Problemas para la evaluación de modelos

1) Utlizaremos el conjunto de datos recopilado por Worthy, S. L., Jonkman, J. N. y Blinn-Pike, L. (2010). Sensation-seeking, risk-taking, and problematic financial behaviors of college students. Journal of Family and Economic Issues, 31(2), 161–170 (overdrawn.csv).

Para este conjunto de datos, los investigadores realizaron una encuesta a 450 estudiantes universitarios de cursos introductorios de gran 'envergadura' en la Universidad Estatal de Mississippi o en la Universidad de Mississippi. La encuesta contenía cerca de 150 preguntas, pero solo cuatro de estas variables se incluyen en este conjunto de datos. (Puedes consultar el artículo para saber cómo las variables adicionales a estas cuatro afectan el análisis). El principal interés de los investigadores fueron los factores relacionados con si un estudiante ha sobregirado alguna vez su cuenta corriente.

El conjunto de datos contiene las siguientes variables:

Age	Age of the student (in years)
Sex	0 = male or 1 = female
DaysDrink	Number of days drinking alcohol (in past 30days)
Overdrawn	Has student overdrawn a checking account? 0 = no or 1 = yes

Crea un modelo basado en un árbol de decisión para predecir el sobregiro de la cuenta corriente del estudiante según la edad, el sexo y los días de consumo de alcohol. Dado que DaysDrink es una variable numérica, es posible que debas convertirla en una categórica. Una sugerencia sería:

si (número de días de consumo de alcohol < 7) = 0
(7 >= número de días de consumo de alcohol > 14) = 1
(número de días de consumo de alcohol >= 14) = 2

2) Hay un conjunto de datos de recopilación de spam de YouTube disponible en la carpeta (YouTube-Spam-Collection-v1). Se trata de un conjunto público de comentarios recopilados para la investigación de spam. Consta de cinco conjuntos de datos compuestos por 1956 mensajes reales extraídos de cinco vídeos. Estos cinco vídeos corresponden a canciones pop populares que se encontraban entre las 10 más vistas del período de recopilación.

Los cinco conjuntos de datos tienen los siguientes atributos:

COMMENT_ID: ID único del comentario

AUTHOR: ID del autor
DATE: Fecha de publicación del comentario
CONTENT: El comentario
TAG: Para spam 1, en caso contrario 0

Para este problema, utiliza cuatro de estos cinco conjuntos de datos para crear un filtro de spam y utilízalo para comprobar la precisión del resto del conjunto de datos.

3) Considera el conjunto de datos contenido en el archivo tipjoke.csv para este problema, proveniente del estudio de Nicholas Gueaguen (2002). Los efectos de un chiste en las propinas cuando se proporciona al mismo tiempo la cuenta del consumo. Journal of Applied Social Psychology, 32(9), 1955-1963.

¿Puede un chiste afectar si un camarero de una cafetería recibe propina?

Este estudio investigó esta cuestión en una cafetería de un famoso resort de la costa oeste de Francia. El camarero asignó aleatoriamente a los clientes que pedían café a uno de tres grupos: al recibir la cuenta, un grupo también recibió una tarjeta con un chiste, otro grupo recibió una tarjeta con un anuncio de un restaurante local y un tercer grupo no recibió ninguna tarjeta. Se registró si cada cliente dejó propina.

El conjunto de datos contiene las siguientes variables:

Card	Type of card used: Ad, Joke, or None
Tip	1 = customer left a tip, or 0 = no tip
Ad	Indicator for Ad card
Joke	Indicator for Joke card
None	Indicator for No card

Utiliza un árbol de decisión para determinar si el camarero recibirá propina a partir de las variables predictoras.