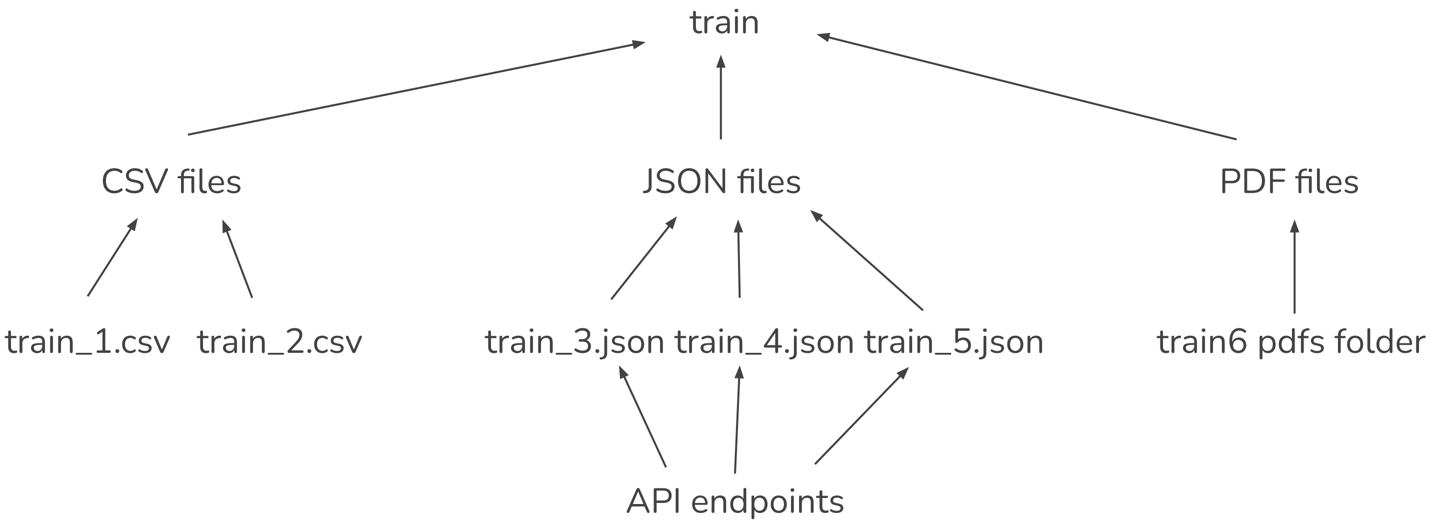
Prediction Models to detect type of Pollutant

Carla Vega
Paulina Campero
Jordi Alfonso

Dataset used

We combined 6 training datasets divided by three types into a single pandas dataframe with 65710 entries.



Modifications to dataset

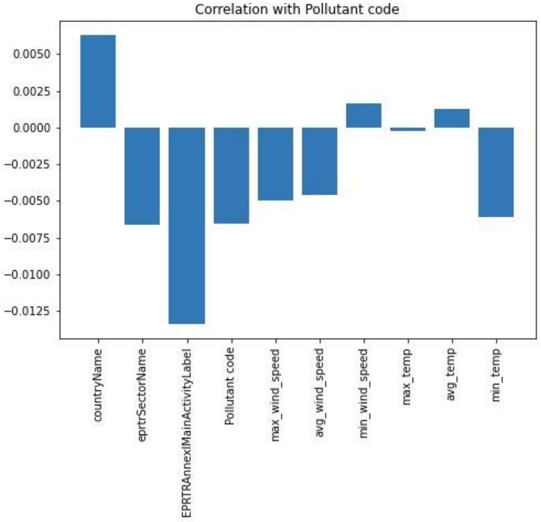
- Columns Removed: City, test_index, EPRTRSectorCode, EPRTRAnnexIMainActivityCode, CONTINENT, FacilityInspireID, facilityName, targetRelease, reportingYear, REPORTER NAME due to not being relevant for the model

- Lowercased strings to make all value combinations align.

