


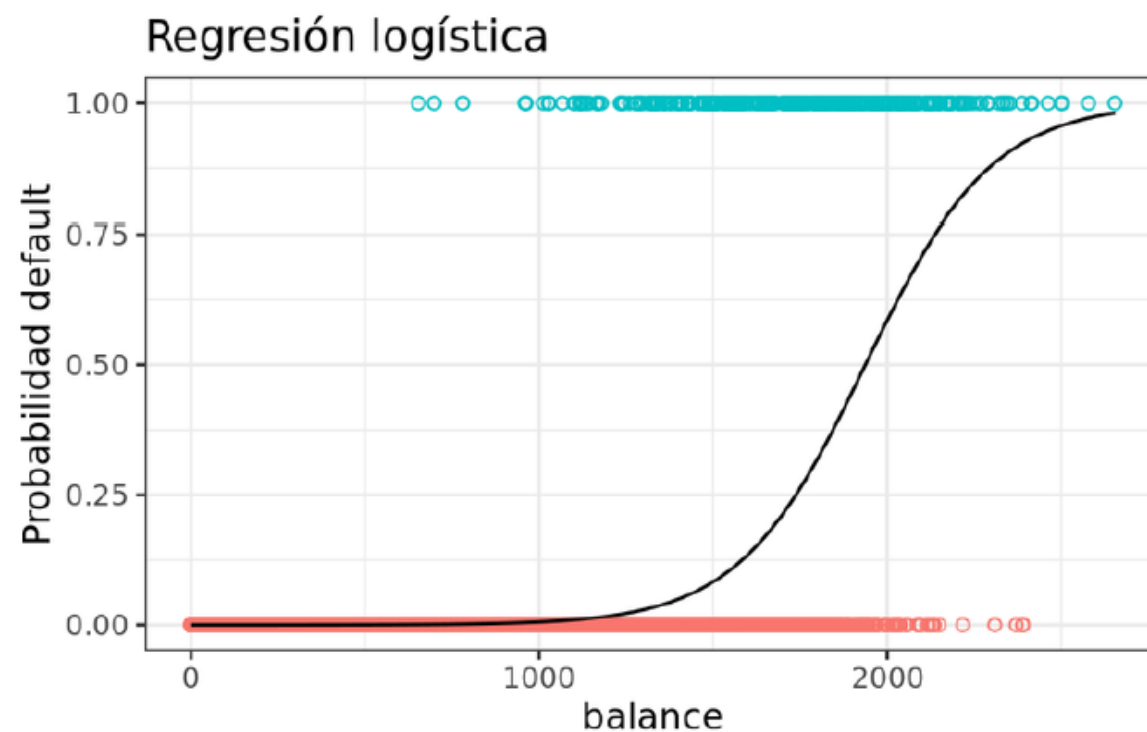
Modelos de Clasificación I: Enfoque basado en Modelos Lineales

 Objetivo: Comprender y aplicar la Regresión Logística para resolver problemas de clasificación binaria en el contexto del aprendizaje automático.



¿Qué es la regresión logística?

La **regresión logística** es una técnica de **análisis de datos** que utiliza las matemáticas para encontrar las **relaciones entre dos factores de datos**. Luego, utiliza esta relación para predecir el valor de uno de esos factores basándose en el otro. Normalmente, **la predicción tiene un número finito de resultados**, como un sí o un no.



¿Por qué necesitamos clasificación?

El desafío del Titanic

¿Cómo predecir si un pasajero del Titanic **sobrevive** o no basándonos en sus características personales?

Aplicaciones cotidianas:

- ¿Un radar detecta una amenaza real?
- ¿Un paciente desarrollará una enfermedad?
- ¿Un cliente incumplirá su crédito?



La clasificación permite tomar decisiones automatizadas en situaciones críticas donde cada predicción puede salvar vidas o prevenir pérdidas económicas.



