Estadística en Ciencia de Datos

Ejercicio: Regresión Logística Binaria

Base de datos: *Heart Disease* (Cleveland, UCI)

Contexto

La base de datos *Heart Disease* contiene información clínica y demográfica de pacientes sometidos a estudios cardiovasculares. El objetivo es modelar la probabilidad de presencia de enfermedad cardíaca a partir de un conjunto de variables explicativas.

Sea la variable respuesta

$$Y_i = \begin{cases} 1, & \text{si el paciente presenta enfermedad cardíaca,} \\ 0, & \text{en caso contrario.} \end{cases}$$

—

Modelo de Regresión Logística

Considere el siguiente modelo de regresión logística binaria:

$$\begin{aligned} \log\left(\frac{\mathbb{P}(Y_i = 1 \mid \mathbf{x}_i)}{1 - \mathbb{P}(Y_i = 1 \mid \mathbf{x}_i)}\right) = \beta_0 &+ \beta_1 \,\text{edad}_i + \beta_2 \,\text{sexo}_i + \beta_3 \,\text{colesterol}_i + \beta_4 \,\text{presion\_reposo}_i \\ &+ \beta_5 \,\text{frecuencia\_cardiaca\_max}_i \\ &+ \beta_6 \,\text{angina\_ejercicio}_i + \beta_7 \,\text{depresion\_st}_i. \end{aligned} \tag{1}$$

—

Preguntas

a) Ajuste el modelo de regresión logística binaria mediante el método de máxima verosimilitud.

b) Presente la tabla de estimación de los parámetros e identifique las variables estadísticamente significativas al nivel $\alpha = 0{,}05$.

c) Interprete los coeficientes asociados a las variables `edad`, `angina_ejercicio` y `depresion_st` en términos de *odds ratios*.

d) Calcule e interprete los *odds ratios* y sus intervalos de confianza al 95 %.

e) Evalúe la calidad del ajuste del modelo utilizando:

- la razón de verosimilitudes,
- la matriz de confusión y la tasa de clasificación correcta.

f) Analice la capacidad predictiva del modelo mediante la curva ROC e interprete el área bajo la curva (AUC).

g) Discuta las limitaciones del modelo y proponga posibles extensiones.

—

Indicaciones técnicas

- Trate las variables binarias como factores.
- Justifique la inclusión de las variables explicativas.
- Acompañe los resultados con interpretación sustantiva.

—

Matriz de confusión y medidas de desempeño

Con el objetivo de evaluar el desempeño predictivo del modelo de regresión logística, se considera un umbral de clasificación de 0,5 para la probabilidad estimada de presencia de enfermedad cardíaca.

La *matriz de confusión* compara las clases reales con las clases predichas por el modelo y se define como sigue:

	Clase predicha	
Clase real	No	Sí
No	Verdaderos Negativos (VN)	Falsos Positivos (FP)
Sí	Falsos Negativos (FN)	Verdaderos Positivos (VP)

A partir de la matriz de confusión se definen las siguientes medidas de desempeño:

- **Exactitud (Accuracy):**

$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN}.$$

- **Sensibilidad** (o tasa de verdaderos positivos):

$$\text{Sensibilidad} = \frac{VP}{VP + FN}.$$

- **Especificidad** (o tasa de verdaderos negativos):

$$\text{Especificidad} = \frac{VN}{VN + FP}.$$

Interprete los valores obtenidos para cada una de estas medidas y discuta su relevancia en el contexto del diagnóstico de enfermedad cardíaca. —

Objetivo pedagógico

Este ejercicio evalúa la capacidad del estudiante para:

- Formular modelos de regresión logística,
- Interpretar parámetros y odds ratios,
- Evaluar ajuste y desempeño predictivo,
- Conectar inferencia estadística y clasificación.