PREDICT POLLUTANTS FROM INDUSTRIAL FACILITIES ACROSS EUROPE

Rubén Cuervo, Elliot González and Eduard Ruiz (Table 30)
Schneider Electrics Hackathon 2022 – Data Science
21st May 2022

PROCEDURE

DATA IMPORTATION

1. **Importation of the JSON datasets** with *urlopen* function from *urllib*.
2. **Importation of the CSV datasets** with *read_csv* function from *pandas and adding additional codes and missing*.
3. **Importation of the PDF files** with PyPDF4 library and adapting titles and formatting values.

PREPROCESSING TECHNIQUES

4. **Normalize** values from PDF dataframe (are in a different scale).
5. **Merge all datasets** to obtain a unique dataset.
6. **Convert categorical features** into labels.
7. **Normalize data** by extracting the minimum and dividing by the whole range of that feature.
8. **Split data** into train and test (10% for test).

MODEL SELECTION

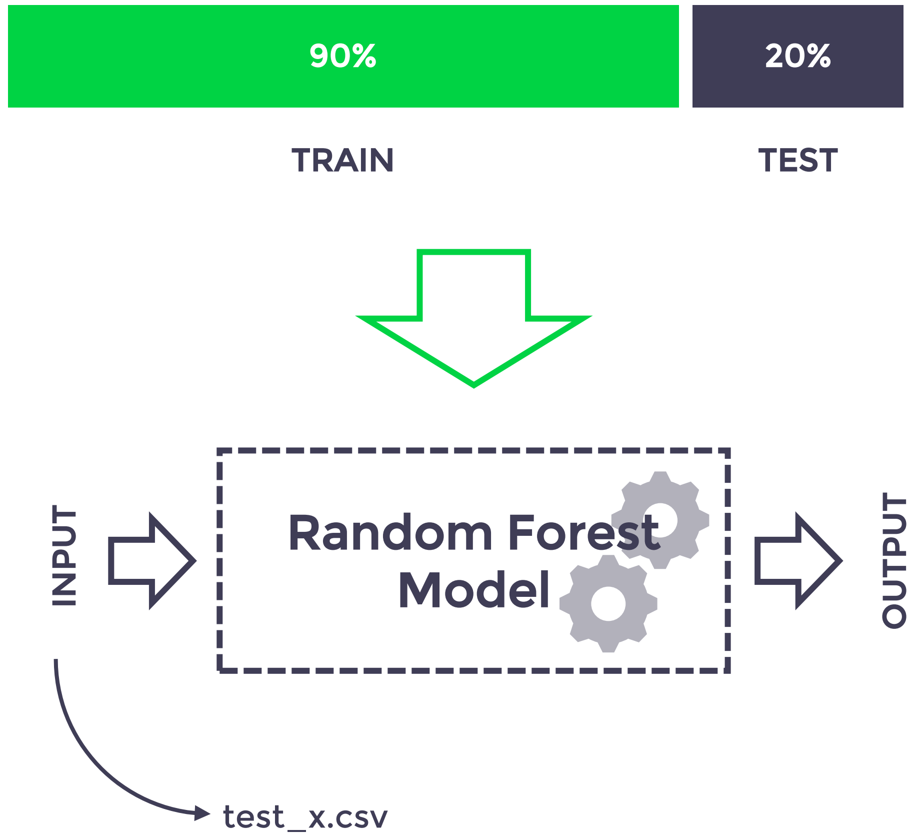
9. **Selecting features** not required performing.
10. **Try different models** to see which is the most optimal (XGBoost, KNN classifier, Random Forest, Deep Neural Networks…).
11. **Adjust hyperparameters** to obtain the best performance.
12. **Evaluate** the model.
13. **Return to point 9** until the best value is achieved.

SELECTION FEATURES

Feature	Procedure	Reason
countryName	Codify	Convert categorical.
EPRETRSectorCode	Codify	Convert categorical.
eptrSectorName	Remove	Code is enough.
EPRTRAnnexIMain ActivityCode	Codify	Convert categorical.
EPRTRAnnexIMain ActivityLabel	Remove	Code is enough.
FacilityInspireID	Remove	Too much unique labels.
facilityName	Remove	Identifier is enough.
City	Remove	City ID is enough.
CITY ID	Remove	Too much unique labels.
targetRelease	Remove	All same value ('AIR').
pollutant	Codify	Convert into categorical int.
DAY	Remove	Improve performance.

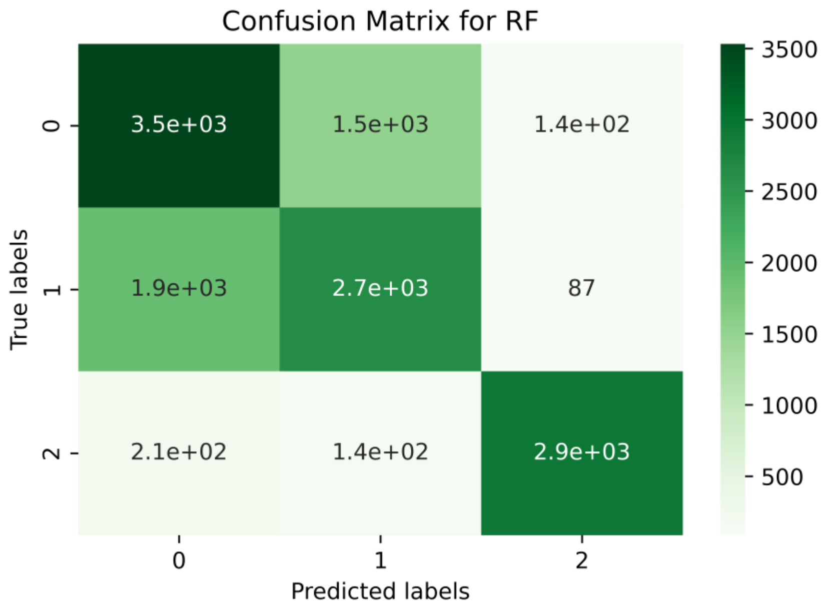
Feature	Procedure	Reason
MONTH	Remove	Improve performance.
reportingYear	Remove	Improve performance.
CONTINENT	Remove	All same value ('EUROPE').
max_wind_speed	Remove	Improve performance.
avg_wind_speed	Normalize	Improve performance.
min_wind_speed	Remove	Improve performance.
max_temp	Remove	Improve performance
avg_temp	Normalize	Improve performance.
min_temp	Remove	Improve performance.
DAYS WITH FOG	Normalize	Improve performance.
REPORTER NAME	Remove	Many unique values.
rng_temp	Add	Combine min and max.
rng_wind_speed	Add	Combine min and max.

RESULTS OBTAINED



0.711

F-1 SCORE
(MACRO)


Confusion Matrix for RF

CONFUSION MATRIX

CONCLUSIONS

- Deep Learning is not always the best method, sometimes more simple ML models provide better performance.
- Not all features are relevant when training a prediction model, it is essential to discriminate between significant and noisy features.
- Even messy data in pdf format can be converted to workable dataframes.
- Schneider Electrics looks for sustainability and this can have an impact on the environment.