Taller de Análisis de datos - Problema de clasificación 0

Jésica Charaf e Ignacio Spiousas

24 de noviembre de 2023

Problema de clasificación 0

El archivo Distrofia-info contiene una descripción de la Distrofia Muscular de Duchenne (DMD), para cuyo diagnóstico se realizó un estudio cuyos resultados están en el archivo Distrofia-Data. La primera fila es

```
38   1     1    1 1007 22 6 0 079 52.0 83.5 10.9 176
```

Las primeras 5 columnas no sirven. "22" es la edad, "6" el mes, "0" no sirve, "079" el año, y las últimas cuatro son CK, H, PK y LD. El objetivo es proponer una regla para detectar la DMD usando las cuatro variables observadas (enzimas), y estimar su error de clasificación. Se plantean algunas preguntas:

a) CK y H son más baratas de medir que PK y LD. ¿Cuánto aumenta el error si se prescinde de estas últimas?

b) ¿Tiene sentido incluir la edad entre los predictores?

c) La sensibilidad y la especificidad son respectivamente las probabilidades de identificar correctamente a sujetos enfermos y sanos. ¿Cómo elegir el balance entre ambas?

d) Se sabe que la probabilidad de que una mujer sea portadora es 1/3200. ¿Tiene alguna utilidad ese dato?

Resolución

Análisis exploratorio