Bigdata **Taller No. 1**
Fecha de entrega: 24 de septiembre de 2025, 12m.

1. Considere el conjunto de datos "*data_cluster.mat*", para lo cual usted debe identificar el número de neuronas en el SOM y de clusters usando K-means que considere que mejor se ajustan al problema.

- Describa la experimentación realizada para cada caso.
- Obtenga el número de grupos o neuronas que considere mejor se ajustan al problema. Explique y justifique su respuesta.

2. Use los datos tomados del repositorio de la UCI (*https://archive.ics.uci.edu/ml/datasets/Avila*), la cual es una base de datos de un problema multiclase para la clasificación del copista de fragmentos de un códice del siglo XII, como lo es la biblia de Ávila. Para hacer el análisis más simple puede eliminar o unir las cuatro (4) clases con menor cantidad de datos.

Cada grupo debe realizar un análisis de la efectividad del agrupamiento; de acuerdo a los procedimientos de aprendizaje no supervisado (Coeficiente de Silhouette y gráfica de Elbow).

- Diseñe dos modelos de clasificación usando modelos no supervisados (K-means y mapas auto-organizados).
- Realice variaciones de número de grupos en el caso de K-means y grillas de salida para los mapas. Mínimo 10 variaciones por cada experimento en el cual establezca la calidad de los grupos propuestos.
- *Enfoque Semi-supervisado:* etiquete la información con respecto al tipo de agrupación que le correspondió, como clases de salida del problema.
- Realice validación cruzada y estime el error de generalización para cada modelo. Adicionalmente, obtenga la matriz de confusión tanto para los datos de entrenamiento como de validación .
- Analice sus resultados y describa los procedimientos en un informe.