1. Objetivo del laboratorio

*Desarrollar de forma autónoma **tres Notebooks** que permitan explicar distintas hipótesis a partir de varios datasets de entrada, mediante la preparación y visualización de estos.*

2. Elementos a utilizar:

- Lenguaje Python
- Librería numérica *NumPy*, pandas, scikit-learn y gráfica *Matplotlib*
- Entorno Anaconda
- Editor Jupyter

3. Práctica 1 (impacto de las restricciones de tráfico en la calidad del aire)

Objetivo

El 30 de diciembre de 2016 el Ayuntamiento de Madrid empieza a restringir el tráfico al centro de la ciudad mediante la aplicación del escenario 3 del protocolo de contaminación. Esta fue la primera vez que se aplicó un escenario tan restrictivo. Demuestra mediante visualizaciones de datos que ha ocurrido después de ello.

Para dicho problema, haremos uso de un dataset con datos sobre la calidad del aire en Madrid desde 2001. El dataset contiene un archivo csv por año. Dentro de cada uno de ellos habrá medidas diarias de distintas estaciones. Estas estaciones miden diferentes parámetros y no siempre todos se registran en cada una de las estaciones.

https://www.kaggle.com/decide-soluciones/air-quality-madrid/downloads/csvs_per_year.zip/5

1.- (1 punto) Lo primero que tendremos que hacer es cargar todos los archivos csv en un mismo DataFrame para poder manipularlos. Habrá que comprobar si existen datos redundantes o anómalos

Para medir la calidad del aire de un día obtendremos el peor valor de cada estación meteorológica ese día y después calcularemos la media de estos. Con dicho valor se estimará como de buena es la calidad teniendo en cuenta la siguiente tabla. Al final habrá que hacer una transformación de valores continuos a categóricos. Proporcionar dicha información en un DataFrame y luego guardarlo en un archivo csv.

Valor del índice	Calidad del aire	Color
0 - 49	Buena	Verde
50 - 99	Admisible	Amarillo
100 - 150	Mala	Rojo
> 150	Muy mala	Marrón

2.- (0,5 puntos) Explica visualmente como se distribuyen las calidades del aire entre buena, admisible, mala y muy mala. Realiza para ello un pie chart. Que se puede concluir de dicho gráfico.

3.- (1 punto) Una vez obtenida la calidad del aire para cada día, calcula la calidad media de cada mes para que sea más fácil saber si la contaminación ha disminuido desde que se restringió el acceso al centro de Madrid. Explica visualmente si los protocolos de restricciones de tráfico han tenido éxito. Utiliza para ello un diagrama de barras. ¿Podemos encontrar la diferencia entre lo que ocurría antes de aplicar el protocolo y después?

4,- (1 punto) Por último, queremos tratar de entender los cambios de la calidad del aire en estos últimos años. Obtén un diagrama de cajas donde cada caja corresponda a un año y haz una interpretación de los resultados.

4. Práctica 2 (el impacto del terrorismo a lo largo de la historia)

Objetivo

El problema del terrorismo es actualmente considerado uno de los más graves para la población. Es un hecho que este ha ido cambiando a lo largo de la historia. Antaño los grupos terroristas actuaban en su propio país o en los países con los que compartía frontera. Hoy en día el terrorismo se considera como un problema global en el que prácticamente cualquier país puede sufrir sus consecuencias. En esta práctica intentaremos entender varias situaciones tanto actuales cómo históricas.

Para ello haremos uso de un dataset con datos históricos de terrorismo desde el año 1970. El dataset contiene es un archivo csv con distintos atributos para cada atentado registrado.

https://www.kaggle.com/START-UMD/gtd/downloads/globalterrorismdb_0718dist.csv/3

1.- (1 punto) Para preparar los datos vamos a crear un DataFrame que luego guardaremos en un csv donde se almacenará sólo la información necesaria. Esta es: año (iyear), país (country_txt), región (región_txt), grupo terrorista (gname) y número de victimas (nkill). Explica si nos encontramos en la época más sangrienta de la historia. Para ello dibuja un diagrama de línea con el total de victimas de cada año. A continuación, vamos a intentar explicar si las regiones del mundo han sido igual de violentas siempre. Para ello crearemos otro diagrama de líneas en el que se reflejen el número de víctimas anuales de cada región. Habrá que comprobar si existen datos anómalos. ¿Podemos indicar cuales han sido los años donde el terrorismo ha tenido más impacto? ¿Cuál ha sido el impacto del 11S en el terrorismo a nivel mundial?

