Práctica 2 (25% nota final)

Presentación

La presente práctica ha sido realizada individualmente por Alejandro Ortega de los Ríos.

En la siguiente tabla, se muestra la programación del trabajo previo y durante el desarrollo de la práctica:

Contribuciones	Firma
Investigación previa	AOR
Redacción de las respuestas	AOR
Exportación del proyecto a Github	AOR
Desarrollo código	AOR

Tabla 1 Programación del trabajo

Descripción del Dataset

Se pretende hacer un estudio de cómo afectan las propiedades del vino a la calidad de este (valor de 1 a 10, evaluado por expertos).

Para ello, se ha escogido uno de los datasets propuestos en el enunciado de la práctica (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009) del repositorio de Kaggle.

El dataset está compuesto por las siguientes variables:

- **fixed.acidity:** ácido tartárico (g/dm^3)
- **volatile.acidity:** ácido acético (g/dm^3).
- **citric.acidity:** ácido cítrico (g/dm^3).
- **residual.sugar:** azúcar (g/dm^3)
- **chlorides:** clorito sódico (g/dm^3).
- **free.sulphur.dioxide:** dióxido de sulfato libre (g/dm^3).
- **total.sulphur.dioxide:** dióxido de sulfato total (g/dm^3).
- **density:** densidad (g/cm^3)**.**
- **pH:** nivel de pH del vino.
- **sulphates:** sulfato potásico (g/dm^3).
- **alcohol:** porcentaje de alcohol.
- **quality:** calidad (de 0 a 10).

Todas las variables anteriores son numéricas, a excepción de quality que es entera.

El dataset consta de un total de 1599 registros y 12 variables.

Dada la motivación del estudio a realizar, se identifica la variable *quality* como variable objetivo.

Se pretende analizar la relación entre la variable objetivo con el resto de las variables, así como generar un modelo que permita predecir la calidad del vino en función de sus características.

A continuación, se muestra un diagrama de barras de la variable objetivo:

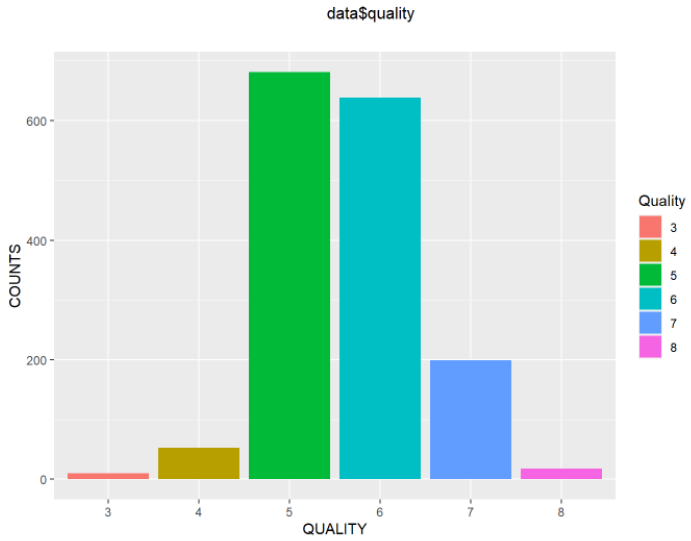

Figura 1 Diagrama de Barras, variable Quality

Como se puede apreciar, casi todos los valores se encuentran centrados en valores de calidad "medios", disminuyendo su frecuencia en los más bajos y en los más altos.

Integración y Selección de Datos

Como se dijo anteriormente, el objetivo último de la práctica es generar un modelo que prediga la calidad del vino. Para ello, será necesario evaluar qué variables están correlacionadas con *quality*, y cuáles no. Esto último, formaría parte de la fase de análisis. No obstante, dado que los coeficientes de correlación obtenidos afectarán a la selección de datos, se ha tomado la licencia de incluir esta parte del análisis dentro de esta fase.

Como se apreciará más adelante (fase de análisis), se realizarán una serie de pruebas de normalidad y homocedasticidad (homogeneidad de la varianza) de las distintas variables.

