Data Science Challenge

1. The Challenge

Demonstrate your data science and engineering skills:
Conduct an exploratory data analysis of the attached dataset and build a machine learning pipeline to train and test ML models for the target variable **soc_stock_t_ha**.

2. Requirements

1. Exploratory data analysis: show covariation/correlations between the different variables of the dataset.
2. Plot the distribution of the soc_stock_t_ha variable and determine the covariates with the highest correlation with this variable.
3. Build a modelling pipeline that let's a user select a subset of covariates to serve as explanatory variables and an algorithm from a preselection of ML models to model soc_stock_t_ha as the target variable.
4. The pipeline should split the dataset into a training and test dataset, train a model, use it to predict on the test data and measure and showcase the accuracy of the results.
5. The modelling pipeline can be contained in a jupyter notebook or run through a script in a docker container.
6. Present your results as a .pdf or .html report

3. Expected time-frame

Time-box your work on this challenge to 3 hours max. We do not expect a perfect solution, but are keen on learning more about your way of working.

4. Deliverables

Send us your challenge submission as a link to a gitlab/github repository or a zip file containing all the code as scripts/notebooks and the report as html or pdf file.

Looking forward to your submission, we thank you and wish you the best of success!