Práctica 2: Limpieza y análisis de datos

Autores:

Pablo Moreno Martínez
Macarena Palomares Pastor

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

Hoy en día, la industria del vino utiliza las certificaciones de calidad de los productos para promocionarlos. Se trata de un proceso que lleva mucho tiempo y requiere la evaluación de expertos humanos, lo que hace que este proceso sea muy caro. Además, el precio del vino tinto depende de un concepto bastante abstracto de apreciación del vino por parte de los catadores, cuya opinión puede tener un alto grado de variabilidad. Otro factor vital en la certificación y evaluación de la calidad del vino tinto son las pruebas fisicoquímicas, que se realizan en laboratorio y tienen en cuenta factores como la acidez, el nivel de pH, el azúcar y otras propiedades químicas. Para el mercado del vino tinto sería interesante que la calidad humana de la cata pudiera relacionarse con las propiedades químicas del vino, de modo que los procesos de certificación y de evaluación y garantía de la calidad estuvieran más controlados.

Este proyecto pretende determinar qué características son los mejores indicadores de calidad del vino tinto y generar un modelo de predicción que permita evaluar la calidad del vino en base a estas características.

El conjunto de datos escogido es de muestras de vino del norte de Portugal, en concreto los vinos de la variedad **vinho verde**. 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.

En el repositorio de Kaggle del que extraemos los datos solo contiene un dataset. Sin embargo, en la fuente original de estos datos, vemos que hay 2 dataset: el utilizado que es de vinos tintos, y otro que contiene muestras de vinos blancos.

Dado que estos tipos de vinos siguen un proceso de elaboración distinto, no es conveniente integrarlos en un solo análisis. Los datos a analizar son todos los que incluye el dataset de vino tinto, 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.

Antes de dar comienzo a la limpieza, se modifican los nombres de los atributos para que no contengan espacios en blanco.

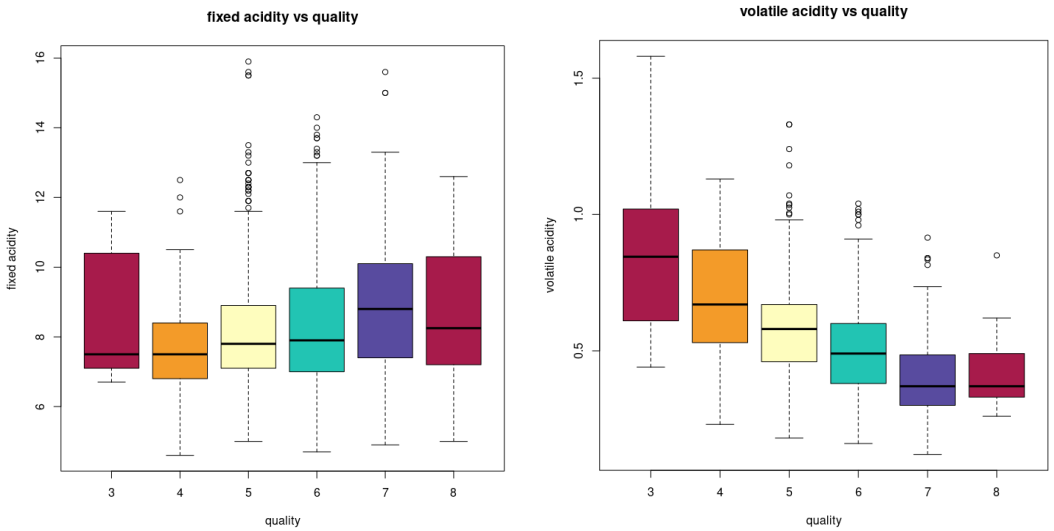
Los datos no contienen valores NA. En cuanto a valores 0, tenemos 132 valores a 0 en la variable *citric acid*. Viendo la distribución de valores 0 con respecto a la calidad del vino, consideramos que son valores válidos por lo que se mantienen en el análisis de datos.

Distribución de valores 0 de la variable citric_acid

Var1	Freq
3	3
4	10
5	57
6	54
7	8

2) Identifica y gestiona los valores extremos.

