Schneider Electric Hackathon 2022

By: Carlos Moreno Tavira

May 2022

1. Data extraction from .csv, .json, .pdf

- I have extracted the data using **pd.read_csv**, **pd.read_json** and **PyPDF2** for the .zip containing the pdfs.

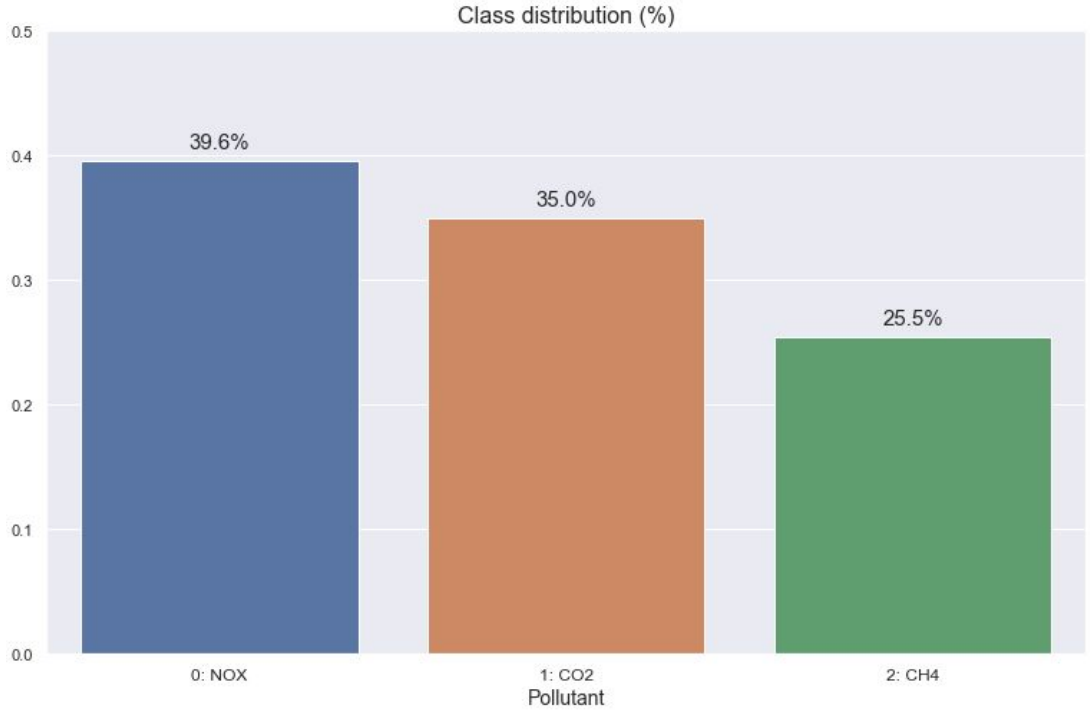
- I have cleaned the dataframes and concatenated them into one big dataframe of **65709 x 21 columns.**

```
df = pd.concat([df1, df2, df3, df4, df5, df6], ignore_index = True)
```

- Then I checked the distribution of the variables and dropped the columns that are not useful (e.g. CONTINENT, targetRelease, some ID's...)

2. Exploratory Data Analysis

- Check the distribution of the data. The dataset is quite balanced.
- Check correlation between variables. Prepare the data for model fitting.



3. Predictive model

- I have compared the performance of **Random Forest, XGBoost, LightGBM and CatBoost classifiers** with default parameters. The best F1 macro score was obtained by using **Random Forest (71.92%).**
- The results can be improved by **hyperparameter tuning with cross validation**, but I had no time left… (note: in-depth explanation and justification of my solution can be read in 'main.ipynb')