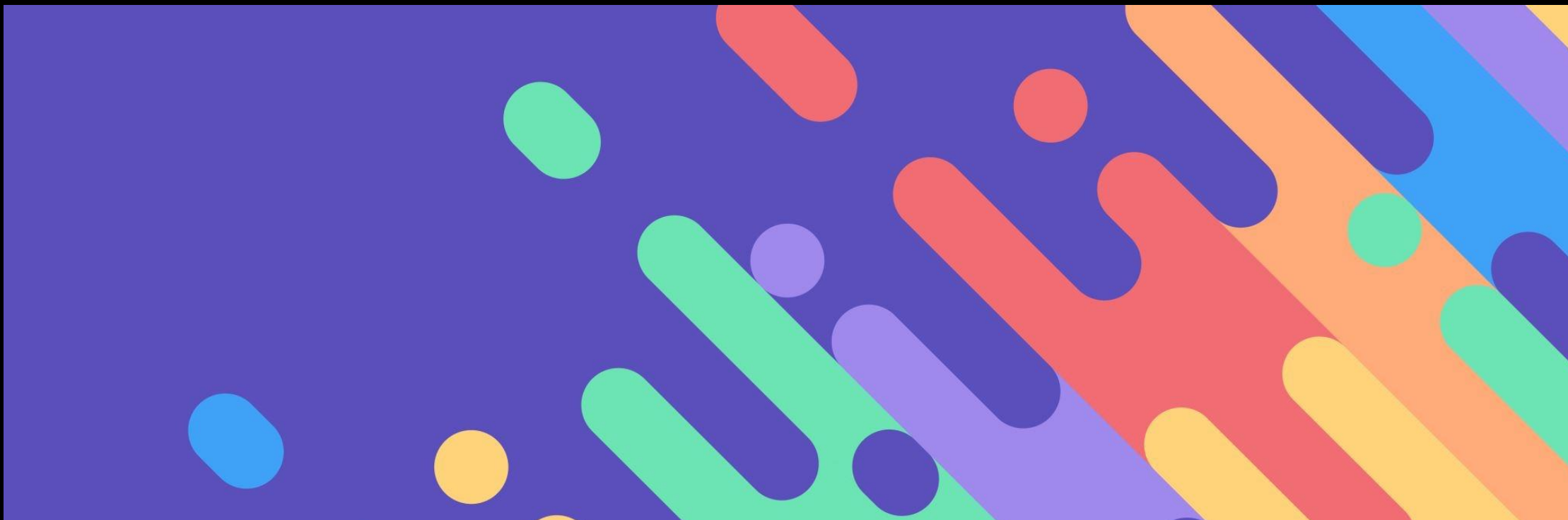
$$F_1 = 2 \frac{precision \cdot recall}{precision + recall}$$

Si queremos darle más importancia a una métrica que a la otra, podemos usar:

$$F_\beta = (1 + \beta^2) \frac{precision \cdot recall}{\beta^2 precision + recall}$$

Donde:

- Si $0 < \beta < 1$, le damos más importancia a la **precisión** (preferimos minimizar los falsos positivos).
- Si $\beta > 1$, le damos más importancia a la **recuperación** (preferimos minimizar los falsos negativos).



CURVA ROC

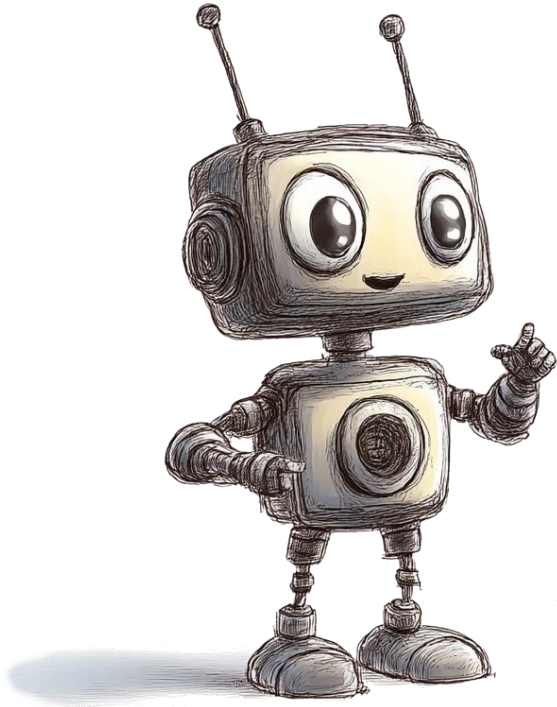
CURVA ROC

Para las métricas que vimos, siempre supusimos que nuestro clasificador nos da la salida 1 si es la clase positiva, y 0 si es negativa. Pero si tenemos una regresión logística, nos da **un valor de probabilidad de qué tan probable es que sea de la clase positiva**.