2.- (0,5 punto) Otro dato interesante sería conocer cual ha sido el grupo terrorista que más victimas ha causado. Para ello obtén un pie chart con la información de los 10 grupos más sanguinarios de la historia. Interpreta los datos.

3.- (1 punto) Siempre es interesante cruzar los datos de un dataset con otro. A priori parece ser que muchas veces el terrorismo surge con más fuerza en épocas de crisis. Podemos dibujar en un diagrama de dispersión el número de victimas con el PIB de un país. Para ello encuentra un dataset que tenga los PIBs de los distintos países. Dibuja dos diagramas de dispersión: uno con los 10 países con el PIB más bajo en el último año y el número de victimas en atentados y otro con los 10 países con más victimas en el último año y sus PIBs. Interpreta los datos.

5. Práctica 3 (Principal Component Analysis)

Objetivo

Existen casos en que las variables no se pueden representar visualmente debido a que necesitaríamos varias dimensiones para ello. Para evitar esto, existe una metodología la cual, un set de datos multidimensional,

podemos transformarlo para poder explicar gran parte de la información en 2 o 3 dimensiones. Dicha metodología se conoce con el nombre de Principal Component Analysis (PCA). Vamos a aplicarlo a un set de datos que está colgado en Moodle y vamos a dar una serie de explicaciones de que ocurre.

1.- (0,5 puntos) Lo primero que habrá que hacer será estandarizar los datos para que las diferencias de rango no supongan un problema a la hora de procesar la información. Usa para ello el método StandardScaler de la librería scikitk-learn.

2.- (2 puntos) El segundo paso será a partir de los datos anteriores, obtener los autovalores (eigenvalues) y los autovectores (eigenvectors) que nos permitan explicar cuantos componentes necesitamos para representar los datos iniciales. Para ello primer habrá que obtener la matriz de covarianza mediante el método cov de numpy y después aplicarle a dicha matriz el método linalg.eig también de numpy. Obten un DataFrame con el porcentaje de varianza y el acumulado por cada componente. Explica que quieren decir estos datos.

3.- (1,5 puntos) Por último queremos representar gráficamente los individuos de nuestro dataset pero usando los valores de los componentes principales obtenidas. Obtén un diagrama de dispersión en 2 dimensiones y comenta que has interpretado en él.

6. Forma de entrega del laboratorio:

La entrega consistirá en un fichero comprimido RAR con nombre LAB01-GRUPOxx.RAR subido a la tarea **LAB1** que **contenga únicamente**

1. **Por cada práctica** un notebook de Jupyter (archivos con extensión **.ipynb**).
2. La **memoria del laboratorio** que se irá construyendo en el Notebook de manera que se explique todo lo que se hace.

Las entregas que no se ajusten exactamente a esta norma NO SERÁN EVALUADAS.

7. Rúbrica de la Práctica:

1. IMPLEMENTACIÓN: Multiplica la nota del trabajo por 0/1

Siendo una práctica de Data Mining, todos los aspectos de programación se dan por supuesto. La implementación será:

- Original: Código fuente no copiado de internet. Grupos con igual código fuente serán suspendidos
- Correcta: El programa funciona y ejecuta correctamente todo lo planteado en los apartados de cada práctica.
- Comentada: Inclusión (**obligatoria**) de comentarios.
- En las gráficas que se realicen proporciona todos los datos que creas necesarios.

2. MEMORIA DEL LABORATORIO

Obligatorio redacción clara y correcta ortográfica/gramaticalmente. Cada paso que se haga tiene que estar justificado.