En función del resultado de la prueba de normalidad, se decidirá si emplear correlación de Pearson (variables normales y con relación lineal), o de Spearman (mide relación de monotonía).

Sin entrar demasiado en detalle (se profundizará en el análisis de datos) se obtiene la siguiente matriz de correlación:

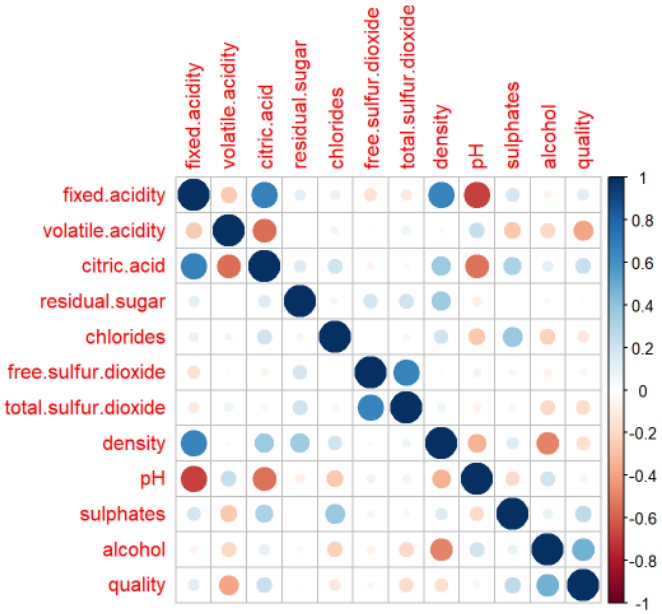

Figura 2 Matriz de Correlación

Se comprueba que se trata de un dataset en el que las variables se encuentran muy poco correlacionadas. En la siguiente tabla se muestra las variables seleccionadas junto con su coeficiente de correlación con *quality*:

Variable	Coeficiente de correlación
volatile.acidity	-0.39
citric.acidity	0.22
sulphates	0.25
alcohol	0.48

Tabla 2 Coeficiente de correlación de Spearman con variable quality

Las variables de la tabla serán las escogidas para la generación del modelo, por ser las que mayor correlación muestran (véase que aunque tengan mayores valores, no deja de ser una correlación muy baja).

Limpieza de Datos

Para la limpieza de datos se van a llevar a cabo una serie de tareas:

- Gestión de valores faltos.
- Eliminación de duplicados.
- Gestión de *outliers*.
- Transformación de datos.

No se encontraron valores faltos, pero sí uno duplicado que fue eliminado.

En las variables seleccionadas se encontraron los siguientes outliers:

```
boxplot.stats(data$volatile.acidity)$out
```

```
##  [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020 1.035
## [13] 1.025 1.115 1.020 1.020 1.580 1.180 1.040
```

```
boxplot.stats(data$citric.acid)$out
```

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

```
boxplot.stats(data$sulphates)$out
```

```
##  [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00 1.08 1.59
## [16] 1.02 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.13 1.07 1.06
## [31] 1.06 1.05 1.06 1.04 1.05 1.02 1.14 1.02 1.36 1.36 1.05 1.17 1.62 1.06 1.18
## [46] 1.07 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03 1.17 1.10 1.01
```

```
boxplot.stats(data$alcohol)$out
```

```
##  [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000 13.60000
##  [9] 13.60000 14.00000 14.00000 13.56667 13.60000
```

Figura 3 Outliers por Variables

En vista de que estos valores se encuentran dentro de un rango completamente válido (véase por ejemplo la graduación del alcohol), se ha decido mantenerlos en el dataset.

Adicionalmente, se podría transformar la variable *density* de g/cm^3 a g/dm^3**.** Sin embargo, en vista al bajo coeficiente de correlación obtenido (Fig. 2), la variable no formará parte del modelo, por lo que no será necesario.

Por último, se convierte *quality* de entero a factor para su posterior análisis.