De forma intuitiva, podemos definir que si la regresión logística *nos devuelve un valor a mayor a 0.5*, lo clasificamos como clase positiva; *si es menor o igual a 0.5*, lo clasificamos como clase negativa. De ahí podemos calcular todas las métricas.

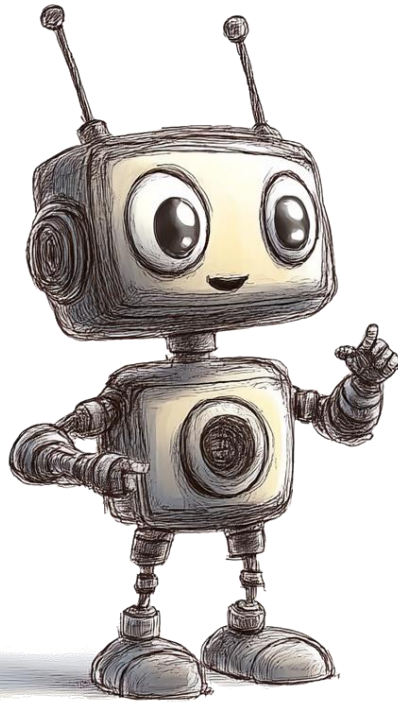
CURVA ROC

¿Por qué este valor de 0.5?



