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

El juego de datos contiene 1599 observaciones y 12 variables, de las cuales todas son variables cuantitativas continuas a excepción de la calidad del vino que se describen a continuación:
1 - fixed acidity: acidez fija, hace referencia al conjunto de los ácidos naturales procedentes de la uva (tartárico, málico, cítrico y succínico) o formados en la fermentación maloláctica.

2 - volatile acidity: acidez volátil, calcula el ácido acético de un vino.

3 - citric acid: ácido cítrico, es un corrector de la acidez en mostos y vinos. Posee además

una acción estabilizante como antioxidante

4 - residual sugar: azúcar residual, es la cantidad total de azúcar que queda en el vino que no ha sido fermentada por las levaduras

5 - chlorides: cloruros. En los vinos bien constituidos, se constata siempre la presencia de una pequeña cantidad de cloruros minerales, especialmente de sodio y de potasio. Estas sales contribuyen a darle el sabor sápido a los vinos.

6 - free sulfur dioxide: dióxido de azufre libre. El dióxido de azufre se utiliza en enología principalmente como conservante, pero también para otros fines como son funciones antisépticas, antioxidantes, antioxidasicas, solubilizantes, combinadas y clarificantes. Tras incorporarla al vino reacciona con algunas de las sustancias presentes, se une con ellas y forma compuestos de adición (dióxido de azufre combinado), y sólo la parte que no combina (dióxido de azufre libre) juega un papel importante en la preservación del vino de las alteraciones oxidantes y de algunos microorganismos.

7 - total sulfur dioxide: dióxido de azufre total.

8 - density: densidad del vino, indica los sólidos totales disueltos en las uvas (azúcares, minerales, proteínas…).

9 - pH: es una unidad de medida que nos indica la acidez o alcalinidad que podemos encontrar en él.

10 - sulphates: sulfatos, son las sales derivadas del ácido sulfúrico que se utilizan como aditivo en el vino.

11 - alcohol: nos indica la concentración de alcohol en el vino

12 - quality: calidad del vino, 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 y se realiza un summary de las variables para resumirlas.

A continuación se suman todos los valores NA del juego de datos obteniendo como resultado valor 0, por lo tanto los datos no contienen valores NA.

```
sum(is.na(data_wine))
```

```
## [1] 0
```

En cuanto a valores 0, encontramos que tenemos 132 en el jueg de datos. Analizamos la distribución de estos valores 0 en las variables:

```
apply(X = data_wine[,1:12] == 0, MARGIN = 2, FUN = sum)
```

```
##        fixed_acidity     volatile_acidity          citric_acid
##                    0                    0                  132
##       residual_sugar            chlorides  free_sulfur_dioxide
##                    0                    0                    0
## total_sulfur_dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
```

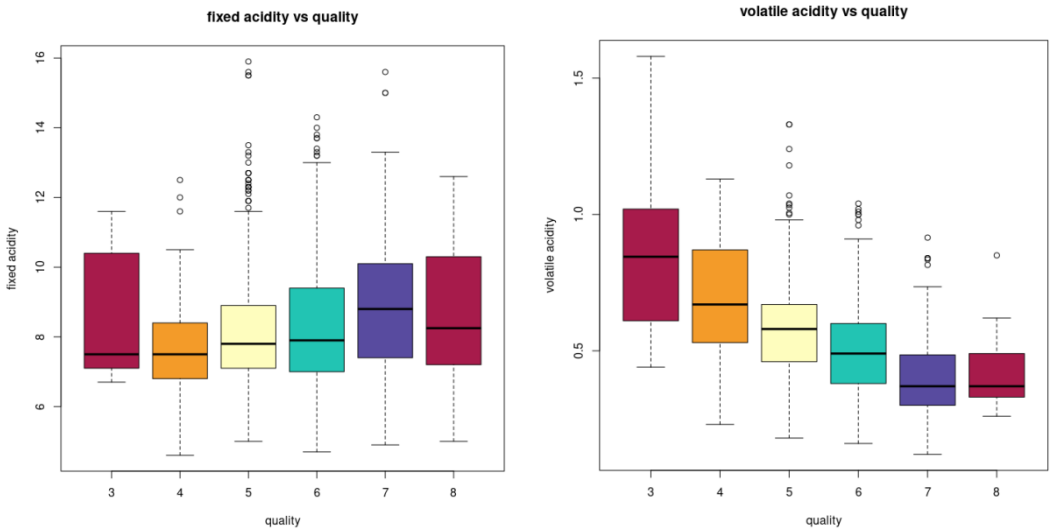
Como resultado obtenemos que los 132 valores a 0 se encuentran en la variable *citric acid*. Viendo la distribución de valores 0 con respecto a la calidad del vino, consideramos que son valores válidos por lo que se mantienen en el análisis de datos.

