01 Schneider Hackathon

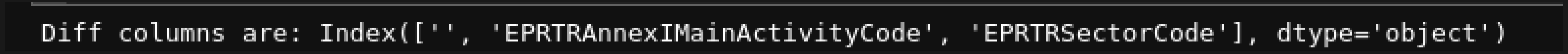
Jordi Arellano Ballestero
Jordi Segura Pons

02

Dataset

We had to deal with Data splitted into CSV and JSON files.

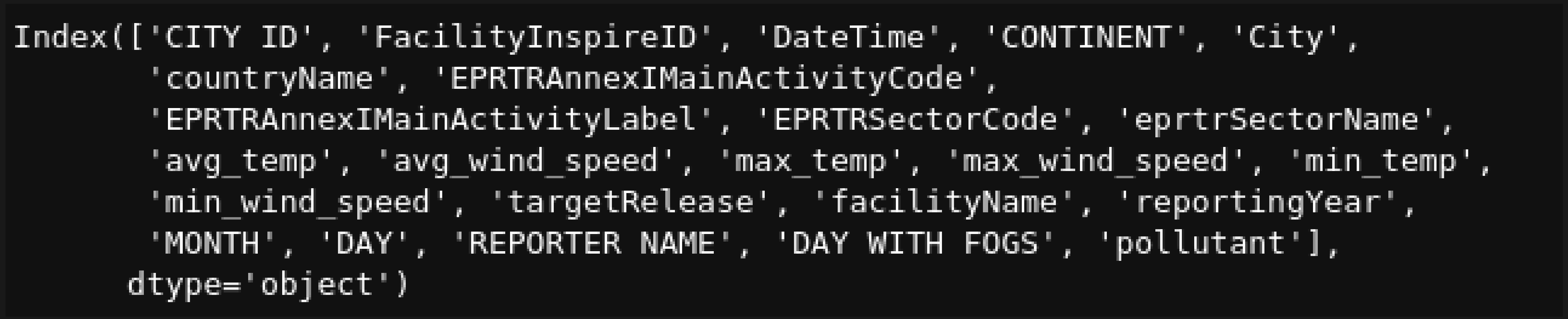
While CSV read was pretty straightforward, JSON contained 2 more columns than CSV.



```
Diff columns are: Index(['', 'EPRTRAnnexIMainActivityCode', 'EPRTRSectorCode'], dtype='object')
```

After having both dataframes unified, we encoded the categorical columns and resorted them.

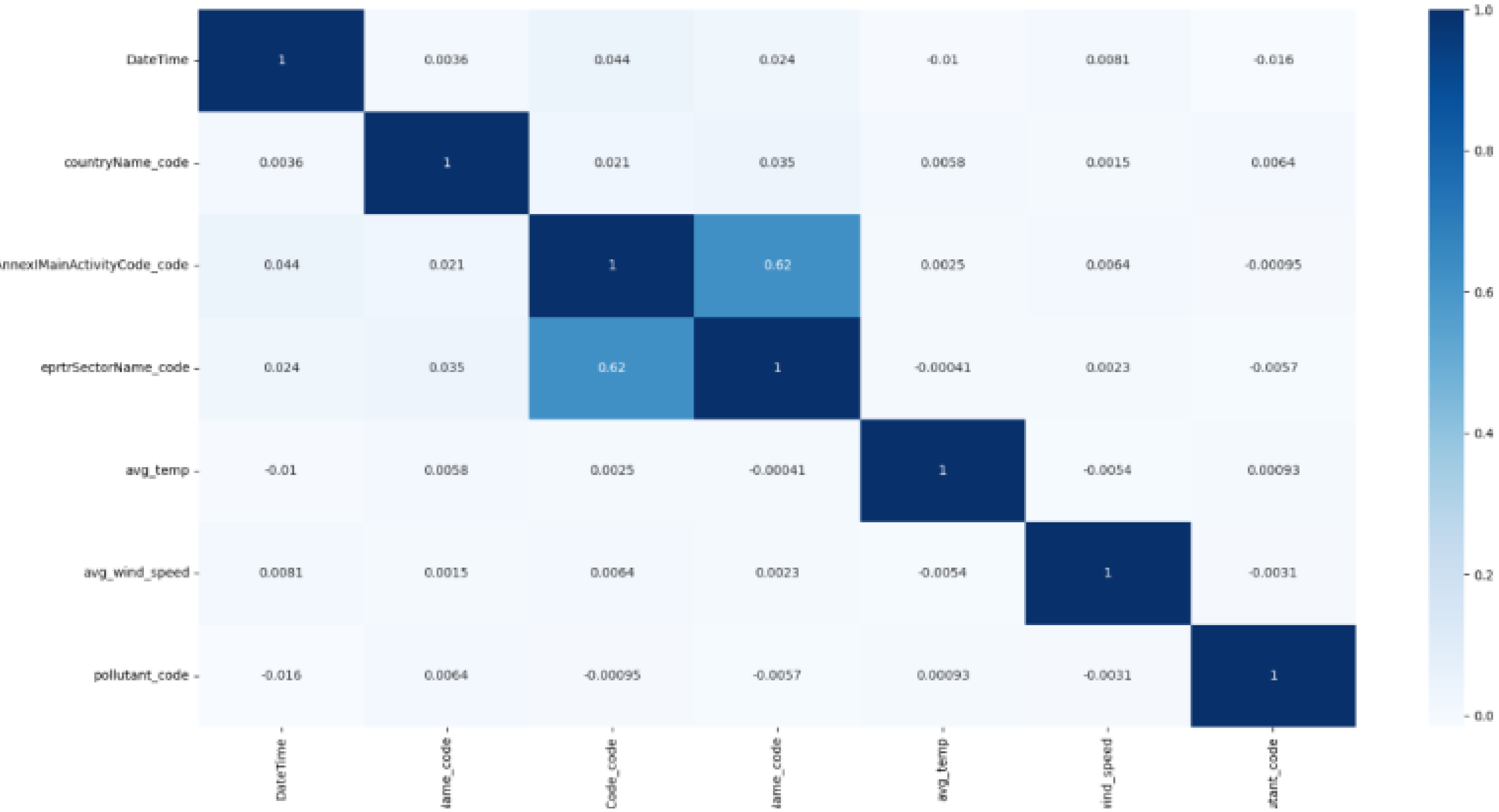
We also changed the date into a datetime due to facility of our models to ingest it.

