



CURSO: CMP 5002 - DATA MINING
COLEGIO: POLITÉCNICO
Semestre: 2do 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:
 - Aplicar la tarea de **normalización min-max** a los datos del conjunto reducido (de acuerdo a las 3 primeras hipótesis empleadas).
 - Utilizar la técnica, *stratified 10-fold cross-validation* (CV) antes del paso de clasificación para dinámicamente crear los segmentos de *training* and *validation* 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 los resultados obtenidos por los clasificadores de acuerdo al promedio y desviación estándar de cada métrica de validación en presencia del *stratified 10-CV*: *accuracy* (ACC), *precision* (PRE), *recall* (REC), *AUC* (area under the receiver operating characteristic curve), F1-score, MCC (Matthews Correlation Coefficient). **(3 puntos)**
 3. Mostrar un plot de *AUC* para cada árbol. **(2 puntos)**
 4. Mostrar un plot de *precision vs recall* para cada árbol. **(2 puntos)**
 5. Graficar los árboles (tres) que mejor clasifican los datos. **(1 punto)**
 - Cargar al D2L los códigos implementados (archivo 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).