

Esquema completo de Regresión Logística y su contexto

1. Modelos Supervisados

- **Definición:** Algoritmos que aprenden a predecir una **variable objetivo Y** a partir de **variables predictoras X_1, X_2, \dots, X_p**
- **Tipos de problemas:**
 - **Regresión:** la variable objetivo es **continua** → ejemplo: precio de una casa.
 - **Clasificación:** la variable objetivo es **categórica** → ejemplo: spam/no spam, aprobado/suspenso.

2. Regresión Lineal

- **Definición:** Método estadístico para predecir una variable **continua** a partir de una o más variables independientes.
- **Objetivo:** encontrar la combinación lineal de las variables que minimice el error.

2.1 Regresión Lineal Simple

- **Fórmula:**

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- **Interpretación:**
 - $\beta_0 \rightarrow$ valor de Y si $X = 0$
 - $\beta_1 \rightarrow$ cambio promedio en Y por cada unidad de X .
 - $\varepsilon \rightarrow$ error/residuo

2.2 Regresión Lineal Múltiple

- **Fórmula:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

- **Interpretación:**
 - $\beta_i \rightarrow$ efecto de X_i manteniendo las demás variables constantes

- **Supuestos clave:** linealidad, independencia de errores, homocedasticidad, normalidad de errores, ausencia de multicolinealidad.

3. Por qué no usamos regresión lineal para clasificación

- Si $Y = 0/1$:
 - Predicciones de regresión lineal pueden **salir menores que 0 o mayores que 1**
→ no interpretables como probabilidades.
 - Necesitamos un modelo que **devuelva probabilidades entre 0 y 1** → **Regresión Logística**.

4. Regresión Logística

- **Definición:** Modelo supervisado para **clasificación binaria**, que predice la probabilidad de que $Y=1$ dada una combinación de variables independientes.
- **Idea central:**

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)}}$$

- **Componentes:**
 - β_0 → intercepto (probabilidad cuando todas las $X=0$)
 - β_i → efecto de cada variable sobre el **log-odds** de la clase 1
 - Función sigmoide → transforma cualquier valor real en **probabilidad 0–1**

4.1 Interpretación de coeficientes

- **Signo de β_i :**
 - Positivo → aumenta la probabilidad de clase 1
 - Negativo → disminuye la probabilidad de clase 1
- **Magnitud de $|\beta_i|$** → mayor impacto en la probabilidad

4.2 Umbral de decisión

- Para clasificar:

$\hat{Y} = \begin{cases} \text{si } P(Y=1) \geq \text{umbral} \\ \text{si } P(Y=1) < \text{umbral} \end{cases}$

$\text{si } P(Y=1) < \text{umbral}$

- Por defecto, umbral = 0.5, pero puede ajustarse según el problema (enfermedades raras, spam, fraude, etc.)

5. Regresión Logística Simple vs Múltiple

- **Simple:** 1 variable independiente (X_1)

$$P(Y = 1|X_1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1)}}$$

- **Múltiple:** varias variables independientes (X_1, \dots, X_p)

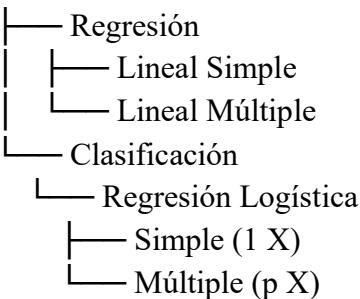
$$P(Y = 1|X_1, X_2, \dots, X_p) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)}}$$

Analogía:

- Simple → “Solo tengo un factor”
- Múltiple → “Tengo varios factores que contribuyen a la probabilidad de que ocurra la clase 1”

6. Resumen visual de la jerarquía

Modelos Supervisados



- **Regresión** → variable dependiente continua
- **Clasificación / Regresión Logística** → variable dependiente categórica (0/1)
- La regresión logística usa la **sigmoide** para convertir la combinación lineal en probabilidad