Estadística en Ciencia de Datos

Ejercicio: Análisis de Conglomerados

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. Más allá del enfoque supervisado (regresión logística), es de interés identificar *grupos homogéneos de pacientes* utilizando técnicas de *Análisis de Conglomerados*.

El objetivo de este ejercicio es explorar la estructura latente de los datos y caracterizar perfiles de riesgo cardiovascular.

—

Variables consideradas

Considere las siguientes variables cuantitativas:

- `edad`
- `presion_reposo`
- `colesterol`
- `frecuencia_cardiaca_max`
- `depresion_st`

Antes de aplicar los métodos de conglomerados, todas las variables deben ser **estandarizadas**.

—

Preguntas

a) Explique por qué es necesario estandarizar las variables antes de realizar un análisis de conglomerados en este contexto.

b) Calcule la matriz de distancias euclidianas entre los individuos a partir de las variables estandarizadas.

—

Conglomerados jerárquicos: método Ward.D2

c) Aplique un análisis de conglomerados jerárquicos utilizando el método de Ward (`ward.D2`).

d) Represente el dendrograma correspondiente e indique el número de conglomerados que considera adecuado, justificando su elección.

e) Asigne cada individuo a un conglomerado y describa las principales características clínicas de cada grupo.

—

Conglomerados no jerárquicos: método *k*-means

f) Aplique el método de k-means considerando el mismo número de conglomerados seleccionado en el método jerárquico.

g) Analice los centroides obtenidos e interprete los perfiles de cada grupo en términos de riesgo cardiovascular.

—

Comparación de métodos

a) Compare los conglomerados obtenidos mediante los métodos `ward.D2` y k-means.

b) Discuta las similitudes y diferencias entre ambos enfoques.

c) Analice la relación entre los conglomerados obtenidos y la variable `enfermedad_cardiaca`.

—

Indicaciones técnicas

- Utilice distancia euclidiana.
- Trabaje con variables estandarizadas.
- Acompañe el análisis con gráficos y tablas descriptivas.
- Interprete los conglomerados desde una perspectiva clínica.

—

Objetivo pedagógico

Este ejercicio permite evaluar la capacidad del estudiante para:

- Aplicar métodos jerárquicos y no jerárquicos de conglomerados,
- Interpretar dendrogramas y centroides,
- Comparar enfoques supervisados y no supervisados,
- Extraer perfiles clínicos relevantes a partir de los datos.