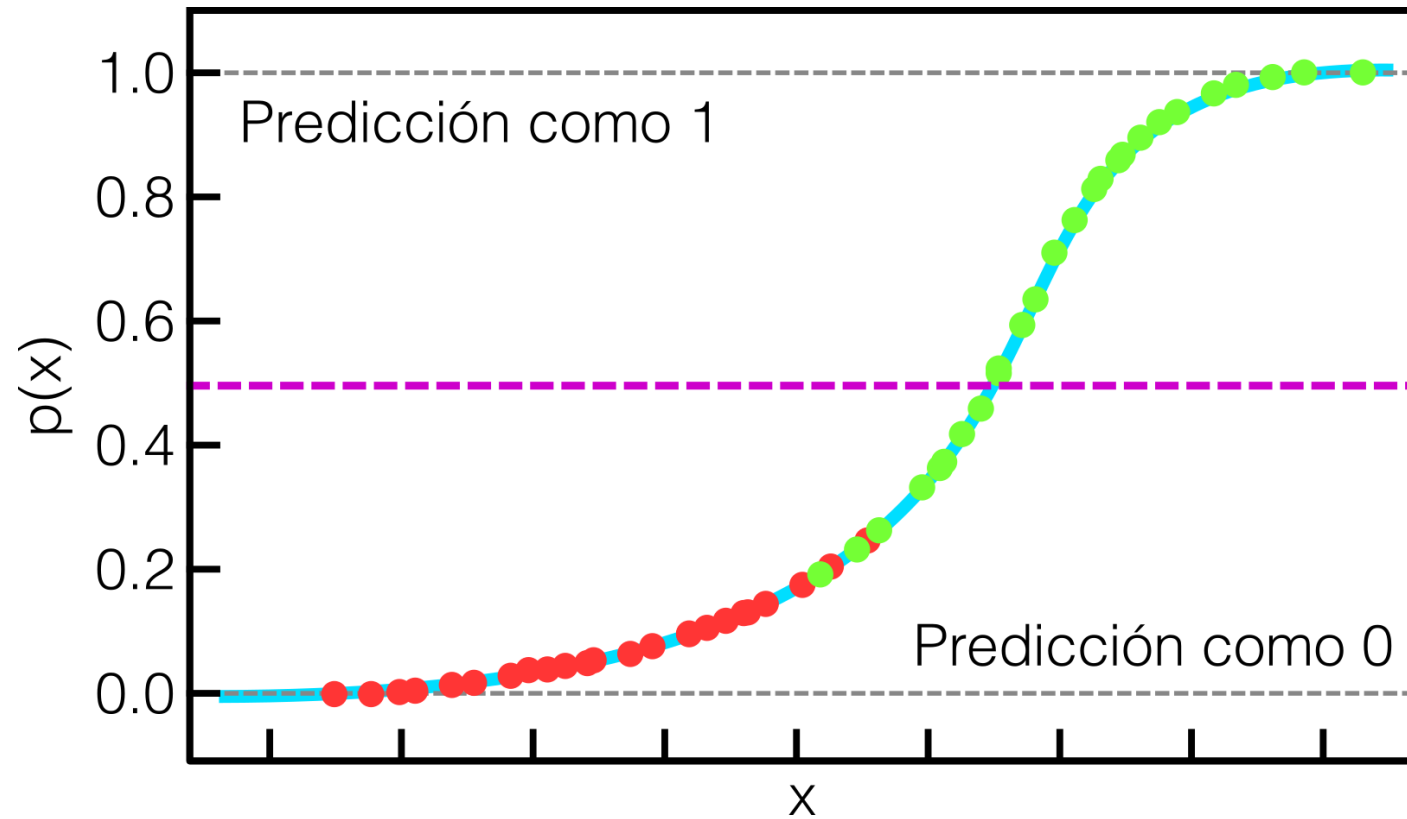
CURVA ROC

¿Por qué este valor de 0.5?