Análisis de Datos

En la fase de análisis se va a realizar lo siguiente:

- Correlación de variables: realizado en la fase anterior para usarlo como criterio de selección de datos.
- Pruebas de normalidad y homocedasticidad: paso necesario para decidir cómo proceder en los siguientes tests y para determinar qué algoritmos son o no válidos para generar el modelo.
- Contrastes de hipótesis: comparación de medias entre dos o más grupos.
- Generación y validación del modelo: buscar un algoritmo adecuado, entrenarlo y obtener métricas de rendimiento.

Para las pruebas de normalidad se llevan a cabo tests de Shapiro-Wilk:

Variable	P-Valor
volatile.acidity	2.6e-16
citric.acidity	2.2e-16
sulphates	2.2e-16
alcohol	2.2e-16

Tabla 3 Resultados de los tests de normalidad

En vista a los resultados obtenidos, dado un p-valor < 0.05 no se puede asumir condición de normalidad.

Al no existir normalidad no se podría haber hecho uso de la correlación de Pearson, por lo que el uso de Spearman está más que justificado.

En segundo lugar, se realizan pruebas de homocedasticidad. Se comprueba si existe una diferencia significativa entre las varianzas de cada una variables anteriores por categorías de calidad. Al ser una comparación de más de dos muestras, y al no existir condición de normalidad, se emplea la prueba de Fligner-Killeen:

Variable	P-Valor
volatile.acidity	1.02e-05
citric.acidity	0.051
sulphates	0.098
alcohol	2.2e-16

Tabla 4 Resultados de los tests de homocedasticidad, Fligner-Killeen

Para las variables en las que se obtiene un p-valor < 0.05, no existe homocedasticidad. En el caso contrario sí que existe (citric.acidity y sulphates).

A continuación, se procede con los contrastes de hipótesis. En este apartado se formularán una serie de preguntas que deberán ser validadas o rechazadas en función del p-valor obtenido.

Para la formulación de hipótesis nos puede ser de ayuda la visualización de las variables a analizar:

