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

Tarea 4: Ejercicio usando el procesamiento de los datos

Problema:

1. Dado un conjunto de datos aleatorio con N variables y dos clases de salida (n=70, c=2). Se desea:
    - El *data set* a usar para este ejercicio se encuentra en: P4
    - Aplicar las tareas de procesamiento de datos: Normalización y Reducción.
        a) Para la normalización:
            i. Usar la técnica *min-max* vista en clase.
        b) Para la reducción:
            i. Se debe implementar un método de selección de características del paradigma *Wrapper,* utilizando la combinación de un modelo metaheurístico con una función objetivo del paradigma *Filter* .
            ii. Los metaheurísticos a considerar son bioinspierados en la naturaleza y se denominan: *Differential Evolution (DE), Genetic Algorithms (GA), Advanced Ant Binary Colony Optimization (ABACO), Particle Swarm Optimization (PSO), Simulated Annealing (SA), and Gray Wolf Optimizer (GWO).* Se recomienda estudiar los distintos métodos a través del uso de bibliografía científica razonable (*scientific papers in journals*).
            iii. Las funciones objetivo a utilizar son: *Information Gain (IG), Gain Ratio (GR), ReliefF, Symmetrical Uncertainty (SU), x^2 (Chi2)-test, Mutual Information (MI)*. Cabe señalar que estas funciones miden la importancia individual de una variable (feature) con respecto a la clase. Por tanto, para medir la importancia de un conjunto de variables (*n1<n features*), se aplicará una modificación basada en el promedio (*Sum[$f(x_i)$]/N_1; i=1:N_1)*. Con esto mediremos la importancia del subconjunto basado en el mérito per cápita. Mientras mayor sea el valor de importancia per cápita, mejor poder de discriminación del subconjunto evaluado.
            iv. Cada equipo de estudiantes debe hacer una investigación sobre el *wrapper* (metaheurístico + función objetivo) seleccionada, de forma tal que puedan entenderlo,

defenderlo e implementarlo. Cada equipo debe implementar *wrapper* diferente.

- Del acto de evaluación y defensa:
  c) Es obligatorio mostrar la trazabilidad de la tarea durante su ejecución:
     1. *Data set* original y normalizado. **(1 punto)**
     2. El método de selección de características empleado. Sus características. Su funcionamiento (ej: como seleccionad y determina la importancia de las características). **(5 puntos)**
     3. *Top five* de subconjuntos de características obtenidos y su importancia per cápita. **(2 punto).** Se debe hacer una pequeña investigación para entender e interpretar la importancia per cápita.
     4. *AUC score-based ranking* de los subconjuntos características obtenidos (resultado del inciso (3)). **(2 puntos).** Se debe hacer una pequeña investigación para entender e interpretar el cáculo de la métrica AUC (*area under the receiver operating characteristic curve).*
  d) Cargar al D2L los códigos implementados (fichero compactado) dentro del plazo de entrega.

Nota: En cada fase de evaluación el profesor aplicará puntos de chequeo sobre el código implementado. Además, esta tarea constituye la base para las restantes tareas de clasificación supervisada.