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 de cada variable con respecto a los resultados de calidad obtenidos, 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?)

El análisis a realizar consiste en la construcción de un modelo de minería de datos que pueda predecir la calidad del vino, es decir, un modelo predictivo.

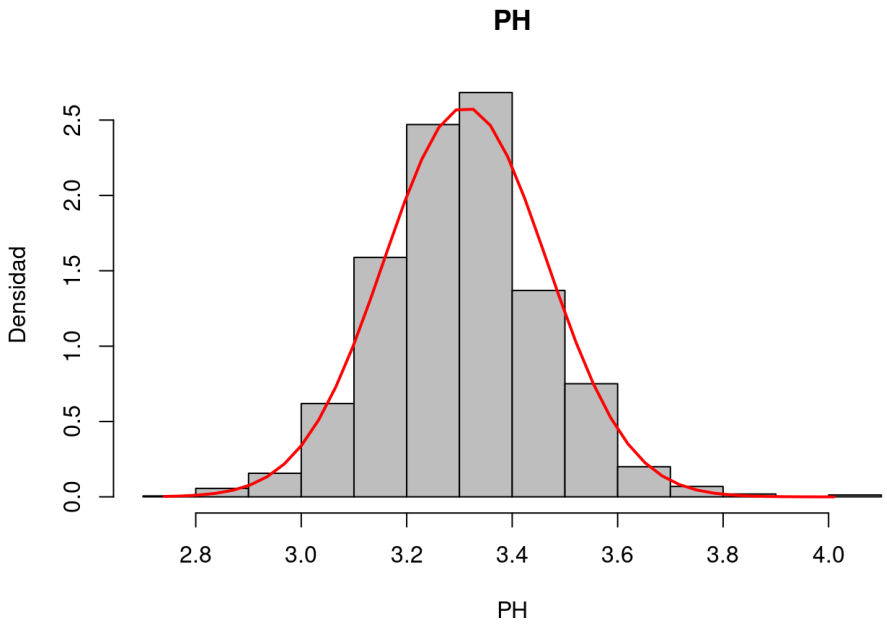
En primer lugar vamos a realizar un análisis mediante regresión lineal, en el que dejamos la variable respuesta (calidad) sin transformar. Asumimos por tanto que la variable predicha tiene valores continuos.

Tras esto, vamos a transformar la variable en binaria, de forma que podamos utilizar un modelo de clasificación binaria para predecir la calidad del vino. En esta transformación vamos a clasificar la calidad en 2 grupos:

- Mala calidad: Calidad < 7
- Buena calidad: Calidad >= 7

2) Comprobación de la normalidad y homogeneidad de la varianza.

Para comenzar representamos el histograma de cada variable, lo que permite ver la distribución de sus valores:

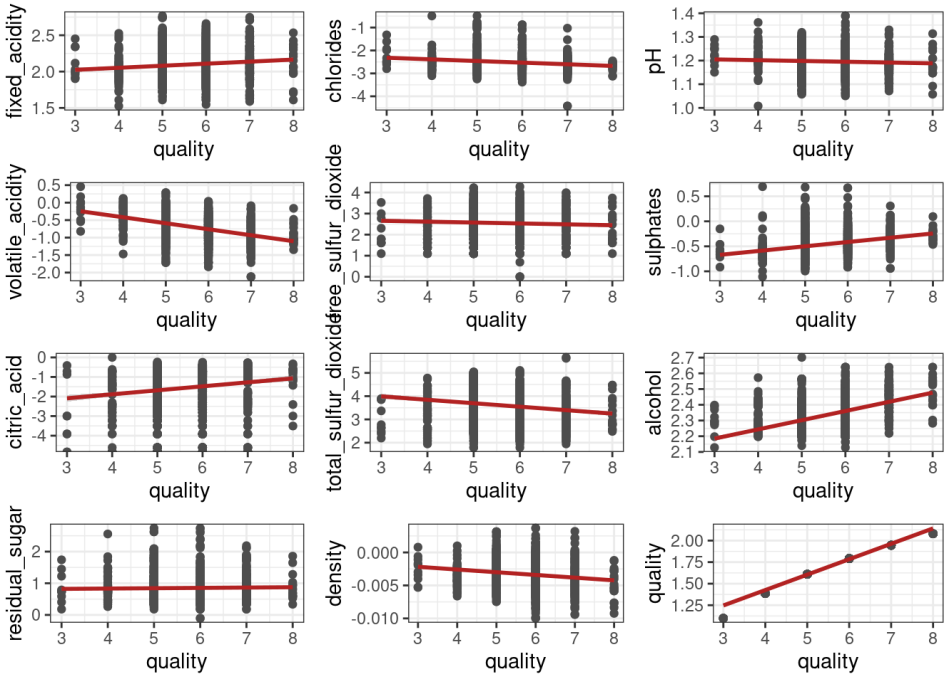


Para la comprobación de normalidad aplicamos test de Shapiro-Wilk. Debemos fijarnos en el p-valor obtenido para cada variable, como todos los p-valores son inferiores a 0.05, todas las variables siguen una distribución normal

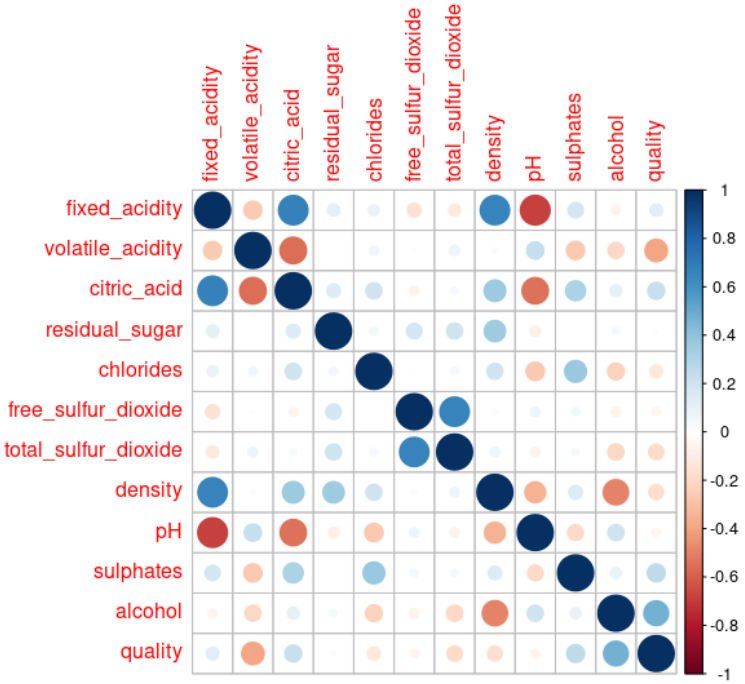
En cuanto a la homogeneidad de la varianza(homocedasticidad), utilizamos el test de Levene. De nuevo hay que fijarse en que el p-valor sea inferior a 0.05. En esta ocasión hay algunas variables que no mantienen homogeneidad de la varianza (*residual_sugar*, *chlorides*, *free_sulfur_dioxide*, *sulphates* 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.