Fundamentos de la Regresión Logística

Variable dependiente

Categorica (binaria o múltiple), no continua como en regresión lineal

Relación clave

Entre variables X y la probabilidad de que ocurra un evento específico

Función Logística (Sigmoides)

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}}$$

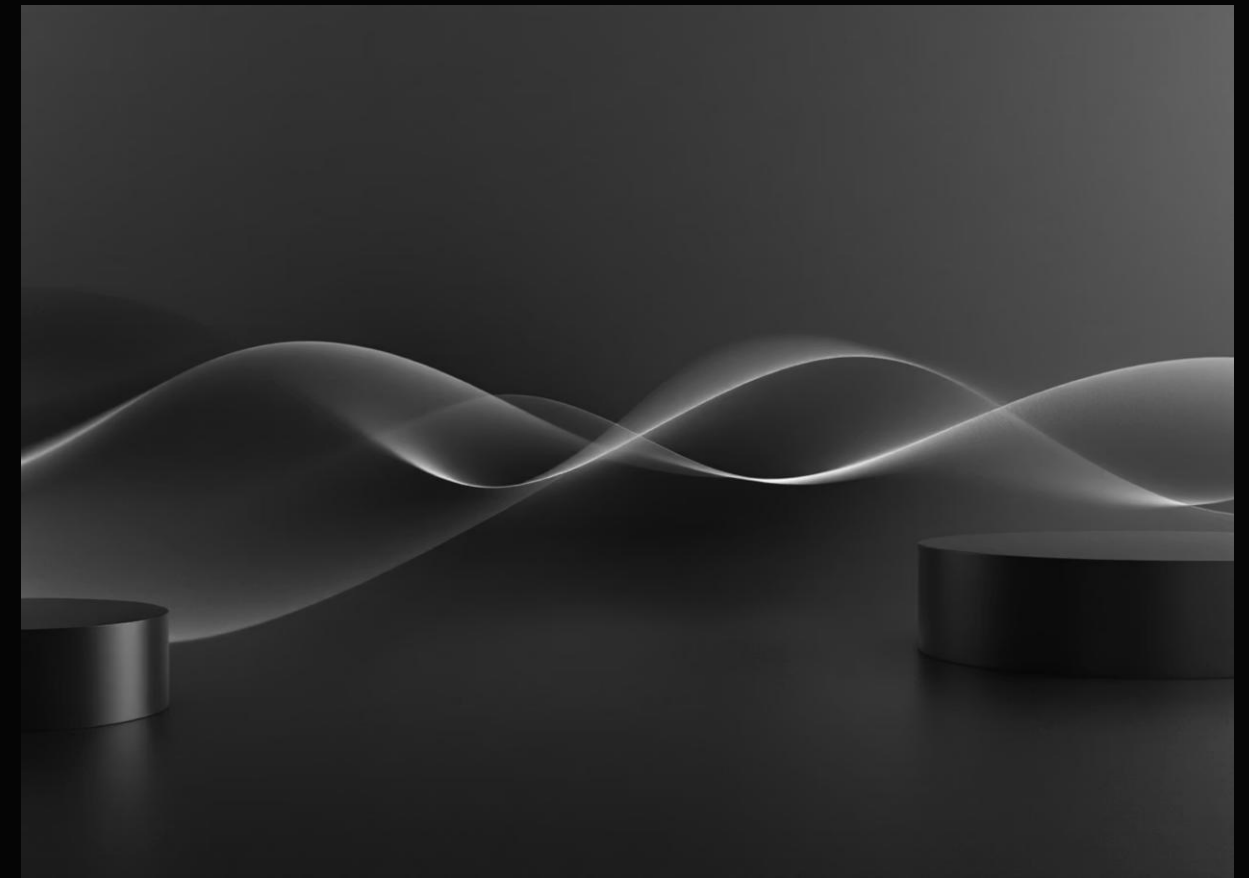
Esta función transforma cualquier valor real en una probabilidad entre 0 y 1, permitiendo interpretaciones probabilísticas claras.

Regresión Logística vs Regresión Lineal



Regresión Lineal

- Variable dependiente **continua**
- Salida: valores reales $(-\infty, +\infty)$
- Interpretación directa del valor



