CURSO: CMP 5002 - DATA MINING
COLEGIO: POLITÉCNICO
Semestre: 1er Semestre 2023/2024

Tarea 6: Ejercicio usando el procesamiento de los datos y la clasificación basada en árboles de decisión.

Problema:

1. Dado el subconjunto de variables obtenidas como resultado de la tarea de selección de características (proyecto 4). Se desea:
    - Teniendo en cuenta el espacio reducido obtenido en la tarea 4, se aplicar la tarea de normalización *min-max* a los datos.
    - Utilizar la técnica, *stratified k-fold cross-validation* (k=10) *with random seed* antes del paso de clasificación para dinámicamente crear los segmentos de *train* and *test* por cada *fold*.
    - Aplicar la tarea de clasificación en conjunto con el *stratified* 10-*fold cross-validation* usando tres árboles de decisión diferentes: ID3 (*information gain*), C4.5 (*gain ratio*) y CART (*gini*).
    - Es obligatorio mostrar la trazabilidad de la tarea durante la ejecución del programa:
        i. El *Dataset* (DATA) original y normalizado. **(0.5 puntos)**
        ii. Los resultados de clasificación obtenidos por los tres árboles de decisión:
            1. Mostrar la matriz de confusión obtenida por cada árbol. **(1.5 puntos).** Investigar como generar la matriz de confusión.
            2. Mostrar el promedio y desviación estándar para cada métrica de validación: *accuracy (ACC), precision (PRE), recall (REC), AUC* (area under the receiver operating characteristic curve), *F1 score*, *MCC* (*Mathew correlation coefficient*), *sensitivity, specificity* para cada árbol de decisión. Se deben obtener resultados en alguna métrica superior al 90% (para ACC) o 0.9 (para las restantes). **(4 puntos)**
            3. Mostrar un plot de *AUC* para cada árbol. **(2.5 puntos)**
            4. Mostrar un plot de *precission vs recall* para cada árbol. **(2.5 puntos)**
    - Cargar al D2L los códigos implementados (fichero compactado) dentro del plazo de entrega.

Nota: Esta tarea depende de la realización del proyecto 4. La no obtención de un conjunto reducido de variables conlleva a la aplicación de los clasificadores sobre el *dataset* completo, lo cual es totalmente ineficiente. Dicha ineficiencia equivale a una penalización del 40% del valor de la tarea (4 puntos).