**PRACTICA 2 – TIPOLOGIA Y CICLO DE VIDAD DE LOS DATOS**

**Eduardo Díaz Villanueva**

**Ignasi Domingo González**

**1. Descripción del dataset. ¿Por qué es importante y que pregunta o problema pretende responder?**

El dataset utilizado “*Red Wine Quality*” (<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>) contiene 12 variables que representan las características que definen a los vinos tintos, así como una variable o parámetro que recoge la calidad de dicho vino.

1 - fixed acidity  
2 - volatile acidity  
3 - citric acid  
4 - residual sugar  
5 - chlorides  
6 - free sulfur dioxide  
7 - total sulfur dioxide  
8 - density  
9 - pH  
10 - sulphates  
11 - alcohol  
12 - quality

Este parámetro es que define el interés de este ejercicio. Basándonos en los parámetros de cada vino analizaremos cuáles de ellos son más representativos para definir la calidad de un vino. De forma que encontremos modelos que puedan predecir la calidad del vino.

Si pensamos por ejemplo en una industria, podríamos reducir el tiempo y coste reduciendo el número de pruebas de calidad a las variables más significativas. Incluso mejorar la calidad del producto final, focalizando esfuerzos y recursos a reducir la variabilidad de las variables que más contribuyan a la calidad final.

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

Seleccionaremos todos los campos del dataset ya que deseamos analizar la influencia y correlación con la calidad final y a priori no podemos descartar ninguno.

**3. Limpieza de los datos.**

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

Se realiza un análisis de los datos para comprobar la existencia de datos vacíos o nulos. Solo encontramos la variable "Citric.acid" con una gran cantidad de valores 0 que consideramos como valores validos ya que en la uva este componente tiene baja presencia.

**3.2. Identificación y tratamiento de valores extremos.**

Después de realizar un estudio graficando el histograma y el boxplot de cada variable para identificar los valores extremos, observamos que el número de valores extremos es muy dispar, siendo bajo en algunas variables y relativamente alto en otras.

Como la eliminación de todos los valores extremos detectados afectaría sensiblemente a la muestra, vamos a buscar cuales de dichos valores son realmente extremos a partir de la distancia de Cook, estimando el grado de influencia de cada uno de los valores al realizar un análisis de regresión por mínimos cuadrados.

Los valores extremos detectados son eliminados de los datos.

**4. Análisis de los datos.**

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

En este apartado presentamos los datos de cada una de las varibles frente a la calidad usando un grafico de boxplot donde observamos relaciones con el resultado final que nos pueden ser útiles para aplicar modelos de predicción.

Las variables "fixed acidity", "citrix acid" , "alcohol" y "sulphates", conforme aumentan, aumenta el valor de la calidad. Por el contrario para que aumente el valor de la calidad es necesario que disminuyan "volatile acidity", "density" y "pH".

Crearemos un subconjunto de datos con estas cinco variables

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

Con la ayuda de las gráficas quantilie-quantile comprobaremos la normalidad de las variables y ejecutamos el test de de Shapiro-Wilk que fijándonos en el valor de W en ningún caso es demasiado pequeño con lo que no podemos rechazar la hipótesis nula. Siendo esta que la población de los datos está distribuida normalmente.

Una vez comprobada la normalidad de los datos, realizaremos un análisis de la varianza.

**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.**

Estudiamos la correlación de las variables usando una matrix de correlación que guardaremos en una matriz para visualizar los datos.

Con las observaciones de los apartados anteriores y la matriz de correlacion pasamos a definir modelos que nos ayuden a predecir la calidad del vino.

El primer modelo que definimos es una regresión lineal multiple.

El segundo modelo utilizaremos una regresión logística.

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

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

**Contribuciones**

Ambos hemos colaborado desde el inicio de la practica en todos los aspectos de la elaboración de la misma. Hemos estado en contacto en todo momento usando la plataforma Github como almacenamiento, así como el email, WhatApp y llamadas de teléfono para resolver dudas y acordar siguientes pasos.

|  |  |
| --- | --- |
| Contribuciones | Firma |
| Investigación previa | ED, ID |
| Redacción de las repuestas | ED, ID |
| Desarrollo código | ED, ID |