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

Proyecto Final

- Actividad obligatoria:
  - Extender uno de los temas visto en clases. Se considerará extensión, el aumento de la tarea previa en un 30% mínimo. Este aumento debe ser basado en temas no obvios y no abordados en clases. Ej: el uso de deep learning para comparar modelos de ANNs vs Deep ANNs, Text Mining vs NLP (cosine similarity vs SVM, ANN, LLM, etc), algoritmos de clustering vistos en clases vs otros nuevos no vistos, etc. Se sugiere que usen los temas de trabajo de titulación. (**6 puntos**)
  - La extensión debe ser aplicada a un *datasets* de la vida real siempre que sea posible (un problema real). No existe límite en el uso de los *datasets* (ej: pueden usar 2 o 3 diferentes para ver como se comporta el desempeño de los modelos analizados en diferentes espacios de características). (**2 puntos)**
  - Para cualquier proyecto de clasificación supervizada se debe usar el método *stratified 10-fold cross-validation*, *mean of ACC, REC, PRE, AUC, F1-score, loss*, y la desviación estandar de cada métrica.
  - Para proyectos de classificación no supervizada, se deben calcular las siguientes métricas para mediar el desempeño y determinar el mejor número de clusters basado en *Rand index, Mutual Information based scores, Homogeneity, completeness and V-measure, Fowlkes-Mallows scores, Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index Contingency Matrix, Pair Confusion Matrix*.
    Para proyectos de *forcasting*, el promedio y desviación estándar de: *Mean Square Error (MSE), Mean Absolute Error (MAE), R-Squared , Mean Absolute Percentage Error, Root Mean Squared Error, Normalized Root Mean Squared Error, Weighted Absolute Percentage Error, Weighted Mean Absolute Percentage Error.* (**1.5 puntos)**
  - Se deben realizar los plots: AUC-ROC, P-R, y Loss vs Epochs (para el training y validation juntos). (**1.5 puntos)**
  - Mostrar la matriz de confusión general. (**1 punto)**
  - Realizar una presentación en PowerPoint basada en los siguientes aspectos (**8 puntos**):
    a. Portada (título del trabajo, nombre del estudiante).
    b. Introducción (Cuál es el problema que se quiere resolver, el objetivo para llegar a la solución y que técnica(s) utilizaron para resolver)

c. Materiales y métodos (descripción de la(s) técnica(s) utilizada(s), metodología usada para evaluarla(s) (incluyendo información del(los) dataset(s) empleado(s) para validarla, métricas de validación, etc)
d. Resultados y discusión (resultados obtenidos por el uso de la(s) técnica(s) y discusión de estos, plots, etc)
e. Conclusiones (los aspectos más importantes obtenidos como parte de su investigación)

Generales:

- Del acto de entrega:
  1. Cargar al D2L el proyecto implementado (un fichero Python bien documentado que ejecute el proyecto completo de inicio a fin) y la presentación dentro del plazo de entrega. El atraso o no envío de la actividad será estrictamente penalizado con nota CERO.
- Del acto de presentación:
  1. La presentación es obligatoria y debe ajustarse a un tiempo de 12 minutos por estudiante (no más de 13 diapositivas). Cuando se exceda el tiempo el profesor detendrá la presentación y recibirá penalización del 30% por el incumplimiento de la actividad. Además, constaremos con 3 minutos para preguntas y respuestas.
  2. Es obligatoria la presentación de cada estudiante. Violar este punto significa que recibe nota CERO en la actividad independientemente de que haya subido al D2L los archivos solicitados.
- Sobre las fechas de entrega:
  1. Tendrán la entrega hasta el cierre de la carpeta en el D2L (antes de la presentación).
- Sobre las fechas de presentación:
  1. En la semana 15 (jueves 30-nov. de 11:30-13:00), En la semana 16 (martes 5-dic. Y jueves 7-dic de 11:30-13:00), el día que nos toca el examen final del curso (miércoles 20 de 9:00 -11:00).