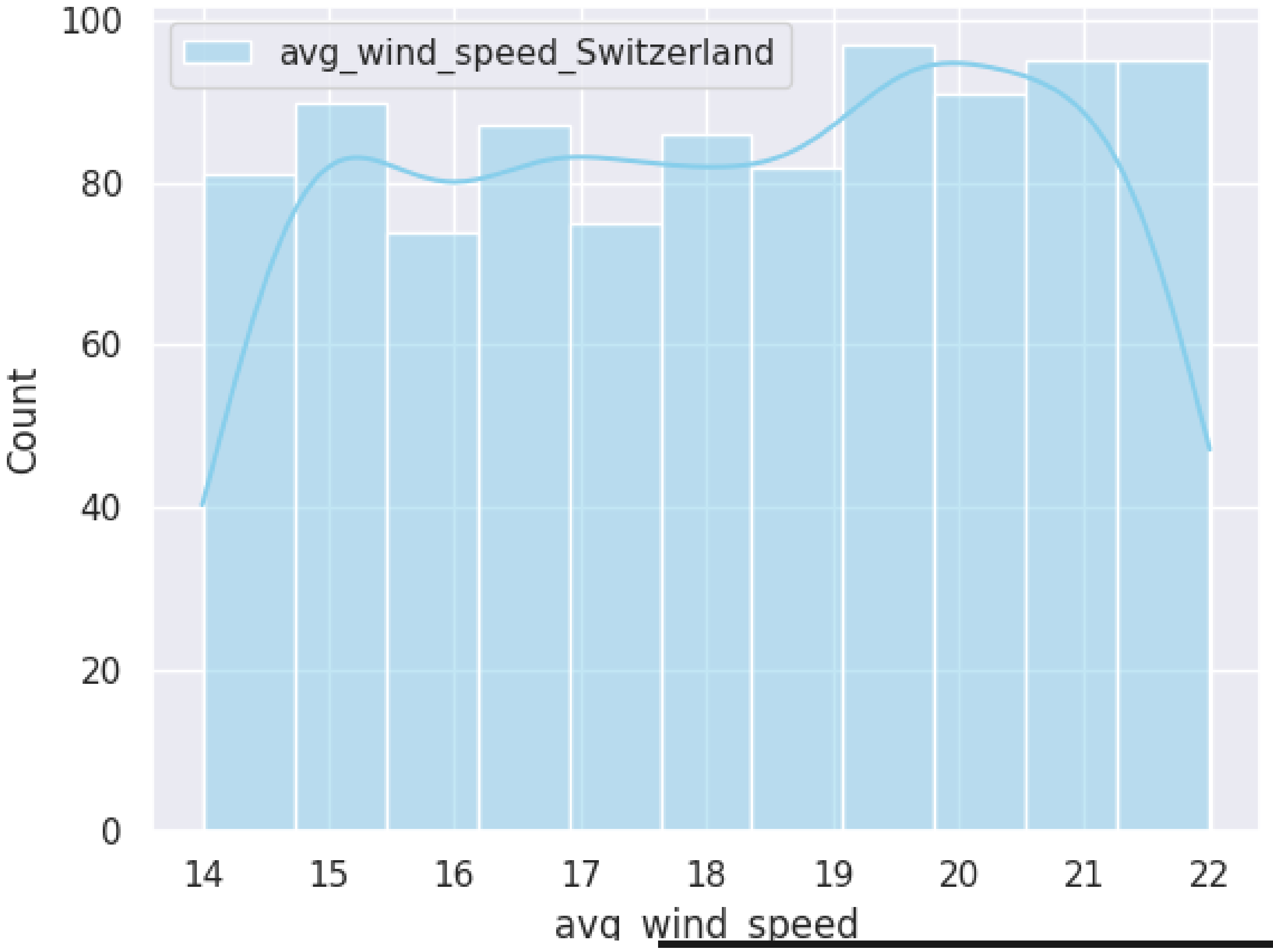


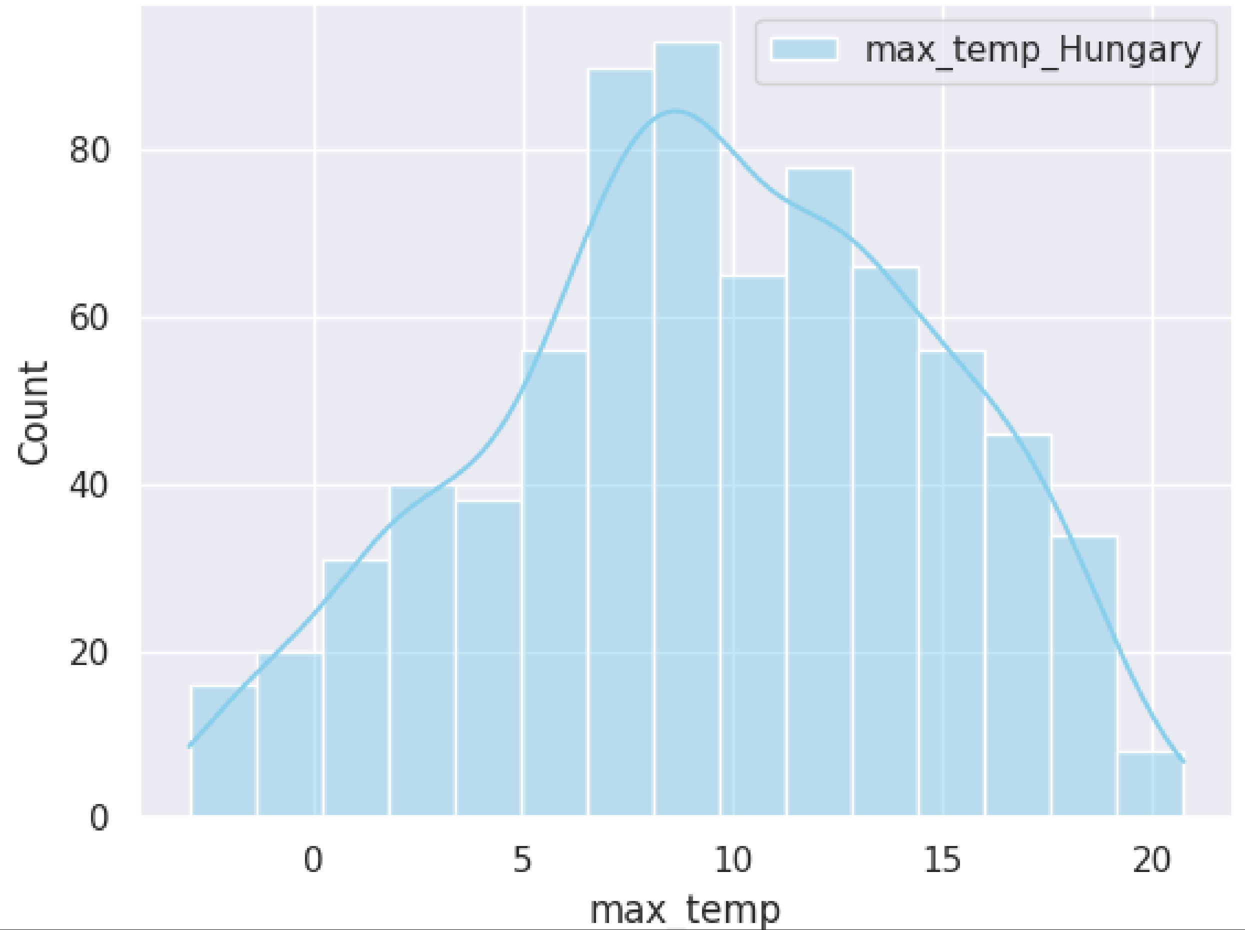
```
Index(['CITY ID', 'FacilityInspireID', 'DateTime', 'CONTINENT', 'City',
       'countryName', 'EPRTRAnnexIMainActivityCode',
       'EPRTRAnnexIMainActivityLabel', 'EPRTRSectorCode', 'eprtrSectorName',
       'avg_temp', 'avg_wind_speed', 'max_temp', 'max_wind_speed', 'min_temp',
       'min_wind_speed', 'targetRelease', 'facilityName', 'reportingYear',
       'MONTH', 'DAY', 'REPORTER NAME', 'DAY WITH FOGS', 'pollutant'],
      dtype='object')
```