Regresión Logística

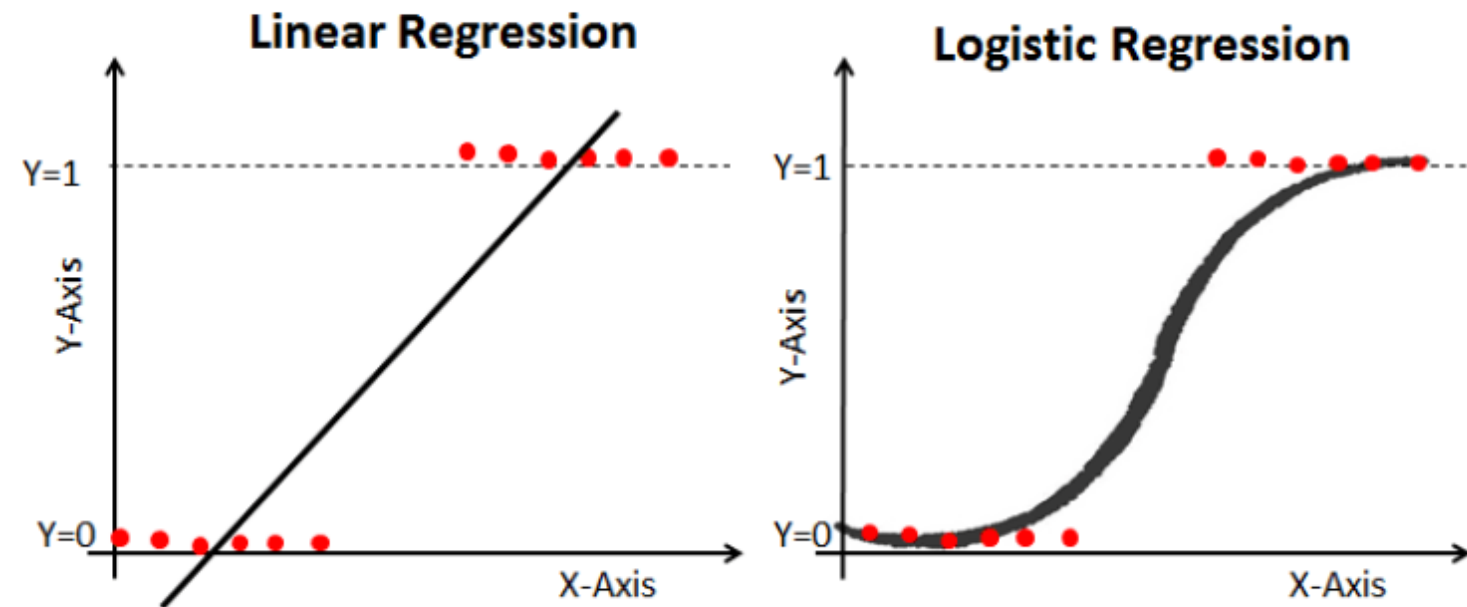
- Probabilidad de **clase**
- Salida: probabilidades $[0,1]$
- Umbral (típicamente 0.5) → Clasificación final



La regresión logística nos permite cuantificar la incertidumbre en nuestras predicciones, proporcionando no solo la clase predicha sino también la confianza en esa predicción.

Regresión lineal vs. Regresión logística

La **regresión lineal** le brinda una **salida continua**, pero la **regresión logística** proporciona una **salida constante**. Un ejemplo de la salida continua es el precio de la vivienda y el precio de las acciones. El ejemplo de la salida discreta es predecir si un paciente tiene cáncer o no, predecir si el cliente abandonará.



Fuente: datacamp.com

Tipos de Clasificación Lineal

01
10

Clasificación Binaria

Dos clases posibles: sobrevivió/no sobrevivió, spam/no spam, enfermo/sano. Es el caso más común y fundamental.



Clasificación Multiclase

Más de dos categorías: tipo de embarcación, género de película, diagnóstico médico específico.



Estrategias de Extensión

One-vs-Rest (OvR): un clasificador por clase vs. todas las demás. **Softmax**: extensión natural para múltiples clases.

Métricas de Evaluación Esenciales

Exactitud

Proporción total de aciertos sobre el total de predicciones

ROC-AUC

Mide la capacidad discriminativa global del modelo en todos los umbrales



Precisión

De las predicciones positivas, ¿cuántas son realmente correctas?

Recall

De todos los casos positivos reales, ¿cuántos logramos capturar?

F1-Score

Media armónica que balancea precisión y recall en una sola métrica

Para encontrar la ecuación de la **regresión logística**, se debe empezar por la ecuación de la **regresión lineal**.



