Universidad del Rosario
Aprendizaje automático de máquina I - Grupo 1
Proyecto 3

Éste proyecto puede trabajarse de forma individual o en parejas. Deberá entregarse en formato .ipynb y adicionalmente en formato .html a más tardar el jueves 1 de Junio

---

El ejercicio consite en implementar algunos algoritmos de clustering en una porción de un conjunto de datos sobre casos de crimen en Cambridge. Los atributos (aunque sólo utilizaremos tres de ellos) son

- File Number: Identificación del reporte.
- Date of Report: Fecha del reporte
- Crime Date Time: Fecha del crimen
- Crime: Tipo de crimen.
- Reporting Area: Código del área reportada.
- Neighborhood: Nombre del barrio.
- Location: Información sobre la calle.

Observación: Los pasos 4 al 8 (inclusive) pesan 2 puntos. Los demás pesan 1 punto.

1. No copiar los enunciados de las preguntas. En lugar de esto, usar celdas de texto entre las celdas de código para describir con sus palabras todo procedimiento llevado a cabo a lo largo del código.
2. Se trabajará únicamente con las columnas **Crime**, **Reporting Area** y **Neighborhood**. Eliminar las demás y convertir a Dummies las variables categóricas. Tome una muestra aleatoria de 5000 registros.
3. Utilizar Análisis de Componentes Principales (PCA) para reducir el conjunto de datos a uno de dimensión 2. Puede guiarse del taller de SVM.
4. Ahora que el conjunto de datos es 2-dimensional, hacer una visualización de los mismos. Puede usarse, por ejemplo, un scatterplot. A simple vista deferían identificarse grupos de datos. Cuántos grupos ve?
5. Implementar el método DBSCAN de SKLearn. Realice una visualización en la que se distingan colores para los diferentes clusters establecidos por el algoritmo. ¿Qué ocurre si damos a épsilon un valor menor que 1? ¿Por qué? ¿Qué parámetros logran una buena agrupación de los datos?
6. Implementar el algoritmo de K-means de SKLearn. Realizar una visualización de los datos al igual que en el item anterior.

7. Implementar alguno de los algoritmos aglomerativos que proporciona SKlearn. Realizar una visualización de los datos al igual que en los items anteriores.

8. Extraer los índices de los registros contenidos **alguno** de los clusters. Estos corresponden con los mismos índices en el dataset al que se le aplicó PCA. ¿Qué tienen en comun los datos de dicho cluster?

9. Escriba sus conclusiones del ejercicio.