PRÁCTICA 3: ÁRBOLES DE DECISIÓN

PARTE I: (Python)

Tome los datos de ejercicio de teoría en el que se discretizaba un valor de la presión. Cree un pequeño script en python que calcule la Ganancia de Información y justifique, de esta manera, la cota escogida para dicha digitalización. Asimismo, usando el método de discretización basado en la entropía, verifique que el resultado es el mismo.

PARTE II: (WEKA)

En el siguiente enlace del repositorio de la UCI:

http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data

Se encuentra un ejemplo de clasificación de pacientes relacionados con enfermedades torácicas, en particular, con el riesgo a un año de padecerlas.

Responder a las siguientes cuestiones:

1. Por qué no se puede aplicar directamente ID3.
2. Pase el algoritmo J48 aplicando el método de retención o Hold-Out. Analice el árbol obtenido sin poda. ¿Se podría prescindir de algún atributo? Si es así, hágalo y compare los resultados de nuevo. Para esta tarea, puede crear un pequeño programa en Python, que se alimente de un fichero proveniente de una exportación del árbol en formato texto que proporciona Weka.
3. Habrá notado que cuando se usa algún atributo numérico, implícitamente se aplica una discretización al plantear las diferentes ramas del árbol a partir de él. ¿Por qué es más eficiente esta técnica que la aplicada en PARTE I?
4. Plantee, entonces, una discretización basada en el punto anterior, aunque no resulte ser binaria, sino en tantos tramos como induzcan los valores usados al formar las ramas del árbol con J48 sin poda (Hold-Out). Introduzca este fichero de nuevo al algoritmo ID3. Compare los resultados con el J48 de 2. Igualmente, para aislar estas cotas, puede construir otro

pequeño programa en Python que las localice también en el árbol en formato texto que proporciona Weka.