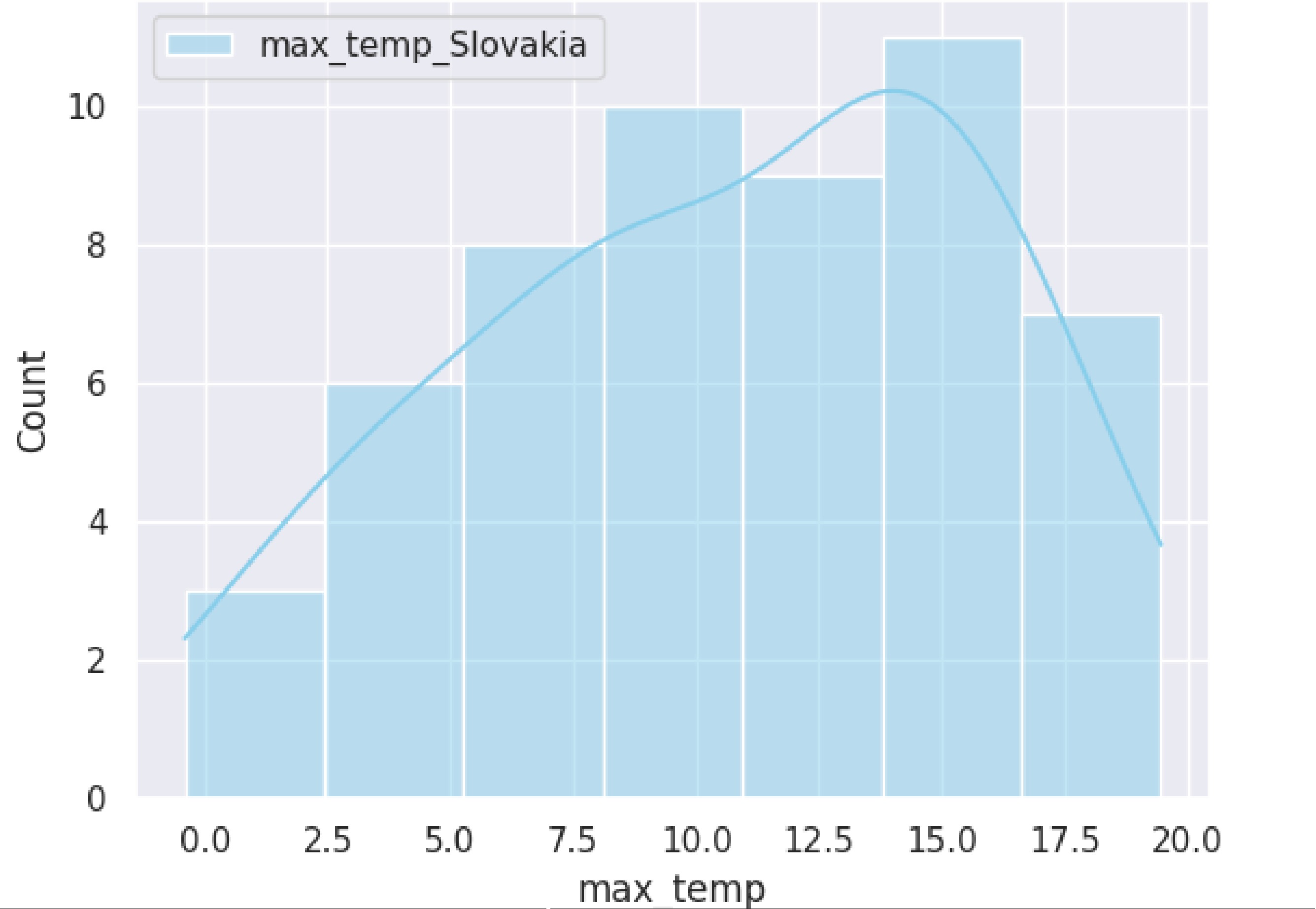
03

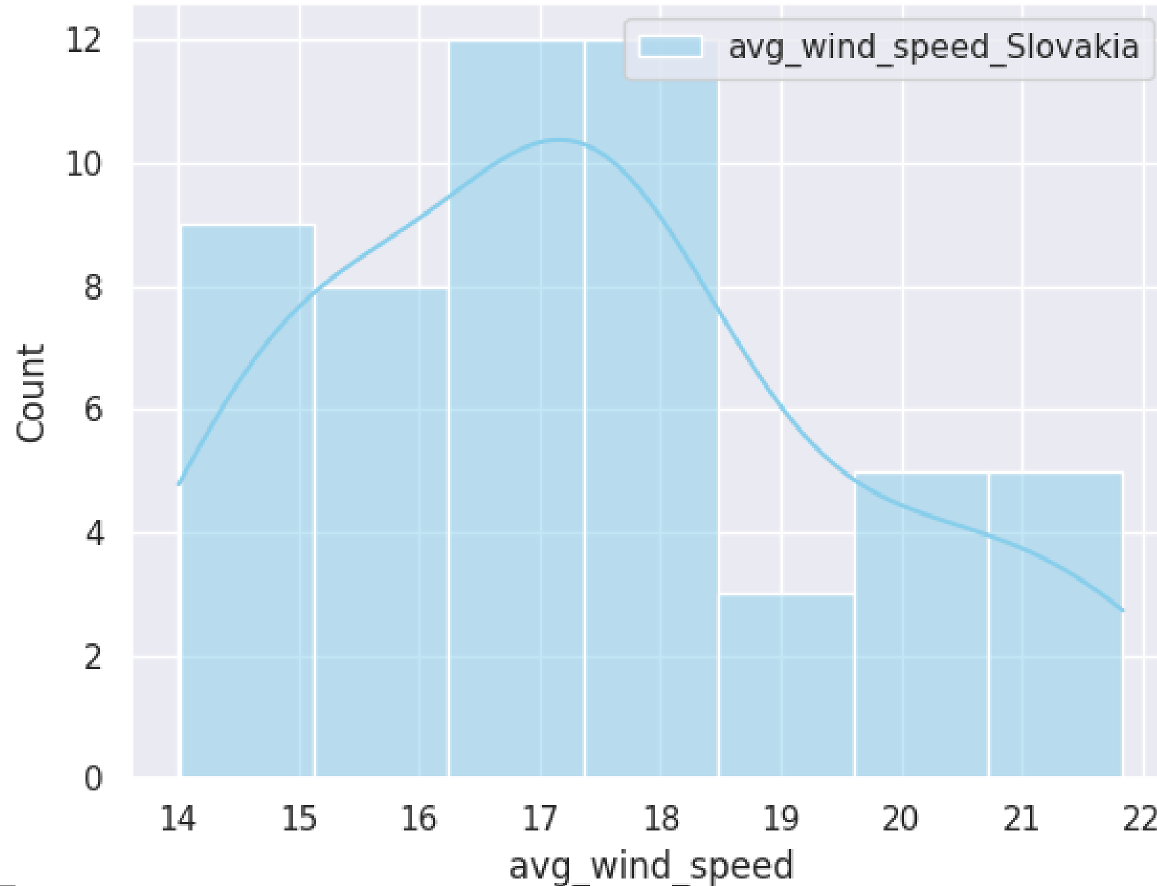
EDA


DateTime - 1 0.0036 0.044 0.024 -0.01 0.0081 -0.016
countryName_code - 0.0036 1 0.021 0.035 0.0058 0.0015 0.0064
annexIMainActivityCode_code - 0.044 0.021 1 0.62 0.0025 0.0064 -0.00095
eprtrSectorName_code - 0.024 0.035 0.62 1 -0.00041 0.0023 -0.0057
avg_temp - -0.01 0.0058 0.0025 -0.00041 1 -0.0054 0.00093
avg_wind_speed - 0.0081 0.0015 0.0064 0.0023 -0.0054 1 -0.0031
pollutant_code - -0.016 0.0064 -0.00095 -0.0057 0.00093 -0.0031 1
DateTime
lame_code
Code_code
lame_code
avg_temp
wind_speed
utant_code
1.0
0.8
0.6
0.4
0.2
0.0