- Dropped duplicates of entries to just leave unique entries to the dataset. it dropped to 57212.

- Swapped min and max values due to the data being the other way around.

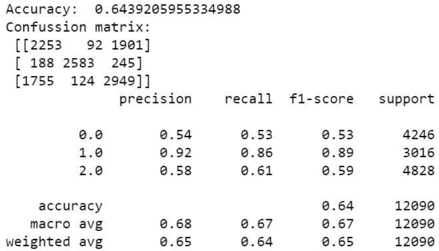
- Encoded labels to be able for the models to process it.



The 3 columns with the least correlation have been dropped too

Our prediction models

Random Forest =>

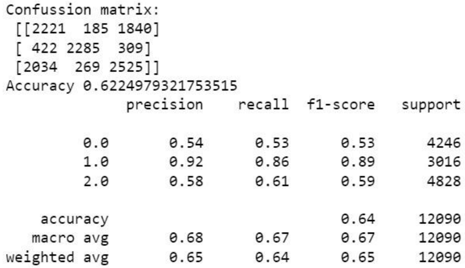


```
Accuracy:  0.6439205955334988
Confussion matrix:
 [[2253   92 1901]
 [ 188 2583  245]
 [1755  124 2949]]
              precision    recall  f1-score   support

         0.0       0.54      0.53      0.53      4246
         1.0       0.92      0.86      0.89      3016
         2.0       0.58      0.61      0.59      4828

    accuracy                           0.64     12090
   macro avg       0.68      0.67      0.67     12090
weighted avg       0.65      0.64      0.65     12090
```

K-Neighbors =>

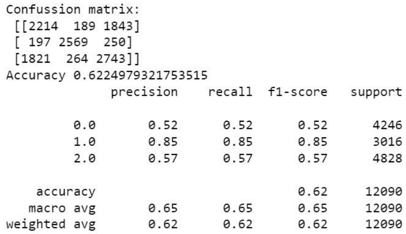


```
Confussion matrix:
 [[2221  185 1840]
 [ 422 2285  309]
 [2034  269 2525]]
Accuracy 0.6224979321753515
              precision    recall  f1-score   support

         0.0       0.54      0.53      0.53      4246
         1.0       0.92      0.86      0.89      3016
         2.0       0.58      0.61      0.59      4828

    accuracy                           0.64     12090
   macro avg       0.68      0.67      0.67     12090
weighted avg       0.65      0.64      0.65     12090
```

Decision Tree =>

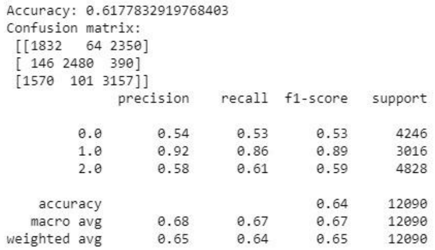


```
Confussion matrix:
 [[2214  189 1843]
 [ 197 2569  250]
 [1821  264 2743]]
Accuracy 0.6224979321753515
              precision    recall  f1-score   support

         0.0       0.52      0.52      0.52      4246
         1.0       0.85      0.85      0.85      3016
         2.0       0.57      0.57      0.57      4828

    accuracy                           0.62     12090
   macro avg       0.65      0.65      0.65     12090
weighted avg       0.62      0.62      0.62     12090
```

Gradient Boosting =>



```
Accuracy: 0.6177832919768403
Confusion matrix:
 [[1832   64 2350]
 [ 146 2480  390]
 [1570  101 3157]]
              precision    recall  f1-score   support

         0.0       0.54      0.53      0.53      4246
         1.0       0.92      0.86      0.89      3016
         2.0       0.58      0.61      0.59      4828

    accuracy                           0.64     12090
   macro avg       0.68      0.67      0.67     12090
weighted avg       0.65      0.64      0.65     12090
```