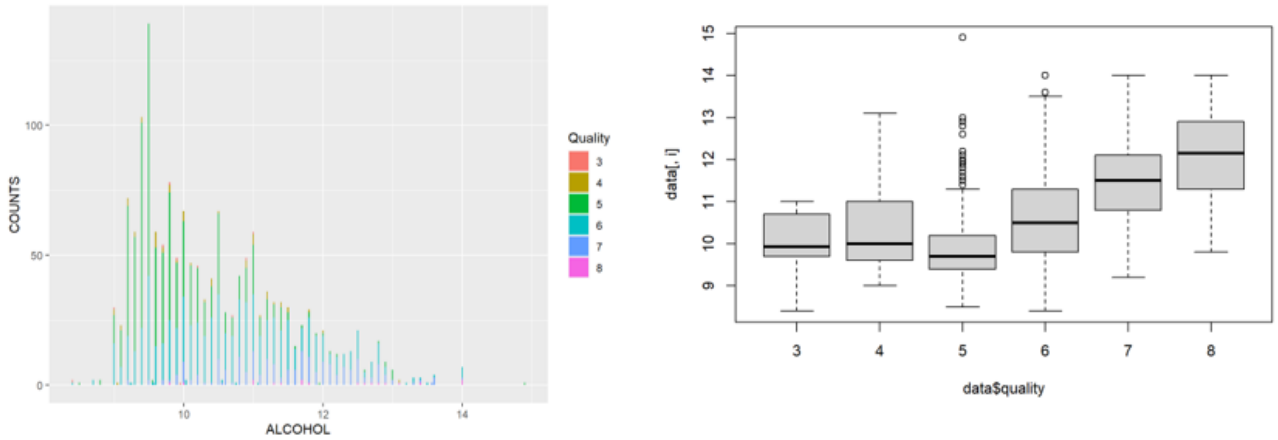

Figura 4 Visualización de la variable Alcohol

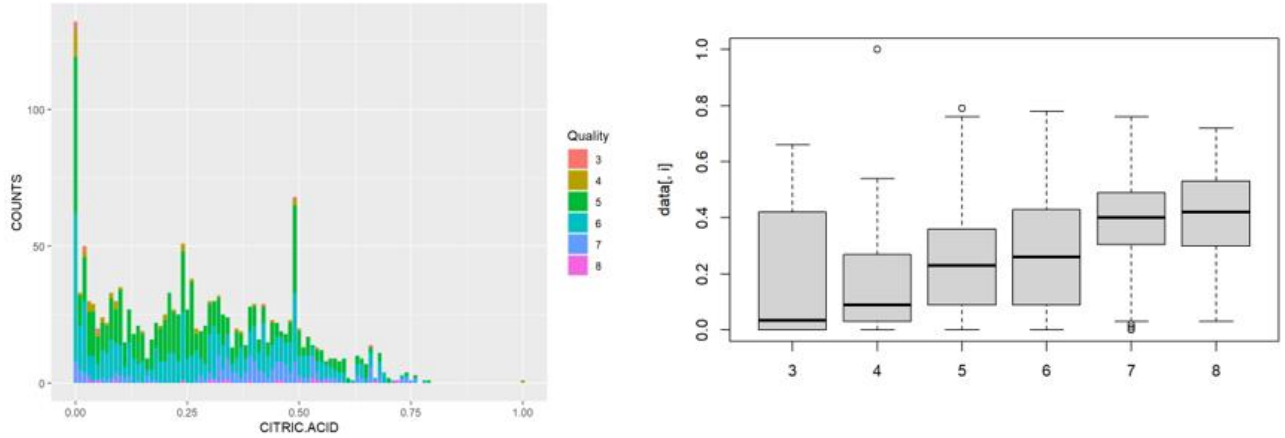

Figura 5 Visualización de la variable Citric.acid

En el diagrama de cajas y bigotes, se puede ver desplazamientos en la media a medida que aumenta la calidad del vino.

Para la comparación de medias, se efectuarán pruebas no paramétricas (t-Student no vale dado que requiere distribuciones normales y condición de homocedasticidad). Se escoge el test de Kruskal-Wallis:

Variable	P-Valor
citric.acidity	2.2e-16
alcohol	2.2e-16

Tabla 5 Comparación de muestras, Kruskal-Wallis

Si bien la prueba de t-Student es de comparación de medias, el test de Kruskal-Wallis es de comparación de distribuciones. En este caso, con p-valor < 0, se puede concluir que ambas variables muestran diferencias significativas según el nivel de calidad del vino.

Por último, se procede con la creación del modelo. Se va a escoger el algoritmo RandomForests. Se trata de un algoritmo dentro del tipo de ensambladores (aplica una serie de algoritmos débiles en conjunción y "votan" cual es la clase correcta para los datos de entrada). Lo que hace es un muestreo con reemplazo (conocido como Bootstrap), y para cada muestra entrena un árbol de decisión. Una vez entrenado el modelo, el conjunto de árboles votan la nueva predicción, siendo el voto mayoritario la predicción del bosque.

Para entrenar el modelo se divide el dataset, aleatoriamente, en un conjunto de entrenamiento y en otro de prueba. El conjunto original se divide en proporción 2/3 y 1/3 respectivamente.

Tras generar el modelo, se obtiene un OOB (Out-Of-Bag Error) de 35.68%, o lo que es lo mismo, una precisión del 64.32%.

En el siguiente apartado se mostrarán en mayor detalle los resultados obtenidos por medio de gráficas.

Representación de Resultados

Para el análisis y representación de los resultados se van a emplear las siguientes herramientas:

- Matriz de confusión
- Curva ROC

La matriz de confusión del modelo mostrada en la Figura 6, nos indica que se ha generado un modelo con una gran cantidad de falsos positivos (véase las predicciones de calidad = 6, que muchas resultan ser 5 o 7). A primera vista parece un modelo poco robusto.

