PROBLEMAS PROPUESTOS DE REGRESIÓN LOGÍSTICA
ANÁLISIS ESTADÍSTICO MULTIVARIANTE
GRADO EN CIENCIA E INGENIERÍA DE DATOS

PROBLEMA 1

El fichero **processed.cleveland.data**, contiene los datos correspondientes a un estudio sobre enfermedad cardíaca realizada por *Cleveland Clinic Foundation*.
El fichero contiene un total de 14 columnas, correspondientes a las siguientes variables: *age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, tal* y *num*. La variable "num" toma valores 0, 1, 2, 3 y 4, indicando el tipo de anomalía cardíaca. El valor 0 indica ausencia de enfermedad, mientras que el resto de valores indican algún tipo de anomalía. Para la descripción detallada de cada variable, puede consultarse el fichero **heart-disease.names.**

Se desea realizar un análisis de Regresión Logística con el fin de predecir la presencia (o no) de enfermedad cardíaca en función del resto de variables (predictores). Se pide:

1) Importar los datos del fichero processed.cleveland.data y poner el nombre de cada variable como se indica en el enunciado. Sustituir la variable "num" por una nueva variable llamada "disease" que valga 0 si no hay enfermedad y que valga 1 cuando haya anomalía cardíaca.
2) Eliminar todas las filas que tengan algún valor perdido. IMPORTANTE: confirmar primero si todas las variables son de tipo numérico para identificar adecuadamente los valores perdidos.
3) Pasar a tipo factor las variables que por naturaleza sean de tipo categórico.
4) Dividir el conjunto de datos en entrenamiento y prueba (70% entrenamiento, 30% prueba). Tomar semilla 123.
5) Con los datos de entrenamiento, obtener el modelo ajustado de Regresión Logística usando todos los predictores. ¿Son todos los predictores significativos?

6) Obtener las predicciones para los datos del conjunto de prueba, es decir, la probabilidad predicha de padecer enfermedad cardíaca para cada individuo del conjunto de testeo.
7) Veamos ahora el problema de Regresión Logística como un problema de clasificación. Usando las predicciones del apartado anterior y tomando como punto de corte la probabilidad de 0.5, obtener la clase predicha para los individuos del conjunto de prueba. Medir la eficiencia del modelo calculando la matriz de confusión, accuracy, sensibilidad y especificidad.
8) Para los datos del conjunto de prueba, obtener la curva ROC del método de clasificación, calcular el AUC (área bajo la curva) e interpretar el resultado.
9) Repetir el análisis (apartado 5 y siguientes) pero aplicando primero los métodos de selección de regresores, con el fin de proponer un modelo más parsimonioso.

PROBLEMA 2

¿Qué sucede en el problema anterior si no se realiza el apartado 4? Es decir, qué sucede si no separamos el conjunto de datos en dos subconjuntos: entrenamiento y prueba.

Puedes intentar repetir todo el ejercicio en este nuevo escenario y ver qué sucede con las medidas de bondad del ajuste o medidas de eficiencia del método clasificador.