Antes de comenzar con la regresión, realizamos un estudio de la correlación entre variables. En primer lugar se busca la correlación entre la calidad del vino y las distintas variables del juego de datos:

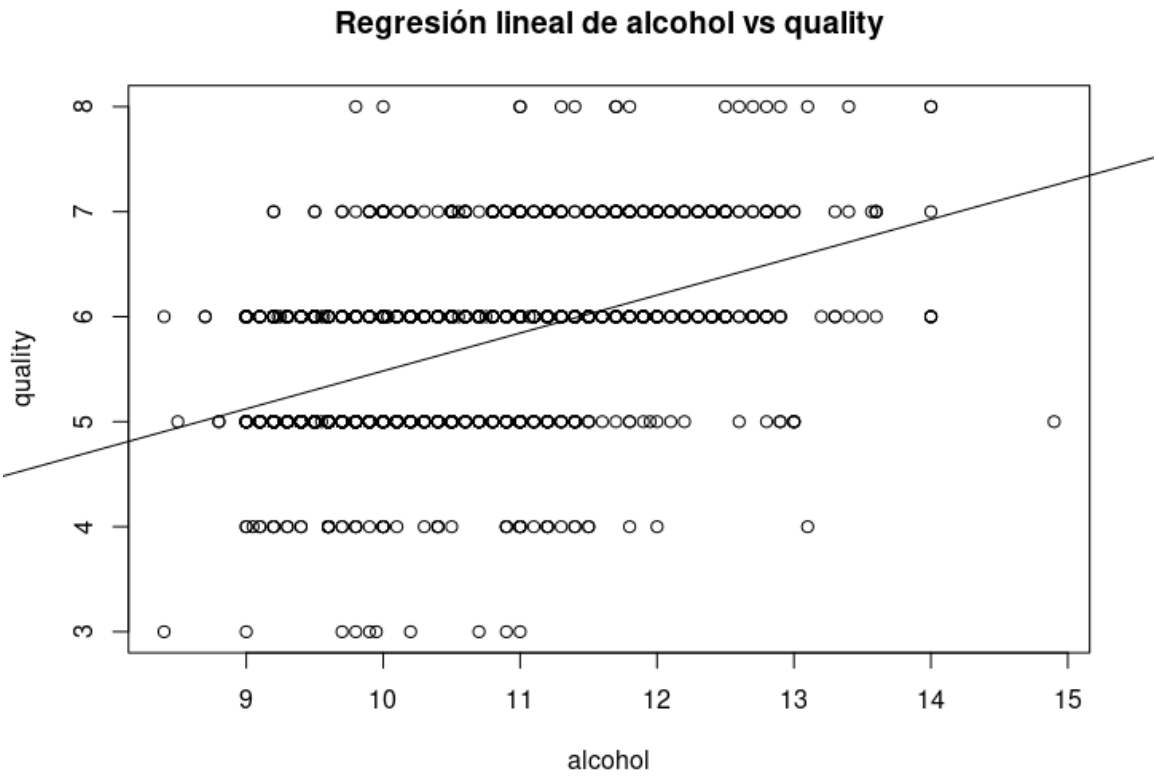


Luego, se representa una matriz de correlación entre todas las variables.



Vemos que la variable *fixed_acidity* tiene mucha correlación con otras variables, por lo que la descartamos para el modelo de regresión. En cuanto a la variable respuesta, vemos que la mayor correlación es con la variable *alcohol*, seguida por volatile_acidity , *citric_acid* y *sulphates.*