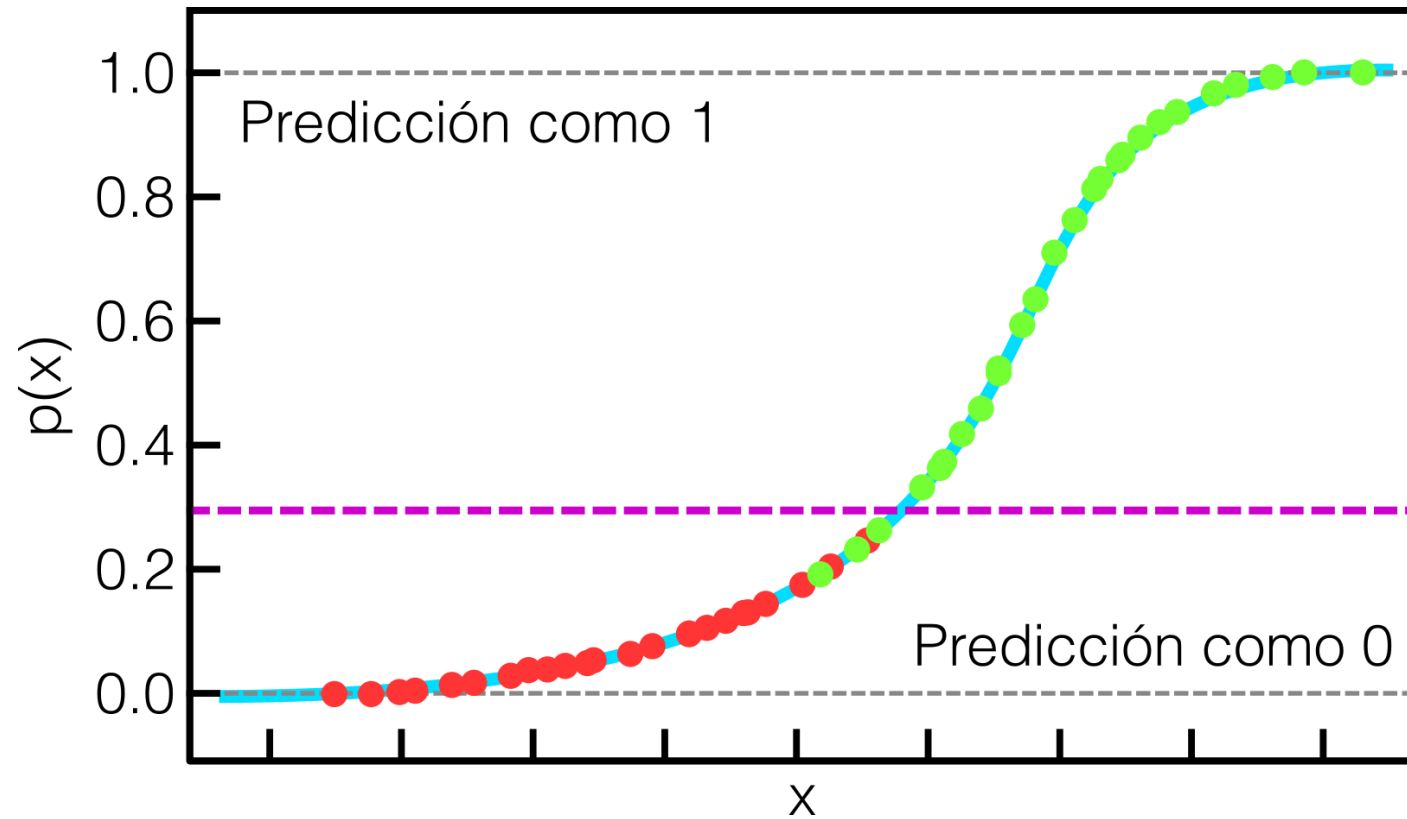


Nos basamos en la idea de que el modelo nos da un valor de probabilidad. Sin embargo, nada impide que el umbral pueda ser definido en otros valores, especialmente cuando las clases están desbalanceadas.

