Alumnos: **Diego Martín Montoro**
Javier Hernández Hernández
PRAC2

PRACTICA 2

Nuestra práctica la hemos realizado sobre fichero winequality-red.csv extraído de página https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009. El fichero ha sido creado por Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009. (P. Cortez, 2009)

Se entrega, junto con este documento, notebook con ejecución de codigo en R. Nos referimos a éste, para dar respuesta a cada uno de los puntos de la práctica.

1. Descripcion del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?.

Se da respuesta a esta pregunta en el punto 2 del notebook "Introducción del conjunto de datos".

El dataset contiene 12 atributos de 1599 observaciones de vino tinto portugués, de la variedad "Vinho Verde". Entre los atributos, encontramos una valoración de la calidad realizada en una cata por expertos en vino. Cada experto valoraba la calidad del vino en una escala entre 0 y 10 donde 0 es muy malo y 10 muy excelente.

Atributo	**Descripción**
fixed acidity	acidez del vino
volatile acidity	La acidez volátil es una parte de la acidez total de un vino, formada por los ácidos primarios que ya están presentes en el mosto de uva (málico y tartárico) y los secundarios que son los generados durante los procesos de fermentación (acético, succínico, málico,...).
citric acid	cantidad de ácido citrico que contiene el vino
residual sugar	cantidad de azúcar que no ha sido fermentada por las levaduras que contiene el vino
chlorides	cantidad de cloruros que contiene el vino
free sulfur dioxide	corresponde al gas sulfuroso disuelto en el líquido
total sulfur dioxide	corresponde al gas sulfuroso disuelto en el líquido combinado con diversas sustancias orgánicas presentes en el mosto o en el vino.
density	densidad del líquido
pH	corresponde con acidez del vino
sulphates	sulfitos del vino
alcohol	cantidad de alcohol que contiene el vino
quality	calidad asignada al vino. Valores de 0 a 10

El fichero esta balanceado, esto es, existe igual proporción de vino buenos, normales o malos. Este es un dato importante para poder extraer conclusiones libres de sesgo.

Se pretende encontrar un modelo de clasificación supervisado que identifique qué niveles de cada uno de los atributos, predominan en cada nivel de calidad de vino. Si alcanzamos este objetivo,

podremos definir modelo que estime la calidad del vino en función de su medición en los atributos dependientes.

2. Integración y selección de los datos de interés a analizar.

El atributo output es quality. Será el atributo de referencia en el estudio de los modelos supervisados de clasificacion y regresión logístico.

Seleccionamos todos los atributos para la realización del estudio. Previamente, procedemos a su normalización, según se cita en el punto 3.3 del notebook "Normalizacion de variables".

Posterior a la normalización, pasamos a discretizar aquellos atributos que presentan mayor dependencia con la variable objetivo, según se cita en el punto 3.4 del notebook "Discretización de variables".

Hemos aplicado técnica de clusterización k-means, según se cita en el punto 3.5 del notebook "Agrupamiento de observaciones mediante clusterización". Se descarta la opción de agrupamiento ante la imposibilidad de separación clara de individuos en clusters separados.

Se opta por categorizar las variables, aplicando método de intervalos por categorías, según se cita en el punto 3.6 del notebook "categorización de variables".

Hemos realizado un análisis para poder reducir la dimensionalidad vertical, mediante la aplicación de técnica de componentes principales, según se cita en el punto 3.7 del notebook "Reducción de la dimensionalidad".

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

El estudio de los ceros o elementos vacíos (NA) se cita en el punto 3.2 del notebook "Imputación y tratamiento de valores NA".

Se asigna a cada uno de los valores NA, la mediana del atributo en cuestión, por considerar que es un estimador más robusto para el atributo, que la media. Existen otras alternativas como pueden ser:

- NN. En este se asigna a cada valor NA, el valor más común de su vecino más cercano.
- KNN. En este se asigna a cada valor NA, el valor más común de sus k vecinos más cercanos.
- Moda, se asigna el valor más común de la población.

