

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
   2. Objetivos
   3. Competencias
2. Resolución
   1. Introducción
      1. Descripción del dataset
      2. Importancia y objetivos del análisis
   2. 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. 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.*
3. 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*