# Air Pollution Emissions

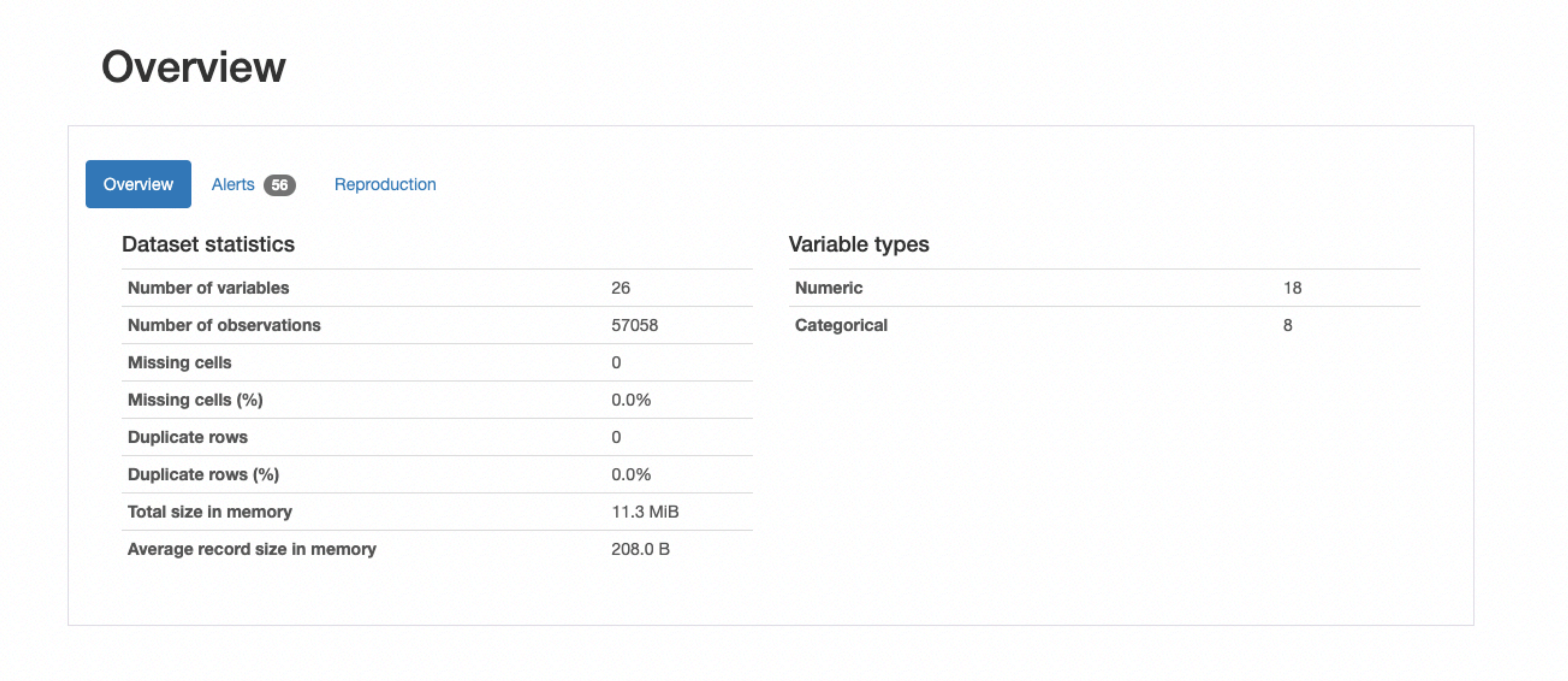
## Predicting Types of a
## Air pollutant

Kyla Juett, Alicja Mańkowska, Leane Macachor - Team 25

# Data Cleaning and EDA

We concatenated all .csv and .json files after solving the API problem with the *delimiter* parameter.
After encoding the categorical variables, we were ready to work with our clean dataset!

Performed a Pandas Profiling Report to help us figure out details about the variables, correlations, missing values, and other statistics.