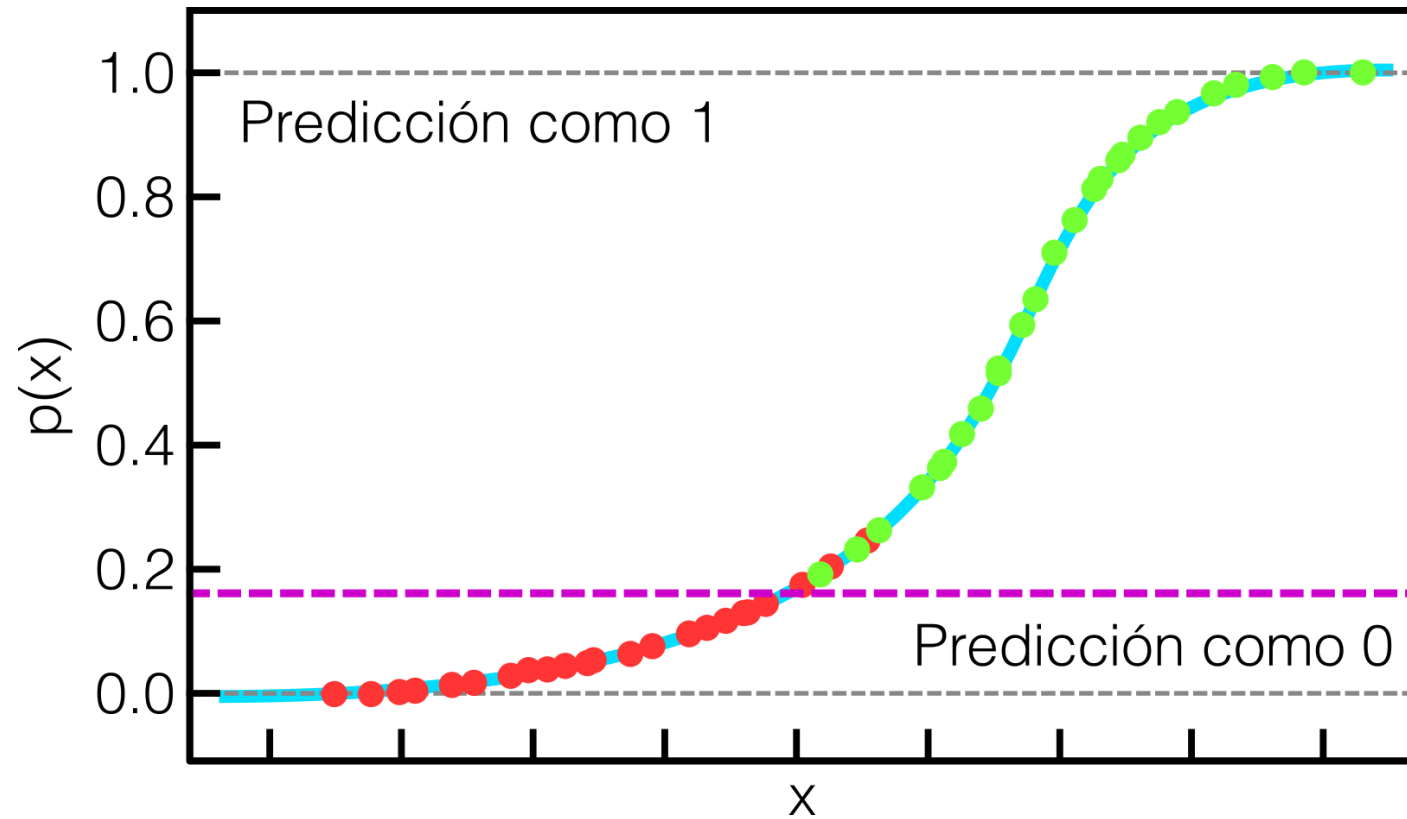
CURVA ROC



CURVA ROC



CURVA ROC

