



CURSO: CMP 0575 - TÓPICOS 2 (DATA MINING)
COLEGIO: POLITÉCNICO
Semestre: Primer Semestre 2019/2020

Proyecto 7: Ejercicio usando agrupaciones (clustering)

Problema:

1. Dado el conjunto de datos “**Dataset(Clustering).csv**” (https://estudusfqedu-my.sharepoint.com/:f:/g/personal/nperez_usfq_edu_ec/EgCEZlzexsJMpC2E3-KpSLYBb7QtX5Ht6Fwv11NtiE5xwA?e=gADxaM) proporcionado en el D2L, 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:
 - Usar **uno** de los tres algoritmos vistos en clase de acuerdo a la distribución por equipo realizada: **k-means**, **BRF** y **CURE**.
 - Es obligatorio mostrar la trazabilidad del método durante la ejecución del programa:
 - i. El valor del **centroide - clustroide** de los **clusters** óptimos de solución.
 - ii. El proceso de selección de **k** (cantidad de **clusters**), o sea como converge la distancia promedio de los **centroides - clustroides** a medida que aumenta el número de **clusters** (optimizar **k** en un rango de 1..10). Un plot parecido al que se presentó en la clase de **K vs D** (distancia promedio).
 - iii. La cantidad de instancias asignadas a cada uno de los **cluster** creados.
 - iv. Es obligatorio realizar una **presentación científica (NO FILOSÓFICA)** al profesor con todos los elementos requeridos.
 - v. La programación del algoritmo utilizado debe ser en el lenguaje **Python , JAVA o C++**
 - Cargar al D2L los códigos implementados (fichero compactado que incluye el ejecutable ej: el .JAR de java) y la presentación dentro del plazo de entrega.