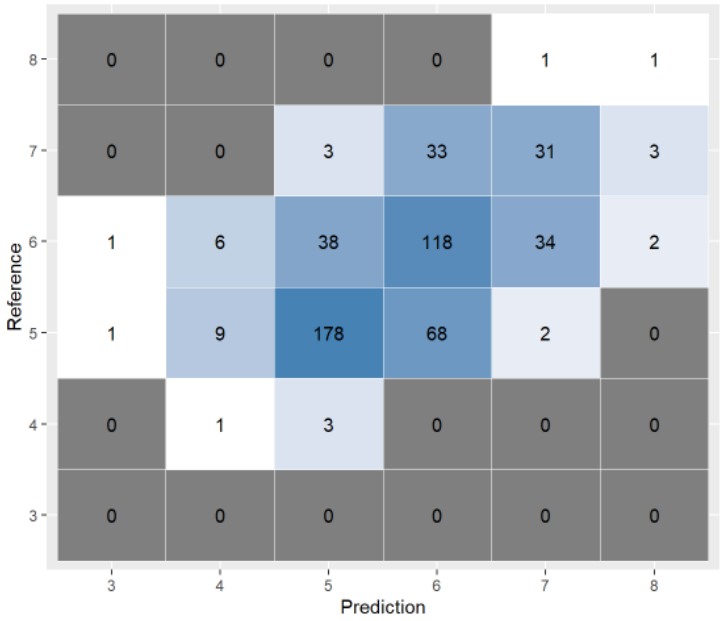

Figura 6 Matriz de Confusión

Por último, se va a emplear la curva ROC. El mejor modelo en términos de ajuste sería aquel con una curva lo más cerca de la esquina superior izquierda de la gráfica. Por otro lado, un modelo no discriminante tendría una curva con forma de diagonal.

Una forma analítica de evaluación sería a través de el área bajo la curva (AUC). De modo que:

- AUC ≤ 0.5: no discrimina.
- 0.6 ≤ AUC < 0.8: discrimina adecuadamente.
- AUC ≥ 0.8: discriminación excelente.

Por tanto, se trata de un modelo que no discrimina adecuadamente.

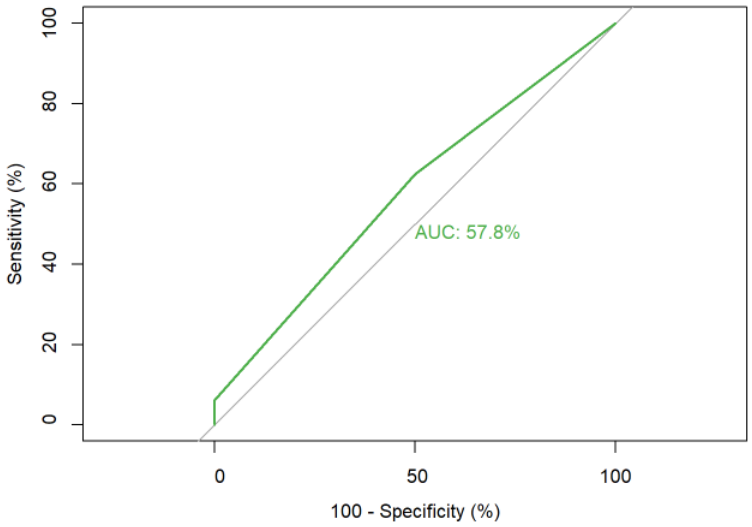

Figura 7 Curva ROC del modelo RandomForest

En vista a los resultados obtenidos tanto en OOB, Matriz de Confusión y especialmente el AUC, se concluye que el modelo generado no es satisfactorio para la predicción adecuada de la calidad del vino.

Esto último no implica que un error en la metodología o en la elección del modelo. Como se indicó en la Figura 2, las variables independientes muestran una correlación muy baja con la variable objetivo, lo que dificulta significativamente la capacidad de hacer predicciones.

La mejora del rendimiento del modelo debe ir de la mano de un dataset robusto, con unas características del vino que permitan asociarlo a un nivel de calidad determinado.

Resolución del Problema y Código

Se adjunta el código generado con *rmarkdown* además del fichero html generado. También ha sido subido al proyecto de Github.

En el proyecto de Github también se adjunta la presente memoria explicativa junto con una breve descripción de la práctica y un vídeo.

Bibliografía

Los siguientes recursos son de utilidad para la realización de la práctica:

- Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
- Guillén M., Alonso M. (2020). Modelos de Regresión Logística. Editorial UOC.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). *Data mining: concepts and techniques*. Morgan Kaufmann.