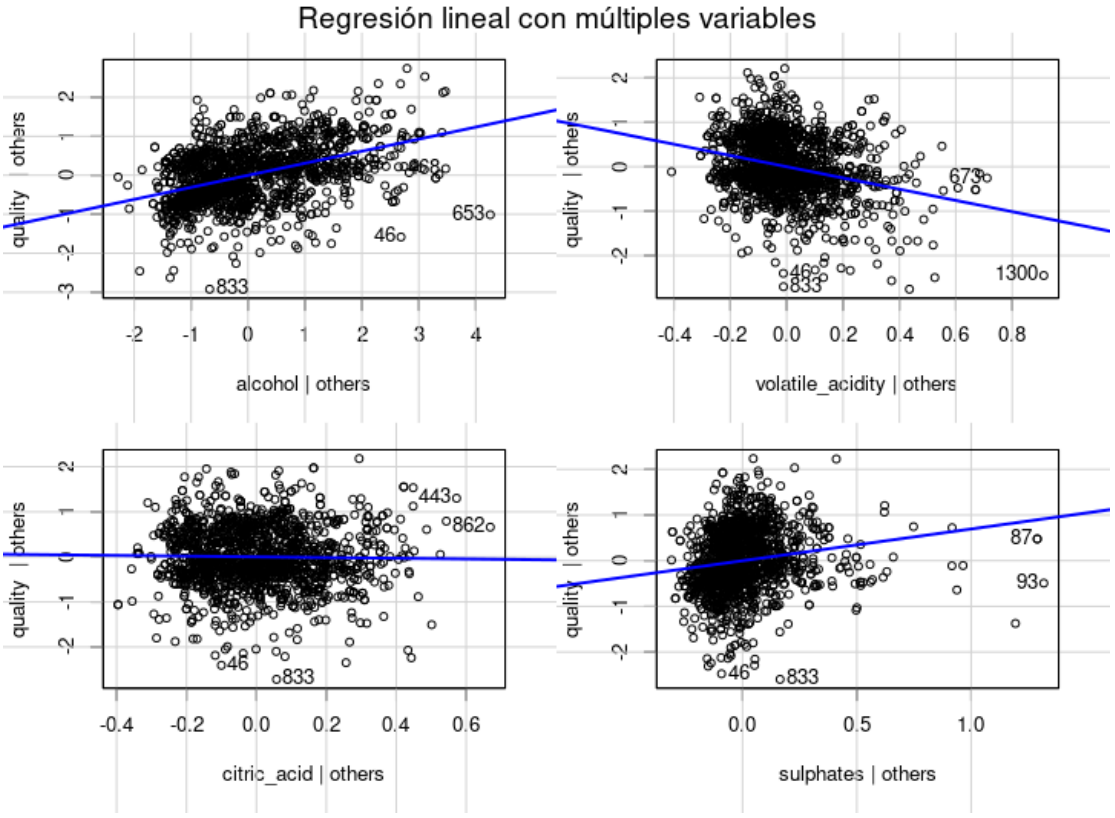
Construimos un modelo de regresión lineal simple que incluye solo la variable más correlacionada:



El resultado del modelo es una R^2 de 0.22 :

```
##
## Call:
## lm(formula = quality ~ alcohol, data = data_wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8442 -0.4112 -0.1690  0.5166  2.5888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.87497    0.17471   10.73   <2e-16 ***
## alcohol      0.36084    0.01668   21.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7104 on 1597 degrees of freedom
## Multiple R-squared:  0.2267, Adjusted R-squared:  0.2263
## F-statistic: 468.3 on 1 and 1597 DF,  p-value: < 2.2e-16
```

Tras esto, construimos un modelo de regresión lineal múltiple que incluye todas las variables encontradas con alta correlación:

