Schneider Datathon

Our team

Marçal Comajoan Cara: mcomajoancara@gmail.com
Pol Puigdemont Plana: polpuigdemont@gmail.com
Jacinto Suñer Soler: jacisuni@gmail.com

Context

The EU contributes 18% of total global warming gas emissions; However, it is increasingly determined to take the lead in the fight against climate change. That is why it has set itself the goal of reaching zero carbon emissions by 2050. To this end, it has put in place a wealth of resources to help achieve this goal over the next few years, and it will need our help to do so. In this context, Scheider Electric has brought us the following challenge.

The challenge

Our goal is to predict the type of pollutant in a factory given a dataset containing information about it.

What have we worked on

All the detailed analysis and justifications of the steps we've taken can be found in the .ipynb file. Please consider it as part of this presentation as well. We thought it made more sense explaining what we did and how we did it following along the code and plots.

We have gone through a full data analysis process starting at data gathering (from all sources including csv, json and pdf), followed by exploration and a meditated preprocessing which has allowed us to train diverse machine learning models aimed for classification.

After testing around with the models hyper-parameters we've selected a final model (a random forest model with a 1000 estimators) from which we obtain predictions on the pollution type for the unlabelled test data which was provided to us.

How we built it

We've built it fully on python using typical tools in ML tasks.

Learning opportunity

We are second and third year students at the Data Science and Engineering Degree at UPC BarcelonaTech which belongs to the FIB, ETSETB and FME. Although this wasn't neither our first hackathon nor our first time working with Machine Learning tools, facing a real problem in this new environment and mission as well as the 24h time constraint has been a great learning experience. We've had to overcome many obstacles, sorting them out by discussing possible solutions and choosing the one we felt was the best. This situations have put us in a great position for learning and growing as Data Scientists.

Repo contents

This repo contains a python notebook with the full detailed process we've gone through plus the predictions for the test data.

Thanks

Developed with :) at the Schneider Electric hackathon organized by NUWE. Thank you to all the people who have made it possible and has offered us guidance during the process.