Ecuación de regresión lineal

$$h_{\theta}(x) = \theta_0 + \theta_1 x_1$$

$$h(x) = \sum_{i=0}^n \theta_i x_i = \theta^T x,$$

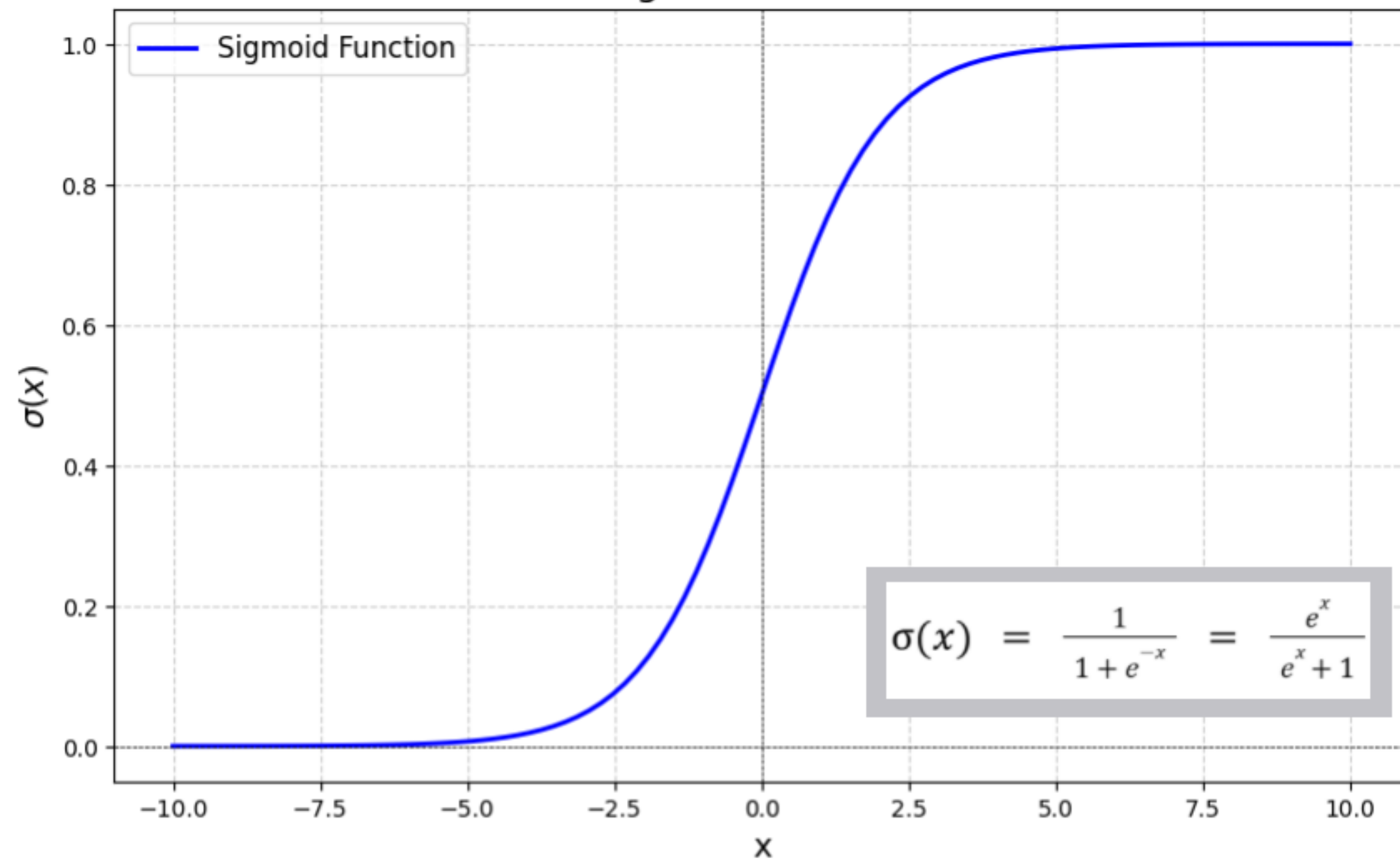
Función sigmoide

$$h_{\theta}(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}},$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

La **función sigmoide** es fundamental en la regresión logística porque se utiliza para transformar la salida de la regresión lineal en una probabilidad.

Sigmoid Function



Aplicando la **función sigmoidea en la regresión lineal**, se obtien lo siguiente:

Ecuación general de la regresión logistica

$$\begin{aligned} P(y = 1 \mid x; \theta) &= h_{\theta}(x) \\ P(y = 0 \mid x; \theta) &= 1 - h_{\theta}(x) \end{aligned}$$

$$p(y \mid x; \theta) = (h_{\theta}(x))^y (1 - h_{\theta}(x))^{1-y}$$

Donde **$P(\mathbf{y})$** es la probabilidad de que ocurra **\mathbf{y}**

Se define la siguiente función de costo

$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X; \theta) \\ &= \prod_{i=1}^m p(y^{(i)} \mid x^{(i)}; \theta) \\ &= \prod_{i=1}^m (h_{\theta}(x^{(i)}))^{y^{(i)}} (1 - h_{\theta}(x^{(i)}))^{1-y^{(i)}} \end{aligned}$$

$$\begin{aligned} \ell(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^m y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) \end{aligned}$$

Regularización en Clasificación



1

Ridge (L2)

Penaliza coeficientes grandes, evitando el sobreajuste. Mantiene todas las variables pero las "suaviza".

