### **Práctica 2** **Autores: Pablo Moreno Martínez Macarena Palomares Pastor**

### Descripción del dataset. Por qué es importante y que pregunta/problema pretende responder

El conjunto de datos escogido contiene variables que describen propiedades fisicoquímicas sobre el vino, y una variable respuesta con la calidad del vino. Es un buen conjunto de datos con el que realizar un proyecto limpieza y análisis de datos, debido a la cantidad de atributos con los que podemos crear modelos de clasificación o regresión.

Este dataset se encuentra en el siguiente enlace del repositorio de github:

<https://github.com/pmm2207/Tipologia_y_ciclo_de_vida_PRA2/blob/main/data/winequality-red.csv>

Los campos que contiene son los siguientes:

1 - fixed acidity: acidez fija

2 - volatile acidity: acidez volátil

3 - citric acid: ácido cítrico

4 - residual sugar: azúcar residual

5 - chlorides: cloruros

6 - free sulfur dioxide: dióxido de azufre libre

7 - total sulfur dioxide: dióxido de azufre total

8 - density

9 - pH

10 - sulphates: sulfatos

11 - alcohol

12 - quality: calidad, puntuación entre 0 y 10

### Integración y selección de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

Los datos a analizar son todos los que incluye el dataset, y no se añaden datos procedentes de otros datasets. (Se podría añadir el vino blanco)

### Limpieza de los datos.

### Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

Los datos no contienen valores NA. En cuanto a valores 0, tenemos 132 valores a 0 en el campo *citric acid*. Viendo la distribución de valores 0 con respecto a la calidad del vino……(decidir qué hacemos con dichos valores, yo los dejaría)

### Identifica y gestiona los valores extremos.

Utilizamos gráficas de tipo boxplot para identificar los outliers. (decidir qué hacemos con los outliers, yo los dejaría)

### Análisis de los datos.

### Selección de los grupos de datos que se quieren analizar/comparar (p. e., si se van a comparar grupos de datos, cuales son estos grupos y que tipo de análisis se van a aplicar?)

Primero analizamos la variable respuesta calidad mediante regresión lineal, dejando dicha variable sin transformar

Luego, en la construcción de modelos de clasificación, vamos a clasificar en 2 grupos. Calidad < 7 mala calidad, calidad >7 buena calidad

### Comprobación de la normalidad y homogeneidad de la varianza.

Para normalidad aplicamos test de Shapiro-Wilk, y para homocedasticidad el test de Levene.

La normalidad se cumple en todas las variables. Sin embargo, no hay homogeneidad de la varianza en *residual\_sugar*, *chlorides*, *free\_sulfur\_dioxide*, *sulphates* y *pH*

### Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Hemos realizado:

* Correlación entre variables
* Regresión lineal simple
* Regresión lineal múltiple
* Árboles de decisión (random forest)

### Representación de los resultados a partir de tablas y gráficas. Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica.

### Resolución del problema. A partir de los resultados obtenidos, ¿cuales son las conclusiones? . ¿Los resultados permiten responder al problema?

### Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

### Vídeo

### Contribuciones

| **Contribuciones** | **Firma** |
| --- | --- |
| Investigación previa | PMM, MPP |
| Redacción de las respuestas | PMM, MPP |
| Desarrollo del código | PMM, MPP |