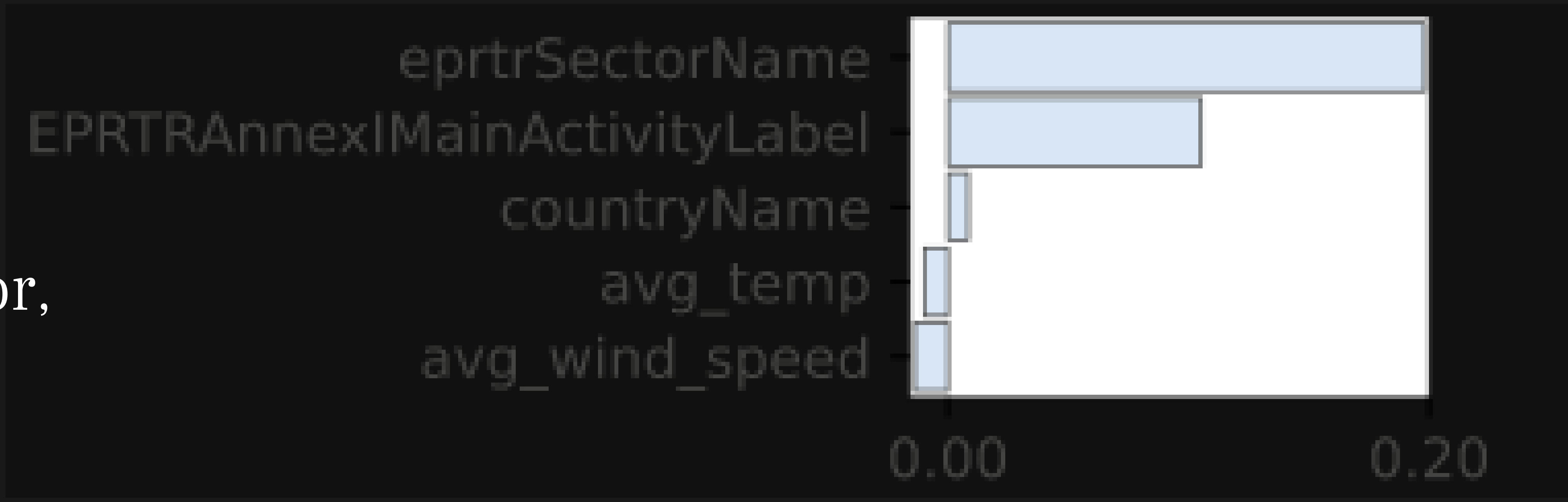

avg_wind_speed_Switzerland
Count
100
80
60
40
20
0
14
15
16
17
18
19
20
21
22
avg_wind_speed


max_temp_Hungary
Count
80
60
40
20
0
0
5
10
15
20
max_temp


max_temp_Slovakia
Count
10
8
6
4
2
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
max_temp


avg_wind_speed_Slovakia
Count
12
10
8
6
4
2
0
14
15
16
17
18
19
20
21
22
avg_wind_speed

