CURSO: CMP 5002 - DATA MINING
COLEGIO: POLITÉCNICO
Semestre: 2do Semestre 2023/2024

Tarea 7: Ejercicio usando la clasificación con redes neuronales artificiales (ANN)

Problema:

1. Dado el subconjunto de variables obtenidas como resultado de la tarea de selección de características (proyecto 4). Se desea:
    - Teniendo en cuenta el espacio reducido obtenido en la tarea 4, se desea aplicar la tarea de normalización *min-max* de los datos.
    - Utilizar la técnica de *stratified k-fold cross-validation* (k=10) *with random seed* antes del paso de clasificación.
    - Aplicar la tarea de clasificación en conjunto con el *stratified* 10-*fold cross-validation* usando las redes neuronales artificiales (ANN).
        i. Explorar 3 topologías del modelo de ANN *feed-forward back-propagation* (sugerencia: variar la cantidad de capas ocultas, cantidad de neuronas por capas, y función de activación)
        ii. Para cada topología, optimizar los hiperparámetros *learning rate* en el rango de 0.1 a 0.5 (con paso de 0.2) y el número de interacciones *EPOCH* en el rango de 1 a 200 (con pasos de 20 unidades).
    - Se debe obtener un resultado de clasificación basado en la media del área bajo la curva ROC cerca de 0.80 (*mean AUC >= 0.80)* para al menos una topología de ANN.
    - Es obligatorio mostrar la trazabilidad del método durante la ejecución del programa:
        i. El *Dataset* original y normalizado. **(1 puntos)**
        ii. Los resultados de clasificación obtenidos por los distintos modelos de clasificación:
            1. Confeccionar una tabla que muestre la configuración y desempeño de cada modelo de clasificación evaluado. La configuración se refiere a: topología + hiperparámetros, y el desempeño a: el promedio y desviación estándar de las métricas de *accuracy, precission, recall F1-score, MCC, AUC y loss* (*invest a Little effort to research about the loss function*). **(4 puntos)**
            2. De la tabla confeccionada, seleccionar el mejor modelo de acuerdo al mayor valor promedio de la métrica *AUC*. **(1 punto)**
            3. Mostrar la matriz de confusión obtenida para el modelo seleccionado. **(1 punto)**
            4. Mostrar el plot del *área bajo la curva ROC* para el modelo seleccionado. **(1 punto)**
            5. Mostrar un plot de *(precission vs recall)* para el modelo seleccionado. **(1 punto)**
            6. Mostrar un plot de *mean of loss function vs epochs* para el modelo seleccionado. Este plot nos permite verificar si existe o no *overfitting* en el modelo. **(1 punto)**

- ¿Que no debo usar durante la fase de entrenamiento?
    i. Optimizadores de aprendizaje (ej: SGD, RMSProp, ADAM, etc).
    ii. Criterios de parada anticipados (early stopping).
- 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).