2

Lasso (L1)

Realiza selección automática de variables, eliminando características irrelevantes al convertir coeficientes en cero.

3

Elastic Net

Combina lo mejor de L1 y L2, balanceando selección de variables con estabilidad de coeficientes.

La regularización es especialmente importante cuando tenemos muchas variables o problemas de multicolinealidad, ayudando a crear modelos más robustos y generalizables.

Caso Práctico: Supervivencia en el Titanic

1

Definición del Problema

Predecir **Survived** (0 = no sobrevivió, 1 = sobrevivió) basándose en características del pasajero.

2

Variables Predictoras

Edad, sexo, clase social, tamaño de familia, tarifa pagada, puerto de embarque, entre otras.

3

Modelos a Comparar

LogisticRegression base, Logistic + L1 (Lasso), Logistic + L2 (Ridge) para evaluar el impacto de la regularización.

Este dataset clásico permite explorar cómo diferentes características demográficas y socioeconómicas influyeron en las tasas de supervivencia durante el desastre.



Flujo de Trabajo Completo

Preparación de Datos

Análisis exploratorio, limpieza de valores faltantes, ingeniería de características y transformación de variables categóricas.

División de Datos

Separación estratificada en conjuntos de entrenamiento (70-80%) y prueba (20-30%) para evaluación imparcial.

Entrenamiento

Ajuste del modelo de regresión logística con diferentes configuraciones de regularización y optimización de hiperparámetros.

Evaluación

Aplicación de métricas múltiples: exactitud, precisión, recall, F1-score, ROC-AUC y análisis de la matriz de confusión.

Validación

Validación cruzada k-fold para asegurar la estabilidad y robustez del modelo en diferentes subconjuntos de datos.



Conclusiones Clave

✓ Eficiencia y Claridad

Los modelos lineales de clasificación son rápidos, interpretables y sorprendentemente efectivos en muchos problemas reales.

✓ Base Fundamental

La regresión logística constituye el fundamento conceptual de clasificadores más complejos como redes neuronales.

✓ Control de Complejidad

La regularización es esencial en datasets con muchas variables o problemas de multicolinealidad, mejorando la generalización.

✓ Evaluación Integral

El análisis debe ir más allá de la exactitud, considerando precisión, recall y el contexto específico del problema.

¡Dominar estos conceptos te preparará para técnicas más avanzadas!



Métricas de Evaluación para Clasificación

Domina las métricas fundamentales para evaluar modelos de clasificación en machine learning

Exactitud (Accuracy)

Definición

Proporción de predicciones correctas sobre el total de predicciones realizadas por el modelo.

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

Ejemplo práctico (Titanic)

Si el modelo predice correctamente 160 de 200 pasajeros:

$$Accuracy = \frac{160}{200} = 0.80$$



⚠ Limitación importante: Puede ser engañosa con clases desbalanceadas (ej: 90% no sobrevivió)



Precisión (Precision)

¿Qué mide?

De todos los casos predichos como positivos, ¿cuántos son realmente positivos?

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

Ejemplo Titanic

El modelo predijo 50 pasajeros sobrevivientes, pero solo 40 realmente sobrevivieron.

$$Precision = \frac{40}{50} = 0.80$$

Interpretación

Alta precisión = pocos falsos positivos

Esencial cuando el costo de falsos positivos es alto

Recall (Sensibilidad)

¿Qué detecta el modelo?

De todos los casos positivos reales, ¿cuántos logra detectar correctamente el modelo?

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

Ejemplo Titanic

De 80 sobrevivientes reales, el modelo detectó 60 correctamente.

$$Recall = \frac{60}{80} = 0.75$$



Alta sensibilidad = pocos falsos negativos. Crítico en diagnósticos médicos.

F1-Score y ROC-AUC

F1-Score

Media armónica entre precisión y recall que proporciona balance entre ambas métricas.

