1. Objetivo del laboratorio

*Desarrollar de forma autónoma **diferentes Notebook** que permitan explicar distintas hipótesis a partir de varios datasets de entrada, mediante la preparación y visualización de los datos.*

2. Elementos a utilizar:

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

3. Ejercicio 1 (Regresión lineal simple)

Objetivo (3 puntos)

Una prestigiosa empresa productora de vinos quiere construir un modelo de regresión que le permita predecir el porcentaje de alcohol de un vino en base a ciertas características (acidez, azúcar residual, azufre, cloruros, densidad, tipo de vino, sulfatos y pH).

Para ello han construido un dataset, denominado **wine_alcohol.csv**, que contiene la información de todos los vinos que han sido producidos en los últimos años con el objetivo de construir un modelo y realizar las siguientes acciones:

1) Realizar todo el proceso de preparación, limpieza y análisis de los datos con el objetivo de identificar aquellas variables candidatas para el proceso de regresión (0.5 puntos).
2) Realizar un análisis sobre los atributos del dataset con el objetivo de entender las posibles relaciones que existen entre ellas y a continuación elige los dos atributos que mayor relación guardan con el atributo objetivo (nivel de alcohol) (1 punto). En necesario justificar la elección de los atributos mediante la utilización de datos empíricos.
3) Construir un modelo de regresión lineal mediante un proceso analítico para una de las variables seleccionadas en el punto anterior (1 punto).
4) Calcular el error en el modelo obtenidos en el punto anterior (0.5 puntos).

4. Ejercicio 2 (Regresión lineal múltiple)

Objetivo (3 puntos)

La empresa productora de vinos cree que es posible construir un modelo de mejor calidad que pueda predecir mejor el grado de alcohol de los futuros vinos si se utilizan varios atributos del dataset. Para ello se solicita la creación de nuevos modelos realizando las siguientes acciones:

1) Crear los diferentes conjuntos de entrenamiento y test para el dataset (0.25 puntos).
2) Construir un modelo de regresión que utilice al menos 4 de los atributos (es posible utilizar más atributos) disponibles en el dataset mediante la utilización de un proceso aprendizaje iterativo. (1.5 puntos).
3) Calcular el error del modelo (0.5 puntos).
4) Evaluar el modelo anterior con el conjunto de test construido anteriormente y explicar el resultado obtenido. (0.25 puntos).
5) Calcular los intervalos de confianza para cada uno de los coeficientes de regresión con un intervalo de confianza del 75% y explica el resultado obtenido (0.5 puntos).

5. Ejercicio 3 (Agrupamiento)

Objetivo (2 puntos)

La empresa productora de vinos no está segura de que la empresa de analiza la calidad de los diferentes vinos esté realizando el trabajo adecuadamente por lo que ha decidido analizar si los diferentes vinos que han sido producidos en los últimos años realmente tienen la calidad identificada.

Para ello se debe aplicar un algoritmo de agrupamiento para comprobar si los vinos realmente se agrupan de manera similar a como han sido valorados por la empresa evaluadora (quality) y realizar las siguientes acciones:

1) Realizar todo el proceso de preparación, limpieza, eliminación y análisis de los datos con el objetivo de identificar aquellas variables candidatas para el proceso de agrupamiento (0.5 puntos).
2) Ejecutar diferentes procesos de agrupamiento utilizando el algoritmo K-Means e identificar cual es el valor de k que mejor resultados ofrece (1 punto). En necesario justificar la selección del mejor valor de k mediante la utilización de datos empíricos.
3) Analiza los clústeres obtenidos para el mejor valor de k y responde a las siguientes preguntas:
   - ¿Existe algún tipo de similitud entre los grupos obtenidos y los valores de la variable quality del dataset? (0.5 puntos).
   - ¿Cómo se diferencian los clústeres entre sí? (0.5 puntos).

   En necesario justificar las respuestas mediante la utilización de datos empíricos.

6. Ejercicio 4 (Agrupamiento)

Objetivo (2 puntos)

La empresa productora de vinos desea mejorar el proceso de producción de sus vinos con el objetivo de mejorar las futuras producciones.

Para ello han pensado en utilizar un algoritmo de agrupamiento jerárquico para segmentar los diferentes vinos en base a sus componentes, calidad y nivel de alcohol.

1) Dado el mejor valor de k del apartado anterior selecciona los 10 ejemplos más representativos de cada clúster para generar un nuevo dataset y aplicar agrupamiento jerárquico y responder a las siguientes preguntas:

- ¿Cuál ha sido tu criterio para seleccionar los ejemplos más representativos de cada clúster? (0,5 puntos).

2) Seleccionar un algoritmo de agrupamiento jerárquico y generar los diferentes clústeres en base a los atributos seleccionados (0.75 puntos).
3) Generar y analizar el dendrograma resultante del proceso de agrupamiento. (0,75 puntos)

7. Forma de entrega del laboratorio:

La entrega consistirá en un enlace al directorio de GitHub donde tengáis el jupyter notebook con la resolución de la práctica 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.