Schneider Electric Hackathon

We import the csv files and concatenate them in a single dataframe.

Then we call the API to get the json files and append them in the dataframe.

Next step is importing the pdf files. Each file is a single row of our dataframe. For each pdf,we extract the text from the pdf first and then we look for every column value,and store it in a list in the same order as the columns of our current dataset.

Now we already have the whole dataset.

We store the target variable “pollutant” apart from our dataset, and we encode it with 0,1 and 2.
We delete all spaces from all variables, because in order to import the data from pdf we deleted all spaces from those rows, so now we delete them also for the rest of rows.
Then we delete a group of variables which we thought were irrelevant or repetitive. Once we have only the variables that we want, we turn our non numerical variables into dummies, for the random forest function of python to be able to use the data.

Then we fit a random forest. Once we have our model, we import the test data and apply to it the same transformations as for the train data.

We then predict with our model and obtain results.

(Due to the dummy variables as input, the number of columns of the train data and test data is not the same(for example if Spain was a country in the train data but there was no observation of Spain in the test data, a column from train will not be in test). Since we had no more time to solve this we simply took the columns from train and put them as the columns in test (deleting and crating columns). This is obviously not the best approach, but just a fast solution to obtain predictions.)