Ejercicios para practicar: Regresión Lineal

El conjunto de datos DatosVino contiene información sobre ciertas características de vinos, junto con las ventas de los mismos.

Las variables contenidas en el fichero son (observa que hay dos variables objetivo diferentes):

Variable	Descripción
Id	Código identificativo del tipo de vino
Beneficio (objetivo)	Beneficio obtenido por la venta de ese tipo de vino
Compra (objetivo)	Variable dicotómica que toma valor 1 si se ha realizado algún pedido de ese tipo de vino, y 0, en caso contrario
Acidez	Características químicas de los distintos tipos de vino: • Densidad y azúcar: sólo valores positivos. • pH: entre 4 y 10. • Restantes: rango ilimitado de valores
Acidocitrico	
Azucar	
Clorurosodico	
Densidad	
Ph	
Sulfatos	
Alcohol	Contenido de alcohol en % (debe situarse entre 0 y 100)
Etiqueta	Percepción del diseño de la etiqueta (MM=muy malo, M=malo, R=regular, B=bueno, MB=muy bueno)
CalifProductor	Calificación (entre 0 y 9) del vino según el productor.
Clasificacion	Clasificación obtenida por un equipo de expertos (4 * = excelente, 1 * = pobre)
Region	Región de la que proviene (toma 3 valores distintos)
PrecioBotella	Precio por botella

Partiendo del conjunto de datos que depuraste en la última clase, el objetivo final de estos ejercicios es construir un modelo de regresión lineal para predecir la variable *Beneficio*. Los ejercicios constan de los siguientes apartados:

1) Determina cuáles serán las variables más útiles para predecir cada una de las variables objetivo a partir de los gráficos y el valor de la V de Cramer.
2) Transformas las variables continuas de input de manera que se maximice la relación con las variables objetivo por separado. ¿Se aplican las mismas transformaciones para las dos variables objetivo?
3) Realiza una partición Entrenamiento-Prueba (80-20) de los datos.
4) Construye un primer modelo de regresión lineal en el que incluyas todas las variables disponibles (sin las transformaciones automáticas ni las interacciones). Evalúa la calidad del modelo resultante e interpreta el parámetro de una variable continua y otra binaria.
5) Basándote en los resultados del apartado 1, construye un modelo de regresión que contenga únicamente las variables detectadas. ¿Este modelo es mejor que el anterior?
6) Basándote en la importancia de las variables del modelo inicial, determina las variables menos útiles para predecir el precio de la vivienda. A continuación, construye un modelo de regresión como el del apartado 4 pero eliminando las variables detectadas. ¿Este modelo es mejor que los anteriores?
7) Partiendo del mejor modelo de los 3 anteriores (ten en cuenta su comportamiento en entrenamiento y prueba, así como el número de parámetros que tienen), incluye las interacciones que consideres puedan ser influyentes y determina si lo son o no.
8) Utilizando validación cruzada (20 repeticiones, 5 grupos), determina cuál de los 4 modelos anteriores es preferible.
9) Evalúa el modelo ganador (estabilidad y bondad del mismo, variables más importantes, etc.).