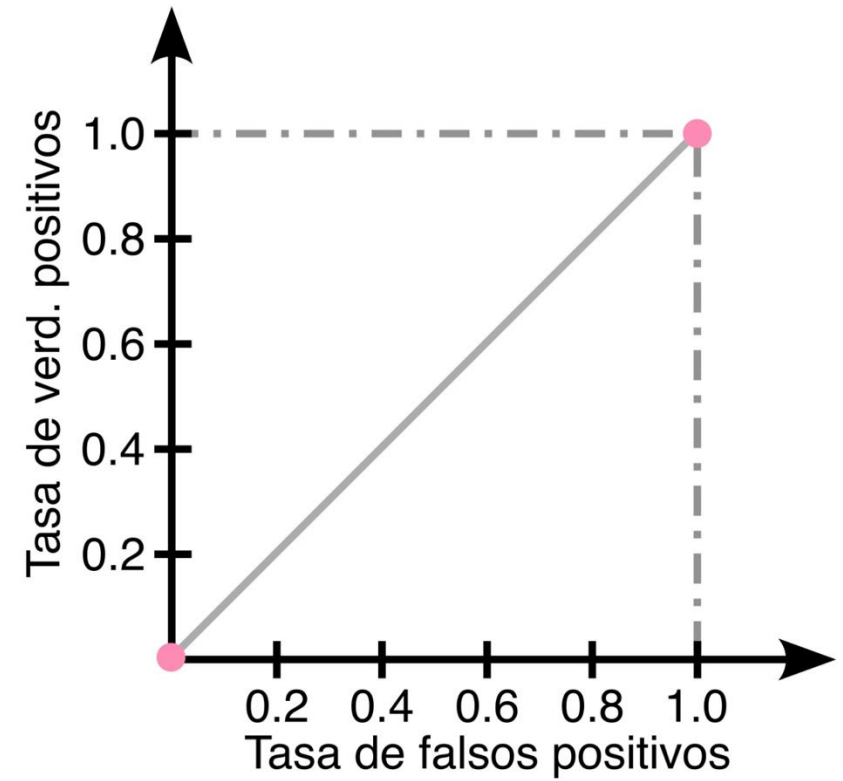
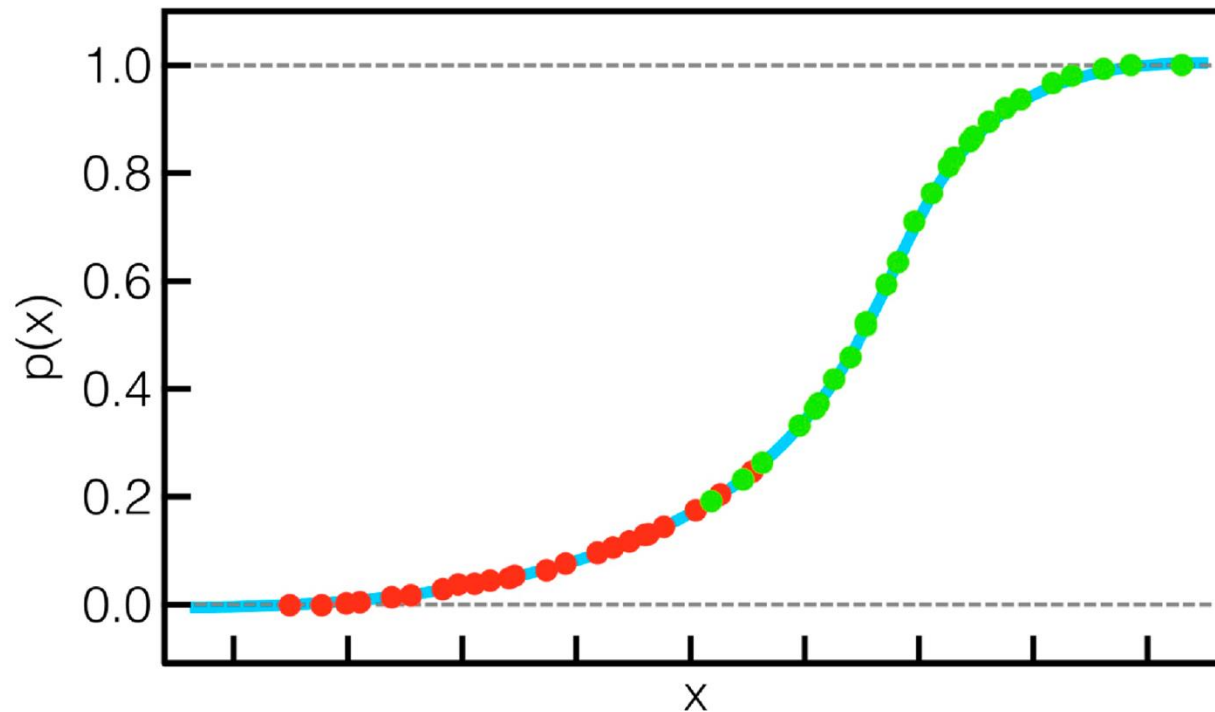


