PRÁCTICA – REGRESIÓN LOGÍSTICA.

Librerías recomendadas: *dplyr, ggplot2, corrplot, ggcorrplot, e1071, GGally, tidyverse, ggpubr,*

base, car, MASS, leaps, hier.part, gvlma, caTools, pROC, ROCR.

Considerar el dataset winequality-white (en formato data.frame, disponible en la URL *"https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv"*), que contiene información relativa a la calidad de determinados vinos blancos en función de características propias de su naturaleza, tales como la acidez, el PH, la densidad, la concentración de alcohol o el azúcar, entre otras.

1.- Realizar una breve auditoría de datos, determinando las cantidades de registros duplicados y vacíos, y eliminando esas duplicidades para generar la tabla protagonista del problema.

2.- La variable a predecir será la calidad (quality) del vino, con lo que se pretende disponer de este atributo en formato factor (inicialmente admite valores entre 3 y 9, indicando así el grado de calidad).

3.- Una vez transformada esa variable a formato factor, codificarla de forma que indique alta calidad para todos aquellos valores por encima de 6, y baja calidad para los casos con etiqueta menor o igual a 6. De este modo, se habrá creado una nueva variable binaria (1/0) que represente al alta o baja calidad de cada registro (el formato deberá ser también factor).

4.- Determinar los porcentajes de vinos con alta y baja calidad en el conjunto de datos sin duplicados.

5.- Partir esa misma tabla sin duplicados en dos conjuntos de datos: uno de training (70% de los registros) para construir un modelo predictivo que permita clasificar vinos en función de su alta o baja calidad, y otro de test (30% de los registros) para realizar las correspondientes pruebas de la eficiencia en la predicción.

6.- Desarrollar con el conjunto de training una regresión logística múltiple que relacione la variable binaria de la calidad del vino con el resto de variables descriptivas de la tabla.

7.- Seguir un proceso step para seleccionar las variables más representativas en la predicción. Describir el modelo final resultante.

8.- Generar las probabilidades asociadas a las predicciones que corresponden al conjunto de test. Crear una nueva variable que asigne a cada registro con probabilidad mayor que 1/3 la etiqueta de buen vino, siendo mal vino todo aquél que muestre una probabilidad por debajo o igual a 1/3.

9.- Mediante una tabla cruzada, mostrar todas las combinaciones de predicciones y valores reales de la calidad del vino (hablar de los resultados obtenidos para conceptos tales como precisión o acierto del modelo, así como de las proporciones de buenos y malos vinos reales que son detectados por el modelo correctamente).