

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

PR2

**Autores:** Alberto Quesada León y Fernando Sevilla Martínez

**Fecha:** 11/05/2020

XX Todo lo que aparece en *gris cursiva* corresponde a comentarios o pautas fuera del índice XX

XX Las variables que aparecen subrayadas se pueden cambiar por cualquier input del modelo XX

1. Detalles de la actividad
   1. Descripción

A lo largo de esta práctica se ha realizado un caso práctico sobre el estudio del dataset Red Wine Quality con el fin de aprender a identificar los datos relevantes de un proyecto analítico, así como utilizar las técnicas apropiadas para la integración, limpieza, validación y análisis de los datos.

* 1. Competencias

A continuación, se muestran las competencias desarrolladas en esta práctica del Máster de Data Science.

* Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
* Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.
  1. Objetivos

Como se especifica en el enunciado de esta práctica, los objetivos son los siguientes.

* Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
* Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
* Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
* Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
* Actuar con los principios éticos y legales relacionados con la manipulación de datos en Tipología y ciclo de vida de los datos Práctica 2 pág 2 función del ámbito de aplicación.
* Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
* Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

1. Resolución
   1. Introducción
      1. Descripción del dataset

Como ya hemos comentado anteriormente, para el desarrollo de la práctica se ha analizado el dataset Red Wine Quality que podemos encontrar en [*UCI Machine Learning Repository*](https://archive.ics.uci.edu/ml/datasets/wine+quality).En este fichero cada registro representa un ensayo de calidad fisicoquímico sobre un determinado vino en base a una serie de atributos.

|  |  |
| --- | --- |
| Formato | CSV |
| Separador | ; |
| *Header* | True |
| Número de registros | 1599 |
| Número de atributos | 12 |

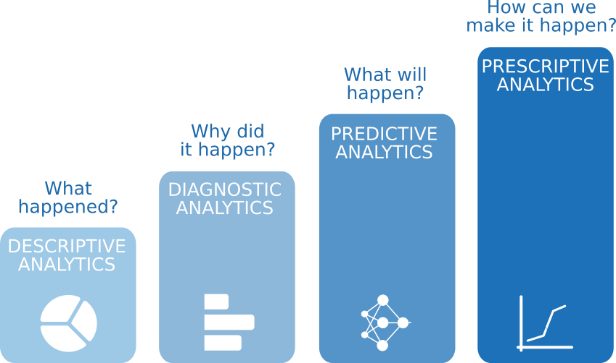
A continuación, se muestra una descripción de los atributos del fichero en estudio.

|  |  |  |
| --- | --- | --- |
| Nombre | Descripción | Tipo |
| fixed acidity | Acidez fija | Numérico |
| volatile acidity | Acidez volátil | Numérico |
| citric acid | Ácido cítrico | Numérico |
| residual sugar | Azúcar residual | Numérico |
| chlorides | Cloruros | Numérico |
| free sulfur dioxide | Dióxido de azufre libre | Numérico |
| total sulfur dioxide | Dióxido de azufre total | Numérico |
| density | Densidad | Numérico |
| pH | PH | Numérico |
| sulphates | Sulfatos | Numérico |
| alcohol | Porcentaje de alcohol en volumen | Numérico |
| quality | Valor de calidad resultante del ensayo | Numérico |

XXX Cargar directamente el dataset desde un csv añadido al repositorio. Si descargamos la última versión automáticamente en el código y se produce alguna modificación en el fichero no podríamos ejecutar el script XXX

* + 1. Importancia y objetivos del análisis

El objetivo es realizar un análisis completo del dataset de modo que trabajemos todas las fases de la analítica avanzada.



* Análisis descriptivo. Nos permite responder a la pregunta *¿Qué ha pasado?* En otras palabras, nos ofrece un resumen de los datos históricos.
* Análisis de diagnóstico. Nos permite responder a la pregunta *¿Por qué ha pasado?*En esta fase se identifican las relaciones entre las variables en estudio.
* Análisis predictivo. Nos permite responder a la pregunta *¿Qué pasará?* En esta fase se generan los modelos y se validan.
* Análisis prescriptivo. Nos permite responder a la pregunta *¿Cómo podemos hacer que pase?* Se aplica el modelo a una situación real para buscar puntos óptimos y responder a preguntas estratégicas para el diseño de nuevos productos.
  1. Análisis del dataset
     1. Análisis descriptivo
        + *Carga del fichero*
        + *Muestra de las primeras líneas del archivo original*
        + *Describe de los datos*
          - *Aquí ya podemos determinar la transferencia de información*
        + *Histogramas*
        + *Boxplots*
          - *Intentar visualizar los outliers identificados también en el hist*
        + *Eliminamos/Imputamos con KNNoutliers/missing values. Responder pregunta 3.1 del enunciado. resuelto en PEC 2*
        + *Comprobación de la normalidad y homogeneidad de la varianza. (QQplot)*
        + *Contraste de hipótesis 1: Es el valor medio de la acidez igual a X?*
        + *Normalizamos/Escalamos los datos*
     2. Análisis de diagnóstico
        + *Scatter chart de variables con recta de regresión + valor de la correlación. Esto se puede disponer en un grid de forma que parezca la matriz de correlaciones, pero en un triángulo de la matriz aparecen los scatter con la recta de regresión y en el otro triángulo de la matriz los valores de la correlación y el p-value o algo así que quede chulo.*
        + *Contraste de hipótesis 2: El valor de la calidad del vino no depende de la acidez.*
     3. Análisis predictivo
        + *Selección de atributos en función de las regresiones lineales.*
        + *Dividimos los datos en train/test*
        + *Modelo de regresión lineal múltiple con datos de training.*
        + *Estudiar colinealidad entre los atributos seleccionados para el modelo.*
        + *Diagnosis del modelo: residuos vs datos ajustados, qqplot, linear regression analysis (real vs predicho), Mean percentage error.*
     4. Análisis prescriptivo
        + *Predicción para un nuevo caso*
        + *Interpretar los coeficientes del modelo.*
        + *Directional output. Mantener uno constante y variar el resto.*

1. Archivo procesado

*Pequeña explicación fichero de salida procesado*

1. Repositorio

*Enlace GitHub*

*Screenshot Github*

1. Tabla de contribuciones
2. Recursos

* *Enlace UCI dataset*
* *Recursos UOC módulo en curso*
* *Recursos enunciado práctica*