CURVA ROC

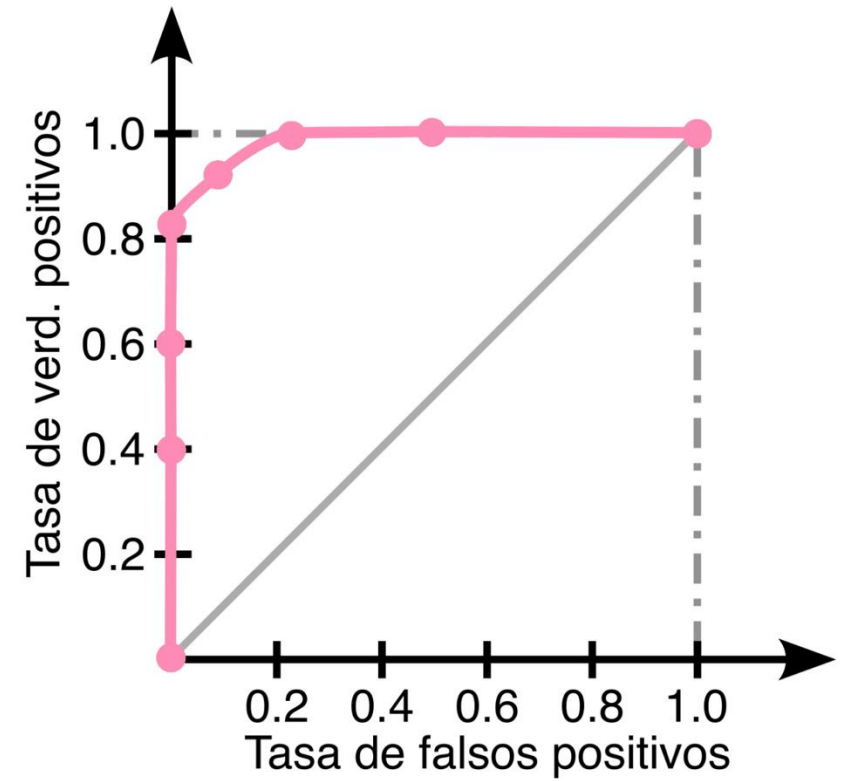
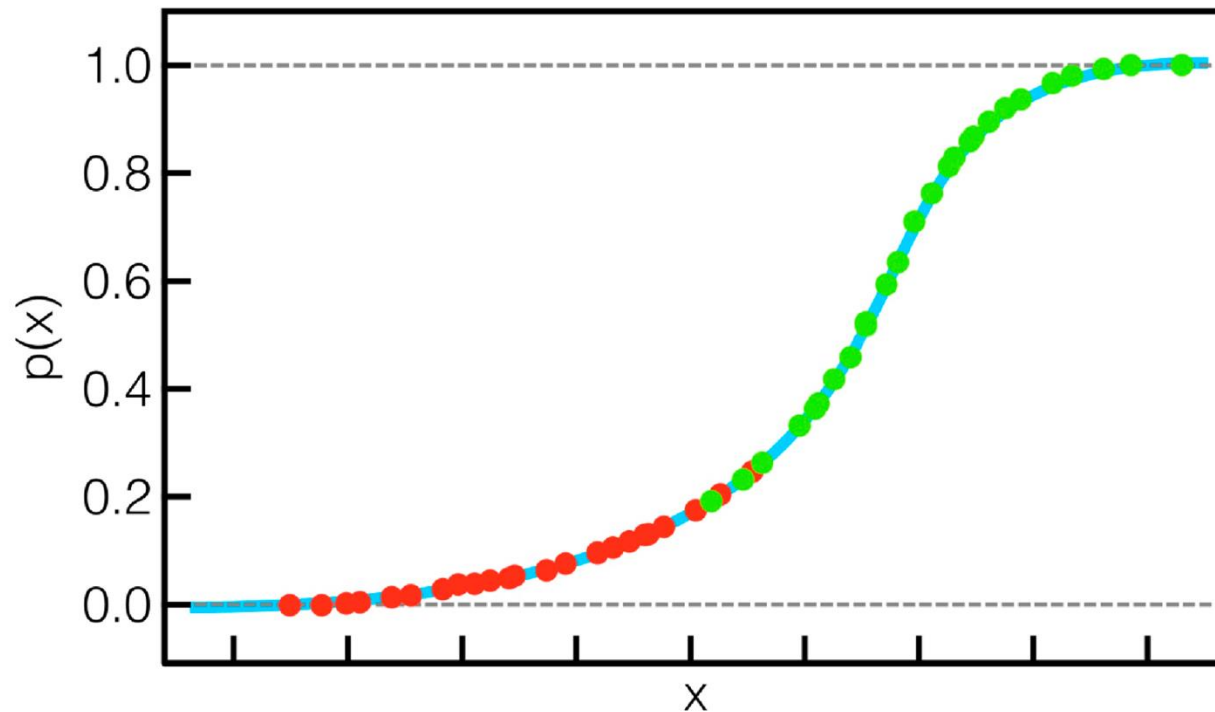
La **curva ROC** (Característica Operativa del Receptor) nos permite ver, para cada valor de umbral, los dos tipos de errores. En el eje de las abscisas se utiliza la **tasa de falsos positivos** (o 1 - especificidad), y en el eje de las ordenadas, **la tasa de verdadero positivos** (sensibilidad).

La curva se obtiene midiendo la sensibilidad y la especificidad para todos los valores de umbral de 0 a 1.

