CURSO: CMP 5002 - DATA MINING
COLEGIO: POLITÉCNICO
Semestre: Primer Semestre 2020/2021

Tarea 4: Ejercicio usando el procesamiento de los datos

Problema:

1. Dado un conjunto de datos extraído del UCI *machine learning repository*, con N variables y dos clases de salida. Se desea:
    - El *dataset* a usar para este ejercicio será el SPECTF (datos completos, baja dimensión, poca numerosidad y salida binaria). http://archive.ics.uci.edu/ml/datasets/SPECTF+Heart
    - El SPECTF *dataset* está dividido en dos subconjuntos, entrenamiento ("SPECTF.train" 80 instancias) y prueba ("SPECTF.test" 187 instancias). Para los efectos de la tarea se deben unir ambos subconjuntos.
    - 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 *Filter.* Se pueden usar funciones objetivo tales como: *information gain, gain ratio, relief, one rule, symmetrical uncertainty, x^2 (chi)-test*. Cada estudiante debe hacer una investigación mínima sobre la función seleccionada, de forma tal que pueda entenderla, defenderla e implementarla.
            ii. Los métodos de filtrado (*filter*) no analizan las inter-relaciones atributo-atributo. Por tanto, se desea aplicar un análisis de correlación de Pearson sobre los resultados obtenidos por el método implementado, de forma tal que el estudiante pueda encontrar un subconjunto N1<N ideal.
    - Del acto de evaluación y defensa:
        c) Es obligatorio mostrar la trazabilidad de la tarea durante su ejecución:
            1. *Dataset* original y normalizado.
            2. El método de selección de características empleado. Sus características. Su funcionamiento (ej: como determina la importancia de las características).
            3. *Ranking* de características obtenido (resultado del inciso (i)) de acuerdo al método implementado.

4. Resultados del análisis de correlación (tabla con los valores de correlación).
5. *Ranking* de características reducido de acuerdo al análisis de correlación realizado (resultado del inciso (ii)).

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.