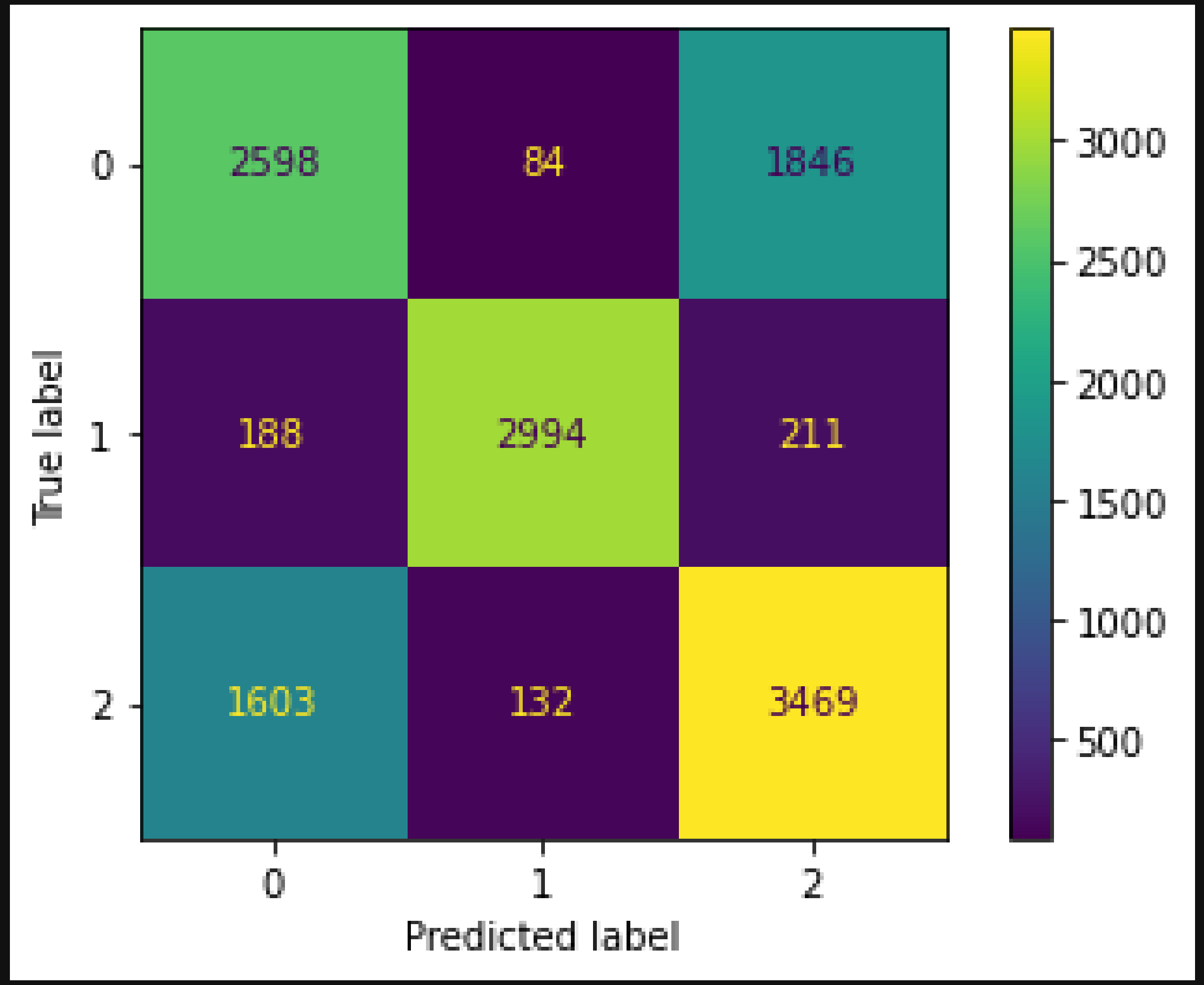
Models



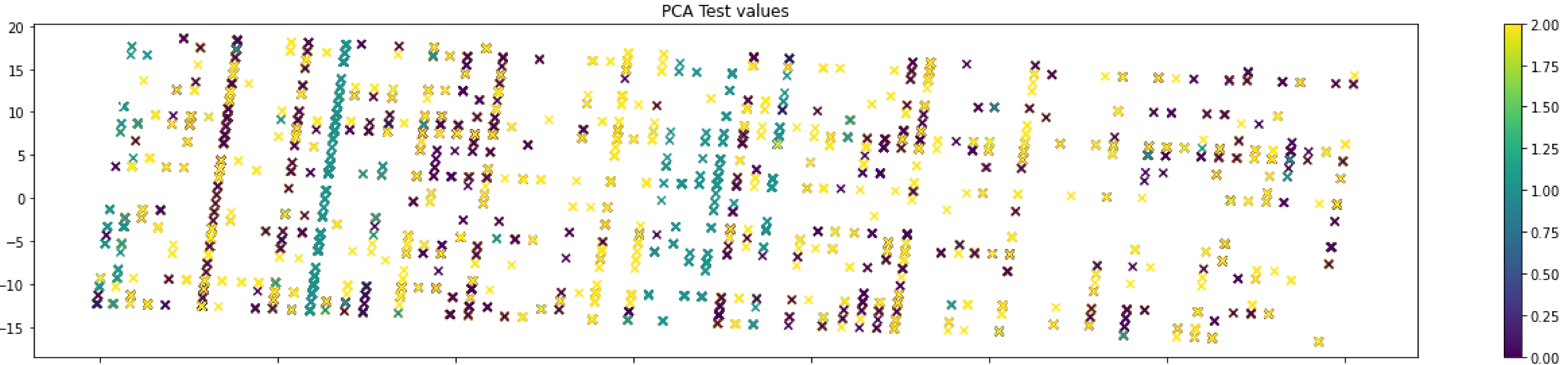
Feature Selection

We end up using only the country, activity, sector, average temperature and wind speed.

