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** $X1, X2, \dots, Xp$
- **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:**

$Y = \beta 0 + \beta 1X + \varepsilon$

- **Interpretación:**
  - $\beta 0 \rightarrow valor\ de\ Y\ si\ X = 0$
  - $\beta 1 \rightarrow cambio\ promedio\ en\ Y\ por\ cada\ unidad\ de\ X.$
  - ε → error/residuo

2.2 Regresión Lineal Múltiple

- **Fórmula:**

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \varepsilon$

- **Interpretación:**
  - $\beta i \rightarrow efecto\ de\ Xi$ 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:**

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

- **Componentes:**
  - $\beta 0$ → intercepto (probabilidad cuando todas las X=0)
  - βi → 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 βi**:
  - 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:

$\hat{Y} = \{$ si P(Y=1) ≥ umbral

si P(Y=1) < 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 ($X1$)

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

- **Múltiple:** varias variables independientes (*X1,...,Xp*)

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

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
│ ├── Lineal Simple
│ └── Lineal Múltiple
└── Clasificación
└── Regresión Logística
├── Simple (1 X)
└── Múltiple (p X)

- **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