## Overview

Dataset statistics

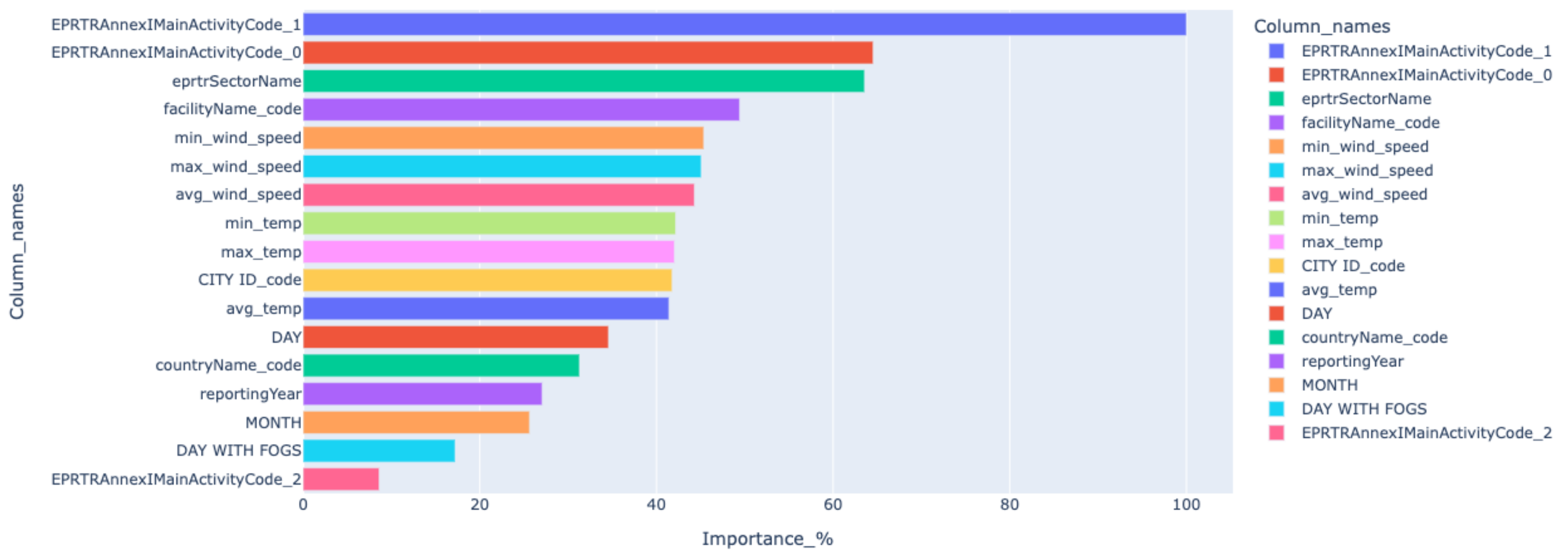
| | |
|---|---|
| Number of variables | 26 |
| Number of observations | 57058 |
| Missing cells | 0 |
| Missing cells (%) | 0.0% |
| Duplicate rows | 0 |
| Duplicate rows (%) | 0.0% |
| Total size in memory | 11.3 MiB |
| Average record size in memory | 208.0 B |

Variable types

| | |
|---|---|
| Numeric | 18 |
| Categorical | 8 |

# Figuring out our Non-Target Variables

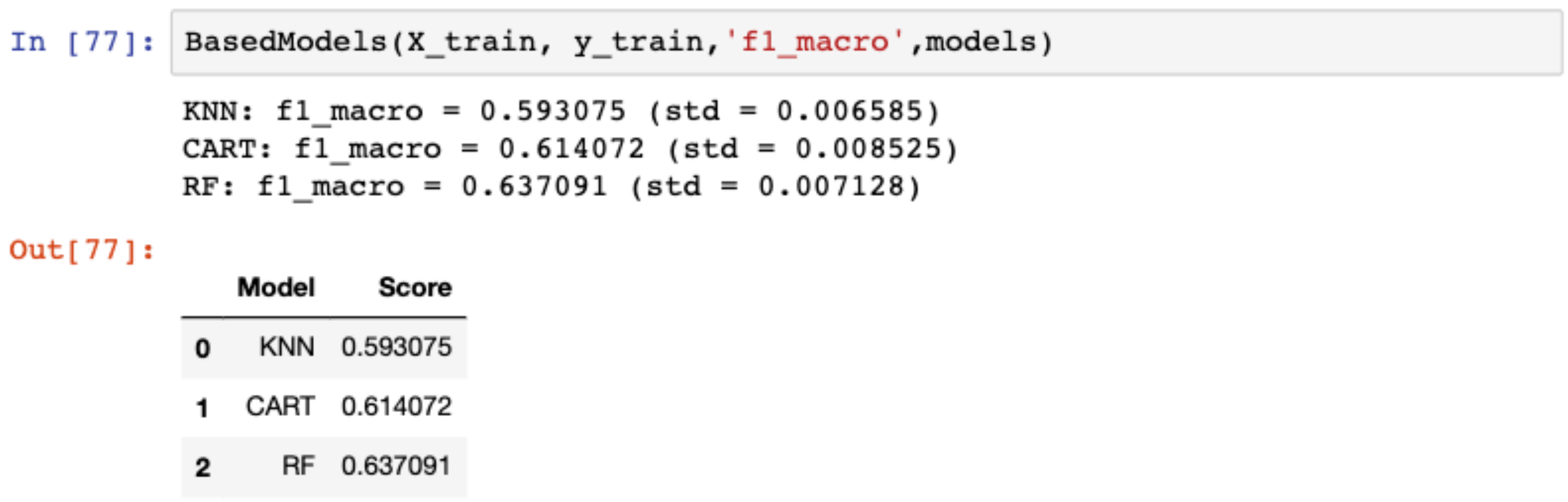
We applied a random forest to understand the ***feature importance*** of our variables.



3

# Pipeline

We built a pipeline to optimize our workflow and help us decide which prediction model type to go for.



```
In [77]: BasedModels(X_train, y_train,'f1_macro',models)

         KNN: f1_macro = 0.593075 (std = 0.006585)
         CART: f1_macro = 0.614072 (std = 0.008525)
         RF: f1_macro = 0.637091 (std = 0.007128)

Out[77]:
```

| | Model | Score |
|---|---|---|
| 0 | KNN | 0.593075 |
| 1 | CART | 0.614072 |
| 2 | RF | 0.637091 |

0.58

F1 score (macro)

4

# Random Forest Classification

We trained our RF to predict the type of pollutant in our test_x

# CONCLUSION

So we have a TERRIBLE test prediction outcome, despite the fairly good train metrics (63-64% F1, 61% accuracy).
This might explain what happened here with our model:

> "F1 is a quick way to tell whether the classifier is actually good at identifying members of a class, or if it is finding shortcuts (e.g., just identifying everything as a member of a large* class)."

Source: Medium (https://medium.com/analytics-vidhya/evaluating-a-random-forest-model-9d165595ad56#:~:text=F1%20score%20is%20a%20little,low%2C%20F1%20will%20be%20low)