



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

**Tarea 5:** Ejercicio usando el procesamiento de los datos y la clasificación basada en el *NB (Naive Bayes)* y *kNN (k-nearest neighbor)*.

**Problema:**

1. Dado el subconjunto de variables (alguna de las hipótesis) 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 la hipótesis empleada).
  - 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*. [Investigar la técnica para aplicarla correctamente.](#)
  - Aplicar la tarea de **clasificación** en conjunto con el *stratified 10-fold cross-validation* usando los clasificadores *Naive Bayes (NB)* y *k-nearest neighbors (kNN)*.
  - El clasificador *kNN* debe ser implementado con dos medidas de distancia diferentes (*Euclidiana* y *Mahattan*) y optimizado en el intervalo de  $[k>1 \ \& \ k<16]$ , considerando los impares.
  - Es obligatorio mostrar la trazabilidad de la tarea durante la ejecución del programa:
    - i. El *Data set* original y normalizado. **(1 puntos)**
    - ii. Mostrar los resultados obtenidos por ambos 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). [Investigar cada las métricas para emplearlas correctamente.](#) **(5 puntos)**
    - iii. Mostrar la selección óptima del valor de *k* para cada métrica de distancia basado en un gráfico que muestre el **ACC** obtenido (eje Y) por el clasificador a medida que varía el valor de *k* (eje X). **(2 puntos)**
    - iv. Se deben obtener resultados iguales o superiores al 80% para la ACC o 0.8 para la PRE, REC, o AUC. **(2 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 *data set* completo, lo cual es totalmente ineficiente. Dicha ineficiencia equivale a una **penalización** del 40% del valor de la tarea (4 puntos).