Practica 2

Integración, limpieza, validación y análisis de datos

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

   He seleccionado el dataset "Wine Quality"[1] (Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)).

   La pregunta que pretendo responder es si existe alguna relación entre los parámetros físico-químicos y el atributo sensorial de calidad. De ser así, sería muy interesante usar el modelo obtenido que represente dicha relación para "predecir" la calidad de un vino a partir de parámetros físico-químicos medibles objetivamente.

   Me hubiese gustado disponer del atributo precio en los registros, ya que así el resultado de podría ser ordenado por calidad y precio, de mucho interés para posibles compradores.

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

   El dataset es auto contenido, por lo que no requiere integración con otros. Respecto a la selección, planteo usar 4/5 de los datos para alimentar el modelaje en sí y usar 1/5 restante para validar el modelo obtenido.

3. Limpieza de los datos.
   a. Los datos contienen ceros o elementos vacíos? Cómo gestionarías cada uno de estos casos?

      Los datos no contienen elementos vacíos. Sí aparecen valores ceros que tendré que analizar para discernir si son reales o corresponden a ausencia de datos.

   b. Identificación y tratamiento de valores extremos.

      He observado que existen valores dispersos en atributos "*sulfur dioxide*" (free y total), pero debe estudiar más su significado.

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

      He planificado la aplicación de regresión múltiple para modelar la relación entre los atributos físico-químicos y la apreciación de calidad. Me planteo identificar si hay atributos correlacionados entre si y, si fuera el caso, eliminar atributos redundantes.

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

      Por definir..

   c. Aplicación de pruebas estadísticas para comparar los grupos de datos.

      Por definir…

   d. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

      Análisis de correlación para identificar si existen vínculos entre los atributos físico/químicos dados y la apreciación de calidad de los expertos. Modelaje de los datos usando regresión múltiple.

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

   Crossplots de atributos físico-químicos vs. el atributo calidad como soporte en el análisis de los datos.

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

   De conseguir correlación entre el atributo calidad y alguno o varios de los atributos físico-químicos, podríamos usar dichos atributos para "predecir" la calidad de un vino de la zona. Quedaría como futuro punto de atención estudiar si los resultados son aplicables a vinos de otras zonas donde, posiblemente, los atributos físico-químicos tengan otra huella. Adicionalmente, siento

curiosidad por saber cómo variarían los atributos aquí estudiados según la cepa del vino.

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.

   Trabajaré en Python. Esta es mi primera materia del master y no conocía antes ninguno de los 2 lenguajes. En la práctica anterior use Python a un nivel básico y quisiera profundizar mi aprendizaje con dicho lenguaje y herramientas disponibles.

8. Referencias

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties.
In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.