3.2. Identificación y tratamiento de valores extremos.

El tamaño de la muestra es suficientemente grande (superior a 30). De modo que, por aplicación del teorema del límite central, podemos considerar que la muestra sigue una distribución normal. El estudio de los outliers se cita en el punto 3.1 del notebook "Estudio de outliers".

Hemos optado por realizar el tratamiento de outliers antes de la imputación de valores NA. Hemos procedido así porque de esta manera evitamos tener en cuenta valores atípicos para la imputación de valores vacíos. Hemos sustituido los outliers por NAs, de manera que luego serán imputados en el tratamiento de estos.

Se han identificado outliers en todos los atributos, entendiendo por outlier todo aquel valor que se aleje más 3 desviaciones típicas de la media.

```
##        fixed_acidity     volatile_acidity          citric_acid
##                   12                   10                    1
##       residual_sugar            chlorides  free_sulfur_dioxide
##                   30                   31                   22
## total_sulfur_dioxide              density                   pH
##                   15                   18                    8
##            sulphates              alcohol              quality
##                   27                    8                    0
```

4. Analisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Para un mejor análisis, vamos a separar el conjunto de datos en dos grupos, en función de la categoría de calidad de vino “normal” o “excelente”. Se identifican los atributos que mayor grado de correlación presentan con el output quality (“vease el punto 3.4 Discretización de variables”), que seleccionamos para la realización de los análisis.

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

El estudio de la normalidad se cita en el punto 5.1 del notebook “Estudio de la normalidad”.

Se ha aplicado el test de Shapiro-Wilk para estimar si los datos siguen una distribución normal. Se rechaza la hipótesis nula para un nivel de confianza del 95%.

Cumple la condición de homocedasticidad, luego la estimación de la varianza es igual para cada variable, indistintamente corresponda a un vino de calidad normal, o calidad excelente, según se cita en el punto 5.2 del notebook “Estudio de la homocedasticidad”.

4.3. 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.

La aplicación de pruebas estadísticas para comparar los grupos de datos se cita en el punto 5.2 del notebook “Comparación de parámetros mediante test de hipótesis”.

Hemos aplicado;

- Agrupamiento por clusterización (k-means) Punto 3.5 del notebook
- Componentes principales (PCA) Punto 3.7 del notebook
- Contraste de hipótesis sobre la diferencia de proporciones y de la homocedasticidad. Puntos 5.1 y 5.2 del notebook.
- Regresión logística. Punto 5.3 del notebook.

5. Representación de los resultados a partir de tablas y gráficas.

En el notebook, se incluyen una serie de tablas y gráficas, resultado de la labor de limpieza y análisis de los datos. Véase las secciones 3.1 del notebook “Estudio de outliers” y sección 4 “Análisis visual del conjunto de datos”. Se incorporan diferentes gráficas y tablas, sobre la estructura de los datos, y resultado de los diferentes tratamientos realizados.

Alumnos: **Diego Martín Montoro**
Javier Hernández Hernández
PRAC2

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

Finalmente, se ha diseñado un modelo de regresión logístico, en el que se relaciona la variable dependiente quality, con las cuatro variables que en el proceso de elaboración de esta trabajo, han destacado por su independencia, y contribución a la explicación del comportamiento de la calidad del vino. Estas son; Alcohol, sulphates, volatile_acidity y citric_acid. Véase la sección 5.3 del notebook "Regresión logística".

La variable citric_acid no es significativo para el modelo de regresión logístico, al aceptarse la hipótesis nula para un nivel de confianza del 95%.

El modelo definido es capaz de explicar el 88,55% de la variabilidad del modelo, lo cual es excelente. Nos permite establecer un patrón para precedir la calidad del vino en base a ciertos niveles de sus variables dependientes.

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.

El código se ha subido a Github, en el siguiente enlace.

https://github.com/vespi-github/Tipologia_y_ciclo_de_vida_Prac2.git

Contribuciones	Firma
Investigación previa	DMM, JHH
Redacción de las respuestas	DMM, JHH
Desarrollo código	DMM, JHH

8. Bibliografía

P. Cortez, A. C. (2009). Modeling wine preferences by data mining from physicochemical properties. *In Decision Support Systems, Elsevier*, 47(4):547-553.