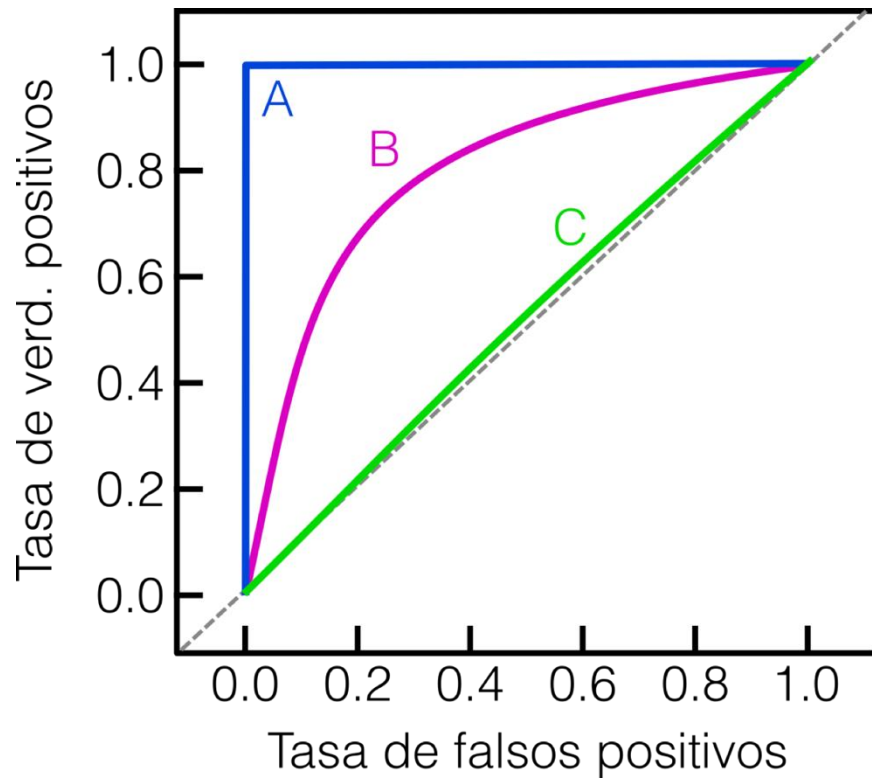
CURVA ROC



CURVA ROC



CURVA ROC

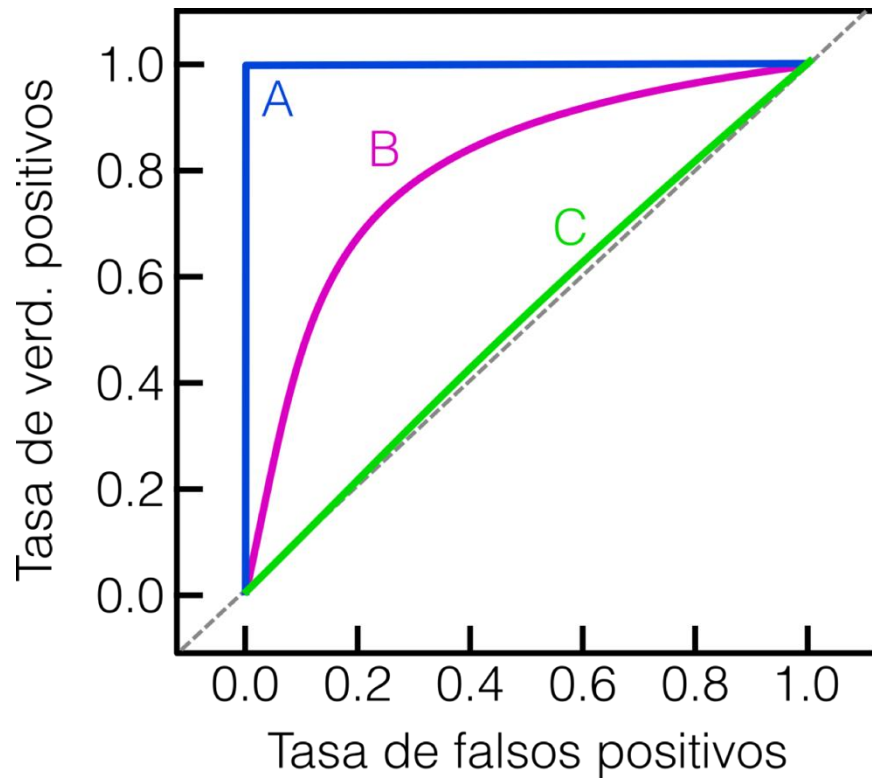


- **A:** La curva de un clasificador perfecto.
- **B:** La curva de un clasificador estándar.
- **C:** La curva de un clasificador que adivina (el peor caso)

La curva ROC permite encontrar el valor de umbral que dé el mejor resultado

Además, permite comparar clasificadores sin preocuparnos por el valor del umbral elegido.

CURVA ROC - AUC



Si queremos resumir esta curva en una métrica, podemos calcular el **área bajo la curva (AUC)**:

- **Modelo A** tendrá un $AUC = 1$
- **Modelo B** tendrá un $0.5 < AUC < 1$
- **Modelo C** tendrá un $AUC = 0.5$