Utilizamos gráficas de tipo boxplot para identificar los outliers, como las siguientes:.



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. Es decir, si eliminamos los outliers estamos perdiendo información sobre las muestras, por lo que el análisis no tendría en cuenta vinos con atributos alejados de los valores habituales.

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?)

planificación detallada de los grupos de datos a comparar o a analizar para identificar los análisis

más adecuados a aplicar posteriormente.

EL ANÁLISIS CONSISTE EN CONSTRUIR UN MODELO DE MINERÍA DE DATOS QUE PUEDA PREDECIR LA CALIDAD DEL VINO

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.

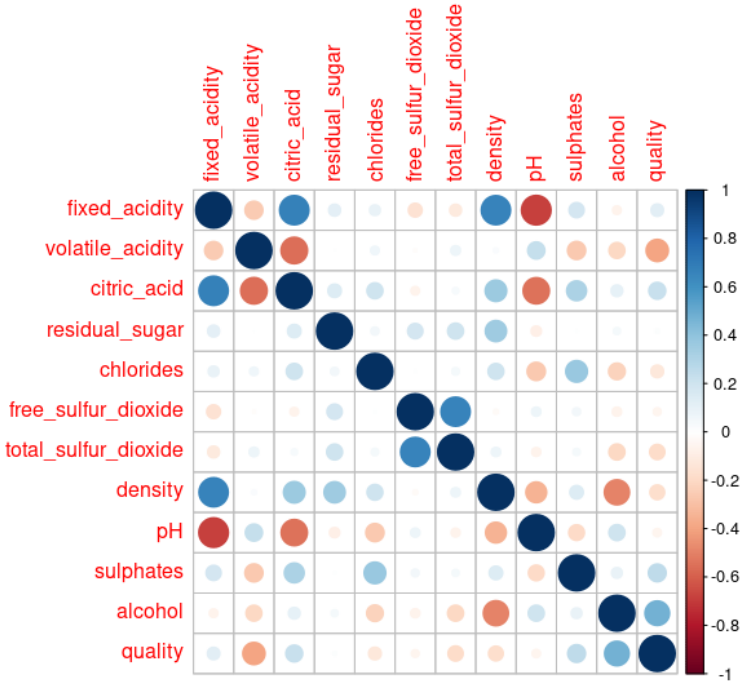
Para normalidad aplicamos test de Shapiro-Wilk, y para homocedasticidad el test de Levene.

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*

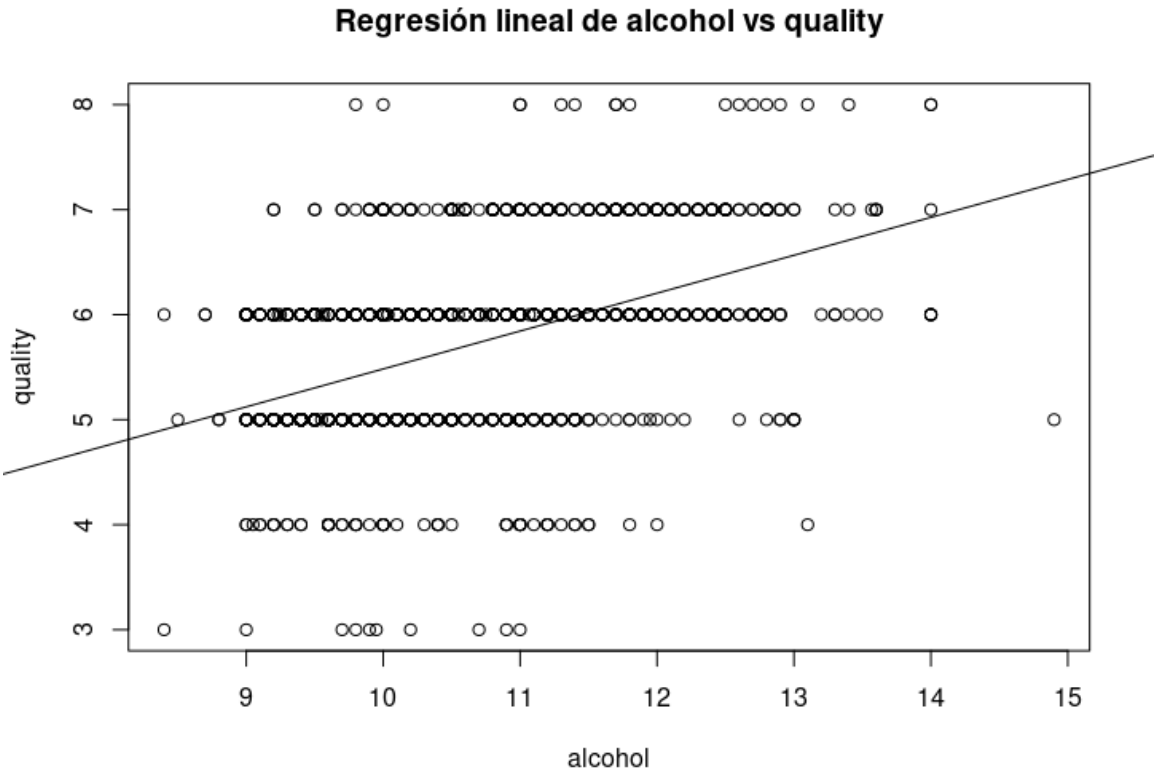
3) Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

