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

3. Práctica 1 (Predictor enfermedades cardiovasculares)

Objetivo (2 puntos)

Un médico se acaba de trasladar a trabajar a una zona rural. Debido a la gran problemática que surge con la dispersión geográfica de los pacientes, se quiere implementar un sistema que pueda predecir varios parámetros. Algunos de ellos pueden influir en adquirir una enfermedad cardiovascular. En el caso de necesitar una intervención se podrá prever parte del material que necesitaría para una intervención. Para cada uno de los siguientes casos: calcula y establece que tipo de relación hay entre las variables, dibuja un diagrama de dispersión con los casos en el que también se incluya el modelo obtenido y por último haz una predicción con un dato cualquiera.

1) La presión arterial es uno de los factores que está relacionado con las enfermedades cardiovasculares. Utiliza el archivo presion.csv para poder calcular la posible presión arterial de un paciente con respecto a su edad. (0,4 puntos)
2) Los altos niveles de colesterol en sangre están también relacionados con este tipo de enfermedades. Utiliza el archivo colesterol.csv para poder calcular la posible cantidad de colesterol en sangre con respecto al peso del paciente. (0,4 puntos)
3) El consumo excesivo de alcohol está también relacionado con la adquisición de enfermedades cardiovasculares. Utiliza el archivo alcohol.csv para poder calcular el consumo de alcohol de un paciente dependiendo de su edad. (0,4 puntos)
4) En caso de que un paciente necesite ser intervenido, se quiere calcular la longitud del catéter. Utiliza el archivo cateter.csv para poder calcular la cantidad de necesaria de materiales a la hora de intervenir al paciente. (0,4 puntos)
5) Explica cómo funcionaría el posible sistema creado para calcular los parámetros descritos arriba. (0,4 puntos)

4. Práctica 2 (Cómo expandir artículos)

Objetivo (1,5 puntos)

El blog sobre Machine Learning https://machinelearningmastery.com/ quiere saber qué características han de tener sus posts para que luego estos sean compartidos lo máximo posible por sus lectores. Para ello tiene información sobre posts anteriores. La información se encuentra en el archivo artículos.csv y está compuesta por: número de palabras, número de links a otras páginas, número de comentarios de usuarios, cantidad de contenido multimedia (videos, fotografías, etc), días desde que se publicó y número de veces que se ha compartido.

1) Encuentra un modelo para dicho set de datos. Haz una interpretación de él. Por último, dibújalo si es posible teniendo en cuenta las variables más influyentes. (0,5 puntos)
2) Haz un análisis de residuos del modelo lo más exhaustivo posible. (0,5 puntos)
3) Utiliza los intervalos de confianza para obtener la mayor información posible con ellos. (0,5 puntos)

5. Práctica 3 (Marketing telefónico)

Objetivo (1,5 puntos)

Un banco quiere establecer un modelo para saber qué tipo de campañas tiene que desarrollar para tener éxito cuando contacta a un cliente vía telefónica. Para ello tiene un historial de varios clientes a los que se les ha ofrecido nuevos paquetes para contratar, en algunos ha sido un éxito y en otro no.

1) Para poder crear el modelo es necesario hacer un preprocesamiento previo de la información (0,75 puntos)
2) Crea un modelo que permita saber que variables afectan más a la hora de predecir si un cliente contratará un nuevo producto o no. Una vez obtenido haz una interpretación lo más completa posible. (0,75 puntos)

6. Práctica 4 (Agrupamiento de jugadores en videojuegos)

Objetivo (3 puntos)

Bluehole, la empresa encargada del videojuego PlayerUnknown's Battlegrounds quiere introducir nuevos paquetes dependiendo del tipo de jugador. Para ello dispone de estadísticas de los 200 mejores jugadores. Aplica un algoritmo de manera que se obtengan dichos grupos.

1) Utiliza varias configuraciones teniendo en cuenta el número de grupos que se creará y cambiando cómo se mide la distancia entre individuos. Crea una tabla donde se incluya toda la información y el número necesario de iteraciones para llegar a dicha solución. Se considera la mejor solución aquella que necesite menos iteraciones. (1 punto)
2) Con la mejor configuración del punto anterior. Utiliza dos criterios para elegir el lugar inicial del punto central de los grupos. Dibuja cómo se van modificando los grupos y cómo van cambiando sus centroides en cada iteración. Obten una conclusión acerca de donde deberían situarse los centroides (1 punto)
3) Estudia que técnicas de preprocesamiento se podrían incluir en base al error cometido en cada clúster. (1 punto)

7. Práctica 5 (Taxonomía de animales)

Objetivo (2 puntos)

El zoo de Madrid pretende hacer una renovación de sus instalaciones. Una de las propuestas es disponer de juegos interactivos que permitan a los visitantes aprender más acerca de los animales que hay en el zoo. Para ello se dispone del archivo csv “zoo” con las características de los distintos animales. Crear un modelo que agrupe los animales y establezca como se relacionan entre ellos jerárquicamente.

1) Utiliza varias configuraciones para el modelo aplicando “single linkage” y teniendo en cuenta los tipos de distancias entre elementos. ¿Cuál es la k del modelo? (1 punto)
2) Dibuja un dendograma con los clústers obtenidos. Explica alguna de las relaciones interesantes que puedas encontrar. (1 punto)

8. Forma de entrega del laboratorio:

La entrega consistirá en un fichero comprimido RAR con nombre LAB04-GRUPOxx.RAR subido a la tarea **LAB4** 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.

9. 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.