



CURSO: CMP 5002 - DATA MINING
COLEGIO: POLITÉCNICO
Semestre: 2do 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:
(https://estudusfqedu-my.sharepoint.com/:x:/g/personal/nperez_usfq_edu_ec/Eelxd5X3zqFMuhqxpov0W54BjqdBHHKsrwFglg8D974N1Q?e=MXDrzp)
 - 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 bioinspirados en la naturaleza y se denominan: *Artificial Bee Colony ABC*, *Ant Lion Optimizer (ALO)*, *Salp Swarm Algorithm (SSA)*, *Cuckoo Search (CS)*, and *Bat Algorithm (BA)*. 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: *Gain Ratio (GR)*, *ReliefF*, *Symmetrical Uncertainty (SU)*, χ^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 ($n_1 < n$ features), se aplicará una modificación basada en el promedio ($\text{Sum}[f(x_i)]/N_1$; $i=1:N_1$). 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 un *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 selecciona y determina la importancia de las características). **(4 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 cálculo de la métrica AUC (*area under the receiver operating characteristic curve*).
 5. Literatura: la literatura por donde estudiaron tanto el método metaheurístico como la función de costo (journals, conferences, blog, etc), origen de la implementación (GitHub, Blog, etc). Evitar literatura no reconocida. **(1 punto)**
- d) Cargar al D2L los códigos implementados (fichero compactado) y la literatura solicitada 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.