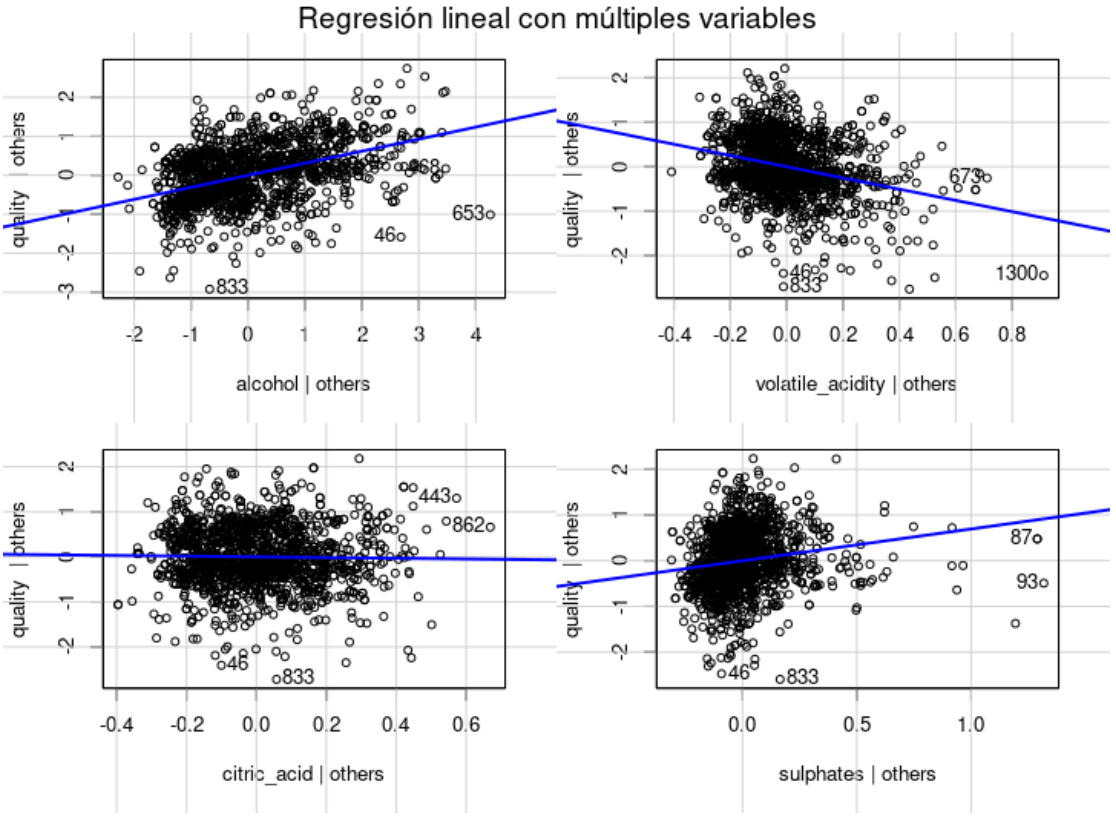
DESARROLLAR ESTO EN DETALLE

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.

