GUÍA DE LECTURA III: Preprocesamiento

1. ¿Qué se entiende por preprocesamiento según Han & Kamber? ¿Por qué es necesario preprocesar los datos?
2. ¿Cuáles son las etapas de las tareas de preprocesamiento?
3. 😁 **Para ir conectando**: ¿Qué paralelo encuentra con lo visto en el curso hasta el momento con la etapa de sumarización y descripción de datos planteada por Han & Kamber? ¿Qué nombre recibe esa etapa o tarea en el proceso de KDD visto?
4. ¿Cuáles son las principales tareas que se realizan en la etapa de limpieza de datos?
5. ¿Qué diferencia existe entre ruido y outliers?
6. ¿Cómo puedo intervenir en una variable que presenta ruido?
7. ¿Cuál es la técnica univariada para identificar valores atípicos?
8. ¿Qué desafíos presenta la integración de datos? ¿Cómo se puede abordar el problema de redundancia?
9. ¿Cuál es el objetivo de la etapa de transformación de datos? ¿Qué procesos involucra esta etapa?
10. 🤓 📏**Situación 1**: Tenemos un conjunto de datos con 5 variables medidas en las siguientes unidades: metros, milímetros, kilogramos, pesos ($) y grados celsius respectivamente. ¿Qué transformación sería necesaria realizar si mi tarea de KDD utiliza un algoritmo basado en distancias?
11. ¿Cuál es el desafío de la reducción de datos? ¿Qué diferencias existe en la selección de atributos con respecto a la reducción de dimensionalidad?
12. 🤓 🍐🍏**Situación 2**: Si tenemos una variable definida en el dominio de los reales R y mi tarea de data mining solo trabaja con etiquetas. a) ¿Qué operación se debe realizar? b) ¿Qué tipo de variable consigo tener al finalizar el proceso?
13. 🥇**Bonus Track**: Si tengo dos variables de igual tipo que la resultante del punto 12 b) que vienen de diferentes fuentes de datos. ¿Cómo puedo saber si están relacionadas entre sí? ¿Con qué etapa de preprocesamiento se resuelve ese requerimiento? ¿Qué técnica es la indicada para esas variables?

Bibliografía sugerida:

* Data Mining. Concepts and Techniques. Jiawei Han and Micheline Kamber. Capítulo 2.
* Brown, M. L., & Kros, J. F. (2003). Data mining and the impact of missing data.Industrial Management & Data Systems, 103(8), 611-621.