Práctica 2

Autores:

Pablo Moreno Martínez
Macarena Palomares Pastor

1. 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: densidad del vino

9 - pH

10 - sulphates: sulfatos

11 - alcohol

12 - quality: calidad, puntuación entre 0 y 10

2. 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.

3. Limpieza de los datos.

1) 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, consideramos que son valores válidos por lo que se mantienen en el análisis de datos.

2) Identifica y gestiona los valores extremos.

Utilizamos gráficas de tipo boxplot para identificar los outliers. Tras revisar cada una de las variables, concluimos que no se pueden eliminar los outliers encontrados, dado que se tratan de valores posibles dentro del rango de valores de cada variable.

4. Análisis de los datos.

1) 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 la calidad en 2 grupos:

- Calidad < 7 mala calidad
- Calidad >7 buena calidad

2) 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*

3) **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.**

Realizamos un estudio de la correlación entre variables, en primer lugar se busca la correlación con la variable respuesta *quality*, y luego se representa una matriz de correlación entre todas las variables.

Tras saber las variables con mayor correlación, las introducimos a un modelo de regresión lineal. Primero construimos un modelo de regresión lineal simple, y luego construimos un modelo de regresión lineal múltiple añadiendo más variables.

Como los resultados de la regresión no son buenos, probamos con modelo de clasificación, un árbol de decisión de tipo random forest.

5. **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.**

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

7. **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.**

8. **Vídeo**

9. **Contribuciones**

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