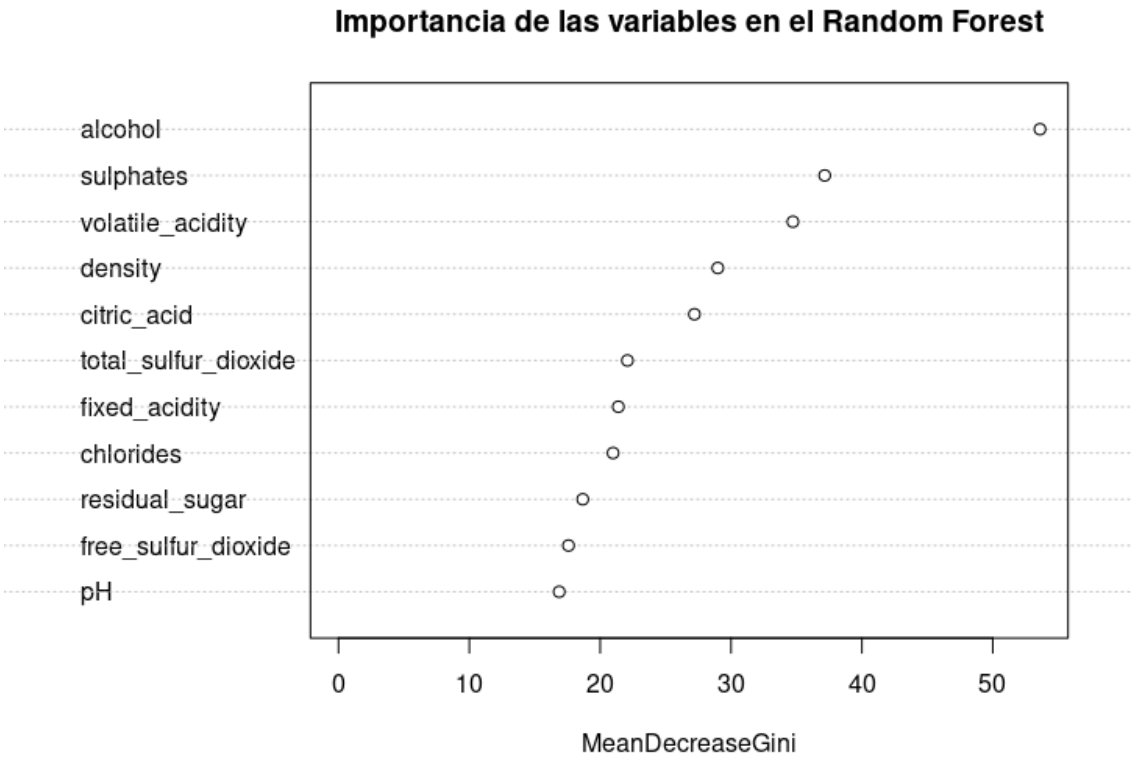


Tras saber las variables con mayor correlación, las introducimos a un modelo de regresión lineal. Primero construimos un modelo de regresión lineal simple, y luego construimos un modelo de regresión lineal múltiple añadiendo más variables.





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.**

   OSEA HAY QUE INCLUIR GRÁFICOS EN TODO EL DOCUMENTO NO? PONER ALGUNO MÁS POR AQUÍ

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

   EXPLICAR LOS RESULTADOS

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.**

   La práctica se ha realizado en código R, el enlace al archivo es el siguiente:

   https://github.com/pmm2207/Tipologia_y_ciclo_de_vida_PRA2/blob/main/code/codigo.Rmd

8. **Vídeo**

   PONER AQUÍ ENLACE AL DRIVE

9. **Contribuciones**

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