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

Proyecto 8: Ejercicio usando agrupaciones (*clustering*)

Problema:
Dado el conjunto de datos "CSV_ETS295_class_smote_5_100(clustering).csv" (https://estudusfqedu-my.sharepoint.com/:x:/g/personal/nperez_usfq_edu_ec/Ef9kRu3FV05EucrdoVb9-qEBR-g0pHLavGHpV56lyG3k6w?rtime=B5zIebXG2kg), se desea aplicar un algoritmo de agrupamiento que permita realizar asignaciones de los datos a los posibles ***clusters*** formados.
Para la realización de la tarea se exige:

- Cada equipo debe usar dos de los algoritmos presentados a continuación y no pueden repetirse entre equipos:
    - *Affinity Propagation, (Equipo1)*
    - *Agglomerative Hierarchical Clustering, (Equipo2)*
    - *BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), (Equipo1)*
    - *DBSCAN (Density-Based Spatial Clustering of Applications with Noise), (Equipo6)*
    - *Mean Shift Clustering, (Equipo2)*
    - *Spectral-clustering, (Equipo3)*
    - *BFR (Equipo3)*
    - *CURE (Clustering Using Representative), (Equipo4)*
    - *Mini batch K-means (Equipo5)*
    - *Expectation-Maximization (Equipo4)*
    - *Gaussian Mixture Models (Equipo5)*
    - *K-means (Equipo6)*
- Cada estudiante debe realizar un ***research*** sobre el algoritmo asignado. De forma tal que pueda presentar y discutir sobre la teoría de *clustering* y a su vez del algoritmo desarrollado.
- Cada estudiante debe realizar un ***research*** sobre el método t-SNE (https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding), de forma tal que pueda implementarlo para la resolución de literales relacionados a la proyección de los datos y resultados del algoritmo de *clustering* empleado.
- Es obligatorio mostrar la trazabilidad del método de *clustering*:
    - Normalización del *dataset* usando el método *min-max.* **(0.5 punto)**
    - Optimización del valor de *k* (numero de *clusters*) en un intervalo de ***k=2..8.*** **(3 puntos)**

- El proceso de selección del valor de ***k*** óptimo de acuerdo al algoritmo utilizado. Se debe proveer una forma de analizar la convergencia del método hacia un ***k*** óptimo. Puede que algunos algoritmos no trabajen con el concepto de *centroide o clustroide* y por tanto, no se pueda usar el *plot K vs D*. Para esos casos, se debe investigar alguna alternativa que nos permita determinar el valor de ***k*** óptimo. **(3 puntos)**
- Imprimir el valor de ***k*** óptimo de acuerdo a su selección. **(0.5 punto)**
- Imprimir el identificador (índice en el fichero) de las instancias (vectores) pertenecientes a cada *cluster*. **(1 punto)**
- Mostrar el plot t-SNE para el espacio original de los datos normalizados. (**0.5 puntos**)
- Mostrar el plot t-SNE (tres plot en total) después de aplicado el método de *clustering* al *dataset* normalizado para los valores de ***k*** óptimo, ***k-1*** y ***k+1***. De esta forma se podrá visualizar los *clusters* en la vecindad del valor de ***k*** óptimo determinado por el inciso (4). **(1.5 puntos)**

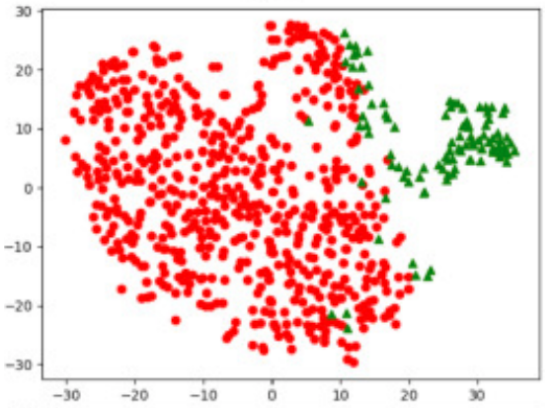

Fig. 1 Ejemplo de representación de dos *clusters* (rojo y verde) usando la técnica t-SNE.

+1 punto: Detectar y eliminar *outliers* usando *clustering*. Se debe demostrar de alguna forma el procedimiento si fuera aplicable.

- Cargar al D2L los códigos implementados dentro del plazo de entrega.