1. Objetivo del laboratorio

*Desarrollar de forma autónoma **un Notebook** 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
- Librerías como *NumPy*, pandas, scikit-learn y *Matplotlib*
- Entorno Anaconda
- Editor Jupyter

3. Práctica 1 (Vacunación COVID)

Valor (2,5 puntos)

A finales de 2020 empezó la vacunación del COVID-19 que ha producido la mayor pandemia mundial que se recuerda. Algunos países en vista de los problemas que puedan causar futuras pandemias quieren saber qué país está llevando el plan de vacunación más eficiente. En España, el Centro de Biología Molecular Severo Ochoa donde participa la Doctora Margarita del Val. Para ello vamos a obtener cual es la evolución del número de vacunados en el tiempo usando el set de datos **vacunaciones.csv**.

1.- (1,5 puntos) Lo primero será evitar los "missing values" de la columna "people vaccinated". Para ello cogeremos los valores de los 3 días anteriores (si existen, en caso contrario dos o uno) y su media para rellenar dicha celda.

2.- (1 punto) De aquellos 5 países que han vacunado más días en total compara su evolución en el tiempo (no importa que en algunas fechas no coincidan). ¿Qué diagrama has usado? ¿Porqué? Teniendo en cuenta que el mejor plan de vacunación es el que se mantiene más constante ¿Cuál es el país que mejor está llevando a cabo la vacunación? Haz una interpretación de dicho plan de vacunación.

4. Práctica 2 (Sensores atmosféricos)

Valor (3,5 puntos)

Se cuenta con los datos del año 2020 de distintas mediciones de un sensor medioambiental situado en el campus de la Universidad Francisco de Vitoria. Para analizar la información recogida haremos uso de un dataset proporcionado por la propia Universidad llamado "ozone".

1.- (0,75 puntos) Llevar a cabo el estudio de los outliers de 3 variables (Wind_speed, Ozone_reading, Visibility) de forma unidimensional. ¿Qué tipo de gráfico es necesario emplear? Interpretar los datos obtenidos en cada caso.

2.- (0,75 puntos) Estudiar la intersección y la unión de outliers entre las variables "Wind_speed" y "Visibility".

3.- (1 punto) Categoriza algunas de las variables y lleva a cabo representaciones que nos permitan relacionar diferentes variables ¿Qué nos dicen los datos?

4.- (1 punto) Se desea saber cómo se distribuyen y cuáles son las frecuencias de las principales variables respecto a los días de medición. Lleva a cabo la representación más útil (sólo una). Justificar la respuesta.

5. Práctica 3 (Reducción de la dimensionalidad)

Valor (4 puntos)

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 en 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 Canvas llamado USA.xlsx y vamos a dar una serie de explicaciones de que ocurre.

1.- (1 punto) 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 scikit-learn.
2.- (1 punto) El segundo paso será a partir de los datos anteriores, obtener los autovalores (eigenvalues) y los autovectores (eigenvectors) que nos permitan explicar cuántos componentes necesitamos para representar los datos iniciales. Para ello, en primer lugar, 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. Obtén un DataFrame con el porcentaje de varianza y el acumulado por cada componente. Explica qué quieren decir estos datos. ¿Cuánta información perdemos con 2 componentes? ¿Cuánta información representamos con 3 componentes?
3.- (1 punto) Por último queremos representar gráficamente las ciudades de nuestro dataset, pero usando los valores de las componentes principales obtenidas. Obtén un diagrama de dispersión en 3 dimensiones y comenta qué has interpretado en él. Es necesario que el diagrama contenga toda la información necesaria. Habrá que interpretar qué información proporciona el eje X, Y y el Z. Por último, elegir al menos 4 ciudades de forma aleatoria y explicar qué pasa con ellas.
4.- (1 punto) Realiza los mismos pasos que en los pasos anteriores usando la librería scikit-learn. Compara los resultados y coméntalos.

6. Forma de entrega del laboratorio:

La entrega consistirá en un fichero comprimido RAR con nombre LAB01_GrupoX.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.