We wanted to try classic ML Multiclassification algorithms, such as: RF, Decision Trees, SVC, kNN and XGBoost.

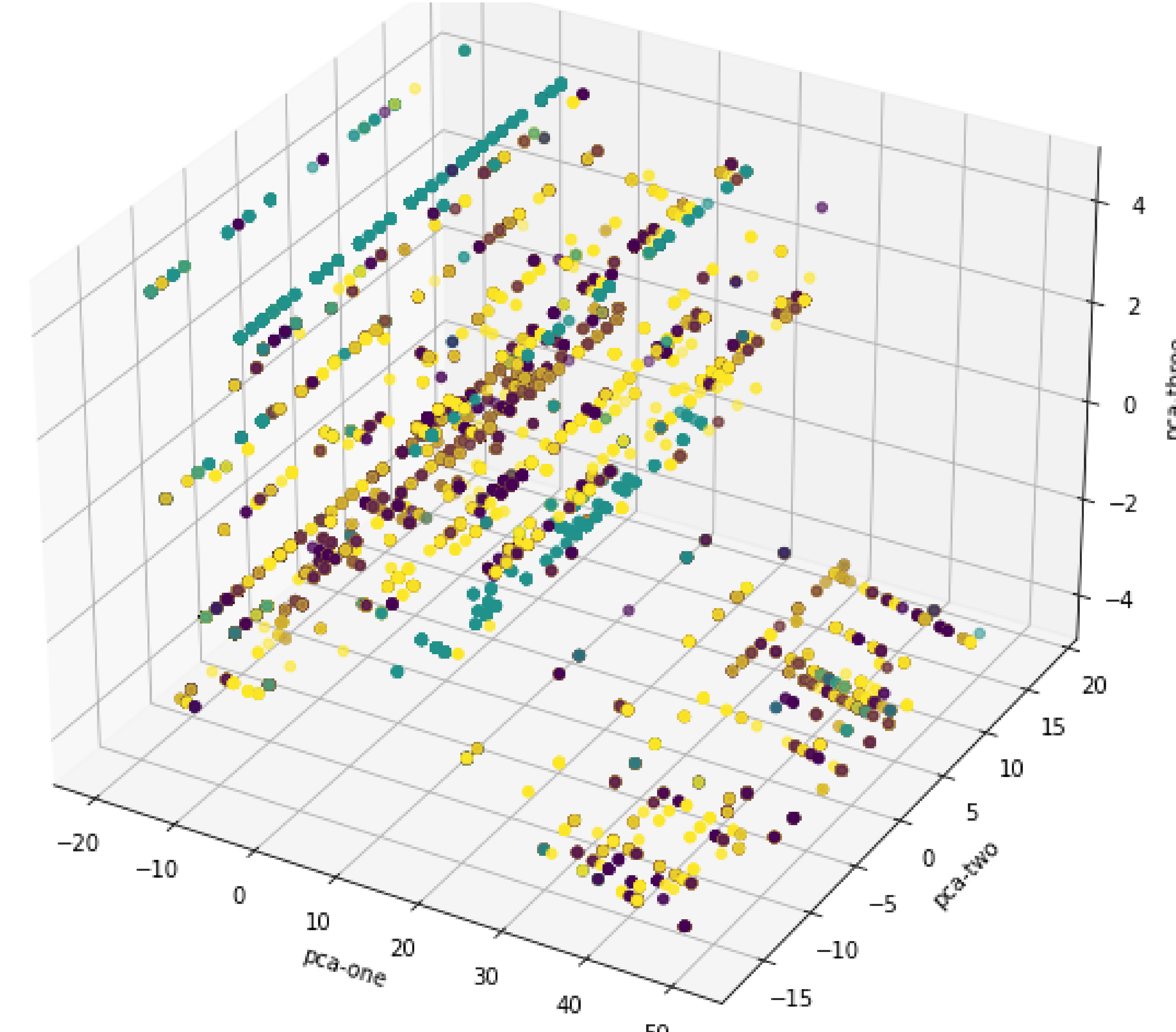


The best results were from a Decision Tree with LightGBM and kNN with PCA.



We can appreciate a structure in the data, where pollunt 1 is really well defined but 0 and 2 are mixed each other. We believe it should be useful to iterate over this idea in order to get more thorough information.

In the future, might be worth trying a resampling and DL algorithms.


pca-one
pca-two