Challenge

The proposed challenge involves developing a predictive model to estimate water consumption **by sector** in the city of **Villarubia** for the **ensuing 7-day period.** The output file should adhere to the same format as 'caudales.csv'. Please be mindful of the frequency of measurements/predictions.

Evaluation criteria

The evaluation criteria for this competition encompass both technical and business aspects, each contributing **50% to the final grades**.

- **Technical criteria** (**50% weight on the final grades**) are meant to assess the general ability of the teams to handle, interpret, and understand the data, build predictive models for the given problem and evaluate their performance according to an objective metric.

- **Business criteria** (**50% weight of the final grades**) are more general in scope, and they involve the ability to devise what value can be extracted from this data, formulate relevant business questions and try to find answers to them in the data. The quality of the presentation and the ability to communicate clear and powerful ideas on the final pitch will also be part of the business criteria.

Teams are required to submit:

1. **Technical materials**: (a) The code they developed to tackle the problem (ideally with comments, or jupyter notebooks alternating code and text explanations). (b) The output file. (c) The model performance metrics.
2. **Business materials**: An executive summary (3-5 pages) with their inspection strategy and value propositions.

The technical criteria will be assessed on the basis of the **first submission**, and the performance metrics obtained by the best model **submitted by Google Forms.** Business criteria will be assessed mostly from the executive summary.

Technical criteria (50% weight on the final grades)

The models submitted will be evaluated based on three chosen metrics: ***MAPE, RMSE and MAE***. ***The superior predictive model will exhibit the smallest values across all the three metrics.***

In addition to this, the code should be clearly structured and results should be interlaced with explanations in jupyter notebooks. The notebooks should be

clearly written, and explain the process followed starting from the raw dataset, cleaning and preprocessing, exploratory data analysis, model formulation, hyperparameter tuning (if needed), final metrics and discussion.

When developing your models, please pay special attention to:

- Overall understanding of the problem
- Exploratory data analysis
- Feature engineering
- Predictive models performance

Overall understanding
Ensure that you understand the meaning of each predictor variable in the different datasets: what it means, in which units is it expressed, how is this data registered, at which moment in the time or day and location, could it contain errors? could it contain outliers ? can we trust the data ? Using common sense, will a given predictor variable be useful to predict our target?

Exploratory Data Analysis
Getting acquainted with the datasets is a first necessary step before any modelling on the data takes place. Explore the data distribution, which variables are categorical and which are numerical, do we really understand the meaning of each variable? Are there any correlations among the variables ? Are there predictor variables with missing values or outliers ? Can we trust the values of the data ? Try to formulate hypotheses and understand your datasets before further exploration is conducted. Create good visualizations that help develop your intuition and understand the patterns. If necessary, decide how to handle missing values by either data imputation or removing rows/columns from the dataset.

Feature engineering
Which features will you use in your predictive model ? Is it legitimate to use all the provided data? Can you imagine how the model will be used in production?
Can you enrich your dataset with external information ? At the very least you will need to merge several datasets and/or derive features to train models. Be creative: anything that you can build on the given data that might have a more direct connection to what you are trying to predict will improve your models performance.

Model performance

Performance metrics and the output file containing predictions should be submitted via **Google Forms**. The proposed challenge involves developing a predictive model to estimate water consumption by sector in the city of Villarubia for the ensuing 7-day period. The output file should adhere to the

same format as 'caudales.csv'. Please be mindful of the frequency of measurements/predictions.

Remember: The models submitted will be evaluated based on three chosen metrics: ***MAPE, RMSE and MAE***. ***The superior predictive model will exhibit the smallest values across all the three metrics.***

Feel free to try different families of models, adjust their parameters, add regularization, go back to your preprocessing cycle and continue iterating, etc. You will use your training data to gauge the performance of your model and you will receive immediate feedback on the test set when you submit your predictions.

Business criteria (50% weight on the final grades)

Extracting value from data

For this task, assume you have a **prediction model that effectively forecasts water consumption for the upcoming 7 days.** Your task is to determine how to apply this model in a business case and how it can bring value to the company.

Knowing the expected water consumption in advance allows the supplying company to ensure the provision of service, as well as to be able to predict anomalous events in water facilities, such as:

1. Prediction of pipe breakages and water leaks.
2. Detection of anomalous situations such as fraud.
3. Anticipation of the network's special needs, for instance, inspections for malfunction prevention.

In this way, the company can improve and assure people access to a resource as vital as water.

You must design a **prevention plan** to minimize water loss, mitigate environmental impacts and avoid shortages for families due to unforeseen breakdowns. You must include clear and realistic strategies to detect and solve possible leaks quickly. Furthermore, your proposals must evaluate the environmental and social impact they provide and present these solutions clearly and understandably, highlighting their viability and potential benefits.
The evaluation will focus on the coherence of your proposal, the clarity and realism of the prevention plan, the calculated impact, and the presentation's clarity. This challenge involves finding technical solutions and considering their broad implications, addressing issues of sustainability and equity in access to water.
Special consideration will be given to proposals that, in addition to solving technical problems, contribute to improving the resilience of communities and

comprehensively address the environmental and social challenges linked to water use.

All participating teams will be assessed on a combination of these two factors **50% technical criteria and 50% business criteria**. A team with a model with poor performance might still qualify for the final if their business model and strategy is outstanding, so try to devote some time to both tasks. And remember to work in parallel and divide your team according to expertise and capacity.

Pitch clarity

The teams that are selected for the final phase will pitch their results in front of a jury. Their technical results will have already been assessed by the technical jury on the basis of their submitted notebooks and model performance. On the pitch you will need to transmit clear and powerful ideas that highlight your results and show your understanding of the problem, your ability to harness value from the data and your ideas to contribute to the problem under consideration. Focus on the large scale goals, while showing evidence that your technical skills are solid, but do not use your time to explain straightforward technical solutions, unless you think that they are really essential.