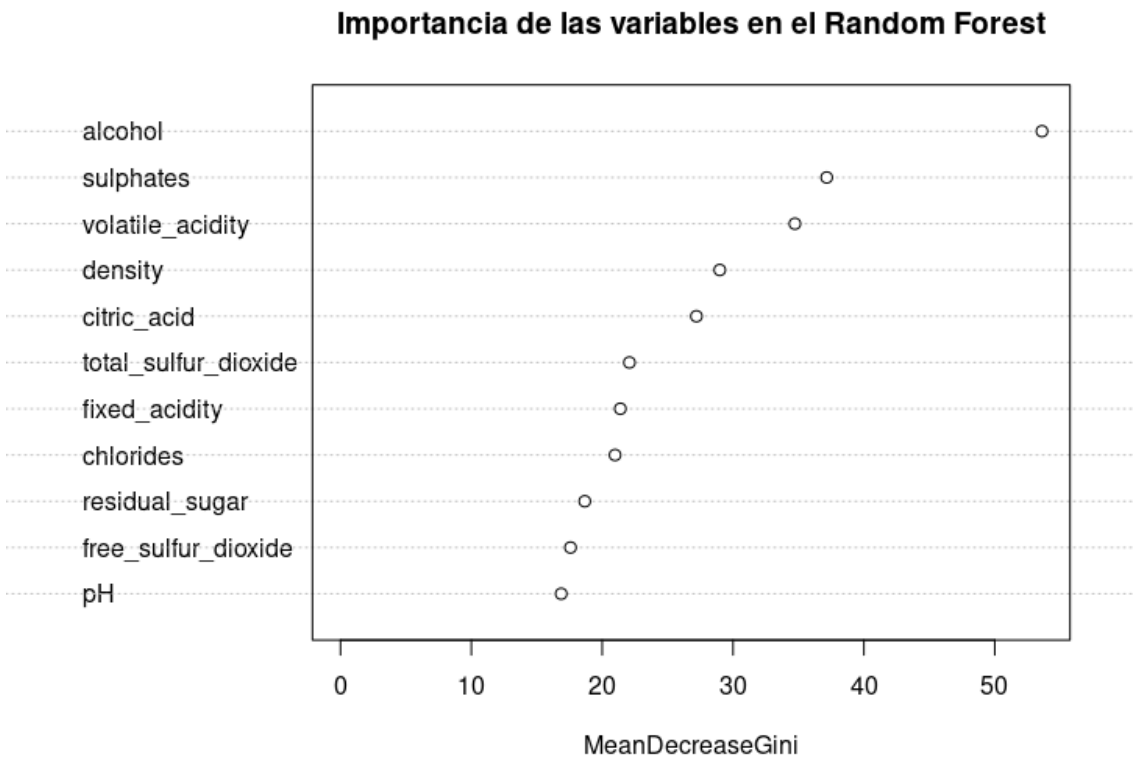


Este modelo da como resultado una R^2 de 0.33 :

```
## Call:
## lm(formula = quality ~ alcohol + volatile_acidity + citric_acid +
##     sulphates, data = data_wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.71408 -0.38590 -0.06402  0.46657  2.20393
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.64592    0.20106  13.160  < 2e-16 ***
## alcohol            0.30908    0.01581  19.553  < 2e-16 ***
## volatile_acidity -1.26506    0.11266 -11.229  < 2e-16 ***
## citric_acid      -0.07913    0.10381  -0.762    0.446
## sulphates          0.69552    0.10311   6.746 2.12e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6588 on 1594 degrees of freedom
## Multiple R-squared:  0.3361, Adjusted R-squared:  0.3345
## F-statistic: 201.8 on 4 and 1594 DF,  p-value: < 2.2e-16
```

Como los resultados de la regresión no son buenos, probamos con modelo de

clasificación, un árbol de decisión de tipo random forest. Este modelo es el que utiliza como variable respuesta la calidad de vino de tipo binario (0 mala calidad, 1 buena calidad). Veamos la importancia de las variables en el modelo resultante:



En cuanto a la calidad de este modelo, tenemos la siguiente matriz de confusión:

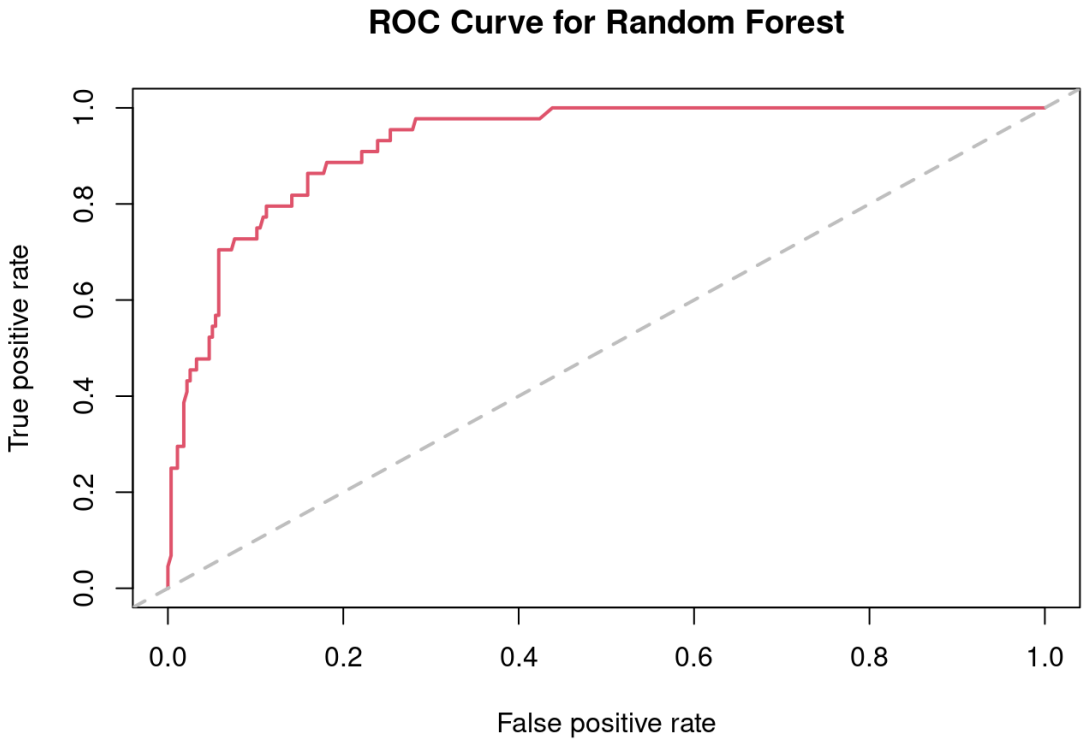
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 263  23
         1  13  21