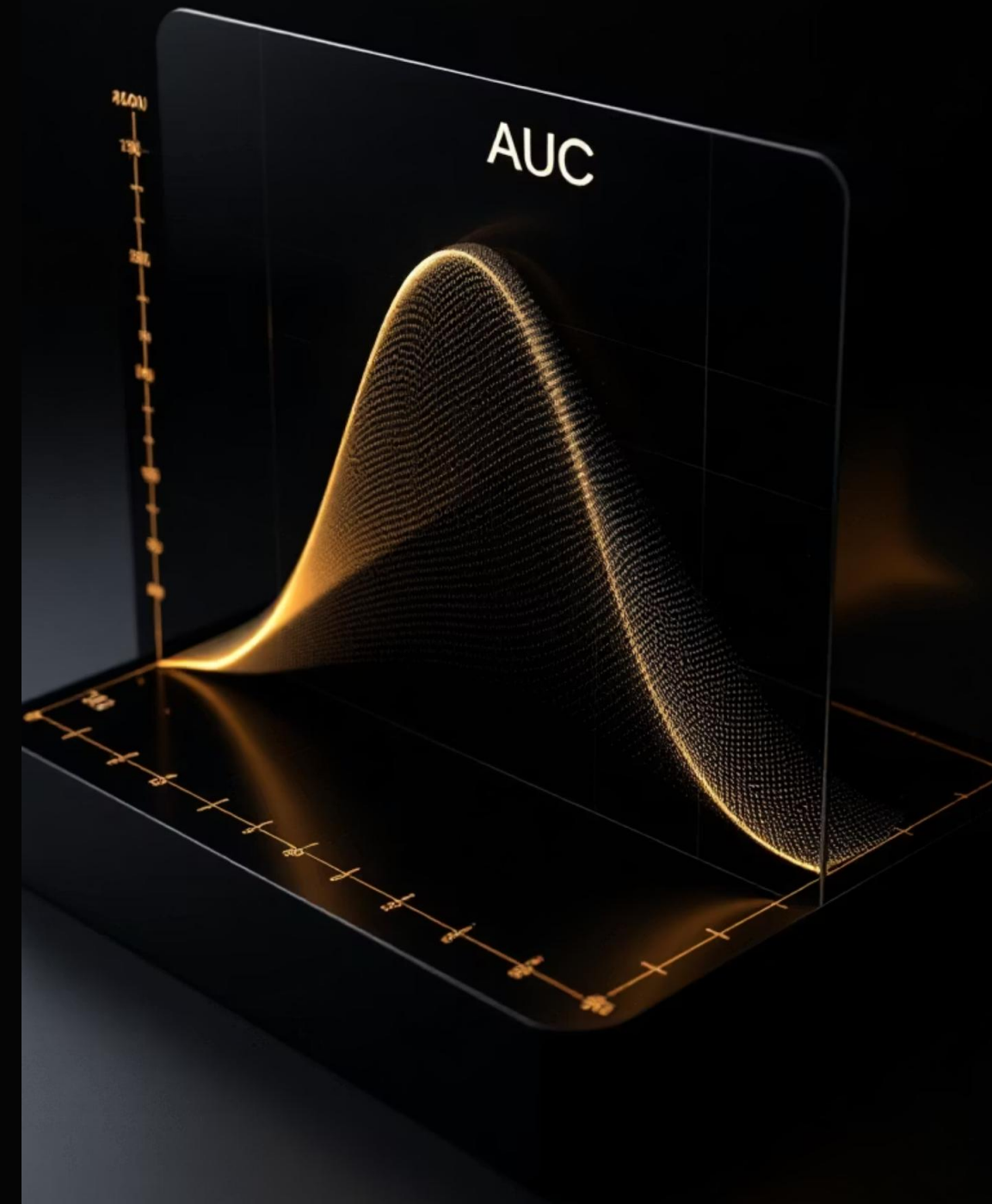
$$F1 = 2 \frac{Precision \cdot Recall}{Precision + Recall}$$

Ejemplo: Precisión = 0.80, Recall = 0.75 → F1 = 0.77

ROC-AUC

Mide la capacidad global de discriminación del modelo en todos los umbrales.

- AUC = 0.5 → modelo aleatorio
- AUC = 1.0 → modelo perfecto
- AUC > 0.8 → buen rendimiento



Matriz de Confusión

Ejemplo con dataset Titanic

	Predijo NO	Predijo Sí	Total
Real NO	TN = 90	FP = 10	100
Real Sí	FN = 20	TP = 80	100



TP (80)

Sobrevivientes correctamente predichos



FP (10)

Falsos positivos: predijo sobreviviente pero no lo fue



TN (90)

No sobrevivientes correctamente predichos



FN (20)

Falsos negativos: sobrevivió pero no lo predijo

✔️ ☒ Ventaja clave: Permite identificar exactamente dónde y cómo se equivoca el modelo