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

Una vez limpiados los datos tenemos 4 columnas que indican los valores detectados de ciertos marcadores (CK, K, PK y LD) y una columna con los valores "Sí" o "No" que indica si la persona es portadora o no.

En la figura 1 podemos ver la medición de cada marcador dependiendo si la persona es o no portadora. A simple vista podemos ver que, en promedio (barra gris), la medición de los cuatro marcadores es superior cuando la persona es portadora. Sin embargo, pareciera que **H** es el que pareciera separar menos eficientemente ambos grupo.

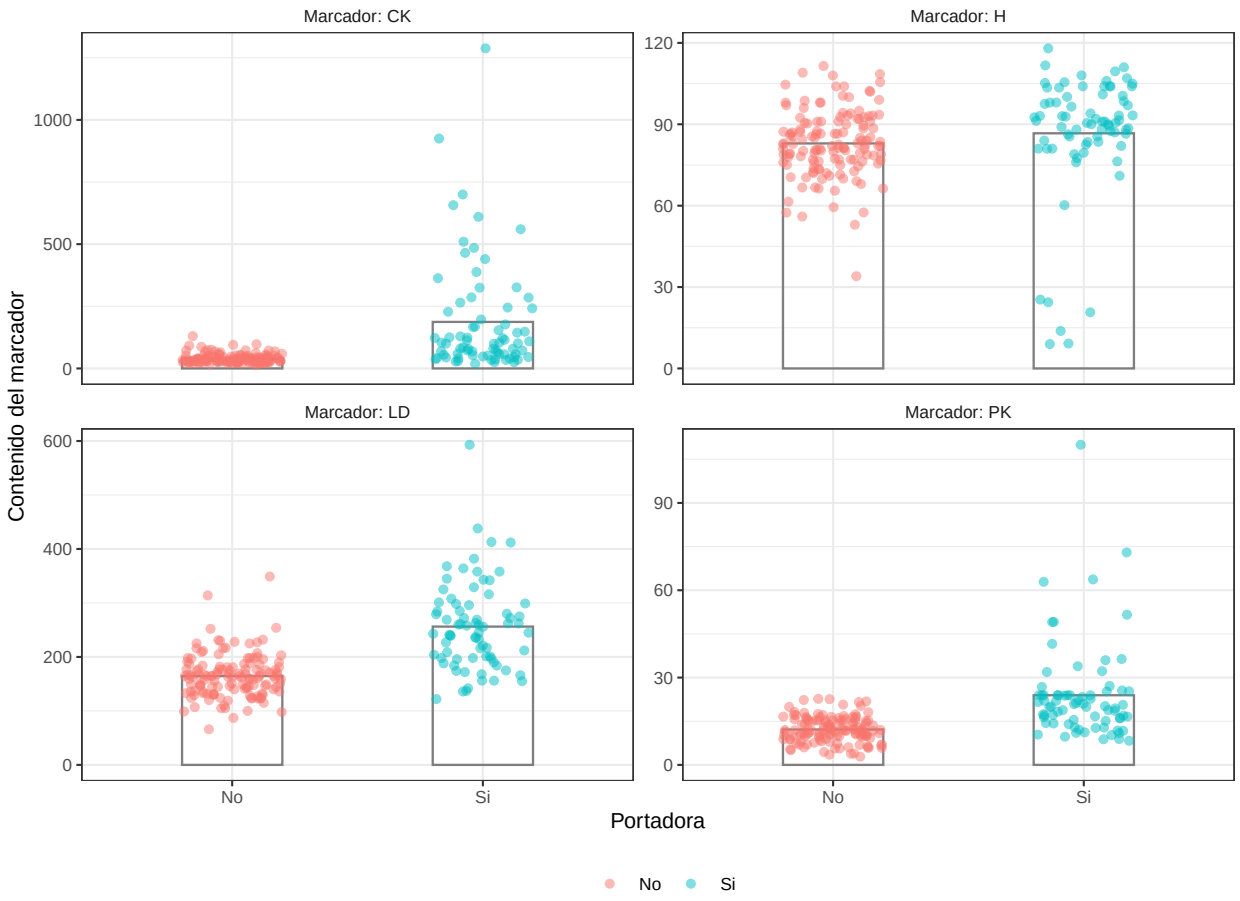

Figure 1: Dependencia de la medición de cada marcador con la condición de portadora de la persona. Los puntos indican los datos individuales mientras que la barra gris indica el promedio para cada categoría.

El objetivo del presente trabajo es entrenar un modelo que, en principio, a partir de las cuatro mediciones de marcadores prediga correctamente si la persona es portadora. Para esto vamos a evaluar un número de modelos de clasificación: Regresión logística, K vecinos vercanos, Naive Bayes y Random Forest. Luego de evaluar cuál modelo es más conveniente para el problema y ajustar sus hiperparámetros y parámetros evaluaremos su capacidad de predicción en el set de *testeo*.