               Accuracy : 0.8875
```

Esto quiere decir que el modelo tiene una precisión del 88%

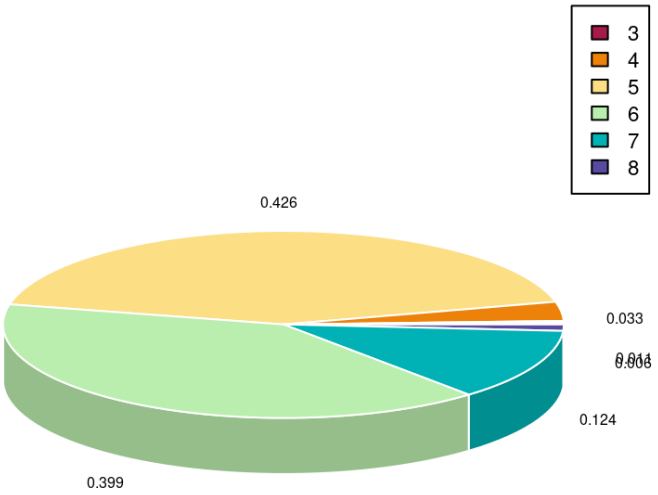
Por otro lado, realizamos un análisis ROC que nos permite evaluar la calidad del umbral de clasificación:



Para saber cómo de buena es la curva ROC, calculamos el área bajo la curva (AUROC), cuyo resultado es 0.92.

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.

Tal y como se sugiere, se han ido aportando representaciones a lo largo de los apartados. Ejemplos de otras gráficas realizadas son el gráfico tipo tarta para la variable calidad:



Esta gráfica ha sido útil a la hora de identificar la clasificación de la variable en grupos binarios.

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

Tenemos un modelo de regresión lineal simple con el que gráficamente ya apreciamos que no hace un gran ajuste. Un coeficiente de determinación R^2 de 0.22 indica que existe algo de relación entre variables, pero el ajuste está muy lejos de ser bueno.

En el modelo de regresión lineal múltiple obtenemos una mejoría en el coeficientes de determinación, pasando de 0.22 a 0.33 . Esto quiere decir que el modelo mejora al incluir variables correlacionadas. Aun con todo esto, el ajuste del modelo sigue sin ser lo bastante bueno.

En nuestro objetivo de obtener un modelo predictor para la calidad del vino, vemos que que la regresión lineal no resuelve el problema.

Por otra parte, el modelo de clasificación sí que tiene buena calidad. Vemos que la importancia de variables sigue cierta similitud a la correlación, siendo la variable *alcohol* la que más importancia tiene en el modelo.

La precisión del modelo del 88% es alta, e indica una buena calidad del modelo. Además el análisis ROC proporciona una forma de saber si el umbral con el que hemos clasificado la

variable respuesta es adecuado. Un valor de 0.92 quiere decir que el modelo discrimina de modo excepcional.

Con todo esto, podemos afirmar que el modelo de clasificación de tipo random forest es capaz de predecir de forma adecuada la calidad del vino, con lo cual se acepta como respuesta 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.

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

Se ha realizado un vídeo de presentación del trabajo. El enlace es el siguiente:

https://drive.google.com/file/d/1IA2fCC_stxUiHX3Je3BPSWTVAfGYTSP7/view?usp=sharing

9. Contribuciones

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