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 (Calidad del vino)

Objetivo (2 puntos)

Una tienda online de venta de vino quiere hacer un estudio de mercado. Han descubierto que muchos clientes valoran como buenos o muy buenos vinos que no son tan conocidos. Con la intención de adelantarse a la competencia quiere encontrar qué factores podrían influir más en esta elección. Usando el archivo vinos.csv y para 5 de las características que consideres que más influyen: calcula y establece qué tipo de relación hay entre ellas, dibuja un diagrama de dispersión con los casos en el que también se incluya el modelo obtenido y por último haz una predicción con varios datos.

1) ¿Qué diferencias hay entre los 5 modelos? (1,5 puntos)
2) Explica cómo funcionaría el posible sistema para clasificar vinos (simplifica la variable "quality" creando 4 clases: muy malo, malo, bueno, muy bueno) creado con los 3 parámetros que más influyen entre los descritos arriba. (0,5 puntos)

4. Práctica 2 (Indicador de diabetes y colesterol)

Objetivo (3 puntos)

La diabetes es una de las enfermedades crónicas más prevalentes en la que las personas pierden la capacidad de regular eficazmente los niveles de glucosa en sangre reduciendo su calidad y esperanza de vida. Se caracteriza porque el organismo no produce suficiente insulina o porque es incapaz de utilizar la insulina producida con la eficacia necesaria. Puede tener complicaciones como cardiopatías, pérdida de visión o amputación de miembros. Aunque esta enfermedad no tiene cura, la pérdida de peso, el ejercicio y una dieta saludable junto con tratamientos médicos pueden reducir el riesgo y los daños producidos por la diabetes. Por ello, es importante predecir el riesgo de diabetes para reconducir la evolución de la enfermedad. Utilizando el dataset diabetes.csv responde a las siguientes cuestiones.

1) Realiza el preprocesamiento necesario para predecir el riesgo de diabetes según el IMC (Indice de Masa Corporal). Realiza un diagrama de dispersión e interpreta los resultados para un IMC alto y otro bajo. (1 punto)
2) Realiza el preprocesamiento necesario para predecir el riesgo de colesterol alto según el IMC (Indice de Masa Corporal). Realiza un diagrama de dispersión e interpreta los resultados. (1 punto)
3) ¿Encuentras alguna relación entre los estudios anteriores que ayuden a la prevención de la diabetes? Justifícalo. (1 punto)

5. Práctica 3 (Puestos de enfermería)

Objetivo (2 puntos)

Una empresa de colocación de trabajadores pretende hacer una aplicación para ser más eficiente en los enfermeros y enfermeras que asignan para cuidados en hogares. Para ello se dispone del archivo csv "enfermeria" con las características de las distintas personas que pueden optar a los puestos. Crear un modelo que agrupe los candidatos y establezca como se relacionan entre ellos jerárquicamente.

1) Utiliza varias configuraciones para el modelo que más se adapte y teniendo en cuenta los tipos de distancias entre elementos. ¿Cuál es la k del modelo? (1 puntos)
2) Dibuja un dendograma con los clusters obtenidos. Explica alguna de las relaciones interesantes que puedas encontrar. (1 punto)

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. Obtén una conclusión acerca de dónde deberían situarse los centroides. (1 punto)
3) Estudia qué técnicas de preprocesamiento se podrían incluir en base al error cometido en cada cluster. (1 punto)

7. 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.

8. 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.