

## Objetivo

La estimación correcta de la demanda de agua potable representa una condición indispensable para la planificación, diseño y operación eficiente y sostenible de todos los elementos que conforman los sistemas de captación, transporte y suministro de agua potable. Esta demanda está sujeta a variaciones interanuales, estacionales, semanales, diarias e incluso horarias, muy significativas y que dependen de múltiples factores como son los ciclos de actividad económica, la meteorología, las situaciones de crisis sanitaria, los cambios en los bloques tarifarios, etc.

Dado lo anterior y partiendo de un amplio dataset con un histórico de consumos y utilizando otras bases de datos abiertos, te retamos a crear el mejor modelo de predicción de consumos en base al cual podamos realizar estimaciones a futuro en cualquiera de los municipios que gestionamos en España.



## Dataset entrada

Contiene la información sobre el consumo de agua de 2.747 contadores, ubicados en el litoral de la Comunidad Valenciana, pudiendo comprender viviendas, locales comerciales o industrias.

El consumo se proporciona con una frecuencia horaria desde el 01/02/2019 hasta el 31/01/2020.

El objetivo es predecir el consumo para cada uno de los contadores en los siguientes horizontes temporales:

- Consumo diario del 1 al 7 de febrero incluidos.
- Consumo de la primera semana de febrero (del 1 al 7 incluidos).
- Consumo de la segunda semana de febrero (del 8 al 14 incluidos).

Para ello deberás usar el dataset "Modelar\_UH2022.txt". Con este fichero deberás construir un modelo predictivo que permita estimar el consumo de agua.

### Variables

- *ID*: Identificador del Contador que registra la medida de lectura.
- *SAMPLETIME*: Fecha y hora del consumo en formato UTC. Momento en el que se produce el mensaje o el contador ha emitido el registro.
- *READINGINTEGER*: Medida registrada por el contador en litros. Parte entera.
- *READINGTHOUSANDTH*: Medida registrada por el contador en litros. Parte decimal.
- *DELTAINTEGER*: Consumo calculado en litros a partir de la medida registrada por el contador. Parte entera.
- *DELTATHOUSANDTH*: Consumo calculado en litros a partir de la medida registrada por el contador. Parte decimal.

Los ID están ordenados de forma ascendente pero no son correlativos

### Formato y estructura

Este dataset tiene extensión txt con la siguiente estructura y formato:

- *Nombres de variables*: incluidos en la cabecera
- *Separador*: "|"
- *Codificación*: UTF-8

Sin nombre de fila.

## Dataset respuesta

Es el fichero solicitado con tus predicciones de consumo. Se denominará "Equipo\_UH2022.txt" donde 'Equipo' será el nombre del equipo con el que te has inscrito.

Sin cabecera ni nombres de filas.

Constará de 2.747 filas con 10 columnas cada fila:

- *ID*: ordenado de forma ascendente
- *Dia\_1*: Predicción para el día 01/02/2020
- *Dia\_2*: Predicción para el día 02/02/2020
- *Dia\_3*: Predicción para el día 03/02/2020
- *Dia\_4*: Predicción para el día 04/02/2020
- *Dia\_5*: Predicción para el día 05/02/2020
- *Dia\_6*: Predicción para el día 06/02/2020
- *Dia\_7*: Predicción para el día 07/02/2020
- *Semana\_1*: Predicción para la semana del 01/02 al 07/02/2020, ambos inclusive
- *Semana\_2*: Predicción para la semana del 08/02 al 14/02/2020, ambos inclusive

Separando campos con "|", el valor de la predicción en litros, y los decimales con ".".

## Se valorará

La calidad y la técnica utilizada para generar un modelo.

Se analizará la técnica analítica utilizada y se compararán objetivamente los valores reales frente a los valores predichos por el modelo. Para ello, se tendrá que minimizar las desviaciones con respecto a los datos reales.

Se calculará el "error cuadrático medio" o RMSE, definido como:

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

Siendo:

"n" el número de casos,

" $\hat{y}$ " el valor estimado,

"y" el valor real

Se aplicará al subconjunto de las predicciones del mismo horizonte temporal, calculando una media para los 7 RMSE diarios y otra media para los 2 RMSE semanales, obteniéndose la métrica final del siguiente modo:

Métrica = 50% RMSE (media diaria) + 50%RMSE (media semanal).

## ¿Qué pedimos?

Además del “dataset respuesta”, te pedimos:

1. Un script (“script exploración”) que contendrá el análisis exploratorio y procesos relevantes testados o ejecutados pero no aplicados en la solución final.
2. Un script (“script predicción”) que contendrá el proceso de extracción, transformación y carga de los datos, el procesamiento aplicado así como la generación de predicciones.
3. Una breve descripción donde se expondrá el proceso y la metodología seguida, las técnicas aplicadas y los resultados obtenidos (en formato presentación, pdf o html, máximo 5 páginas con 3 imágenes).

Un valor menor no conllevará explícitamente una mejor clasificación. El “script de predicción” mencionado debe cumplir que sea generalizable y en el caso de métricas equiparables, se tendrán en cuenta los criterios siguientes:

- el Jurado podrá valorar si la documentación interna aportada (código y comentarios) está correctamente estructurada, expresada y es reproducible.
- los scripts de exploración y predicción deben constituir un proyecto de data science con todas sus fases.