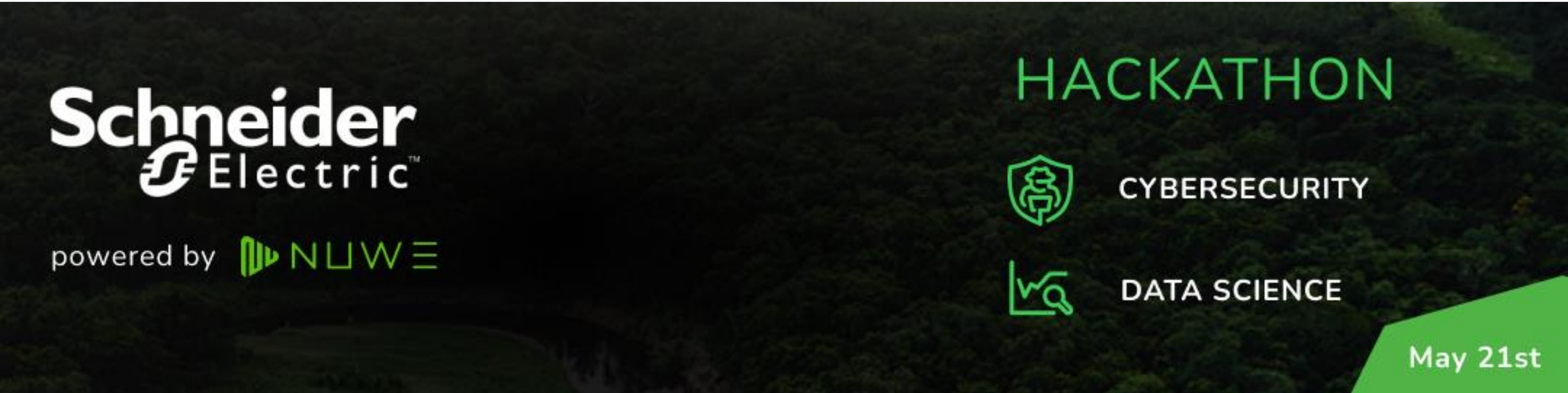
Data Science Hackathon

Lucas Sulzbach Rilho



Introduction

- The main goal of this work was to create a model to predict what kind of pollutant will be produced based on the training data.

Data Descritpion

- Data was collected from 2 CSV files, 3 JSON RESTapi and from 83 PDF reports.
- All sources contained the same kind of information with a little variation on names and field shown.

Methodology

- Firstly the training data was cleaned and joined in a single training dataframe;
- From this dataframe was made a train/test group to test the fields used on the Decision Tree classification;
- It was decided to keep only the fields 'EPRTRSectorCode' and 'EPRTRAnnexIMainActivityCode' as a main indicator of the company production type, that was directly related with the pollutant emited.

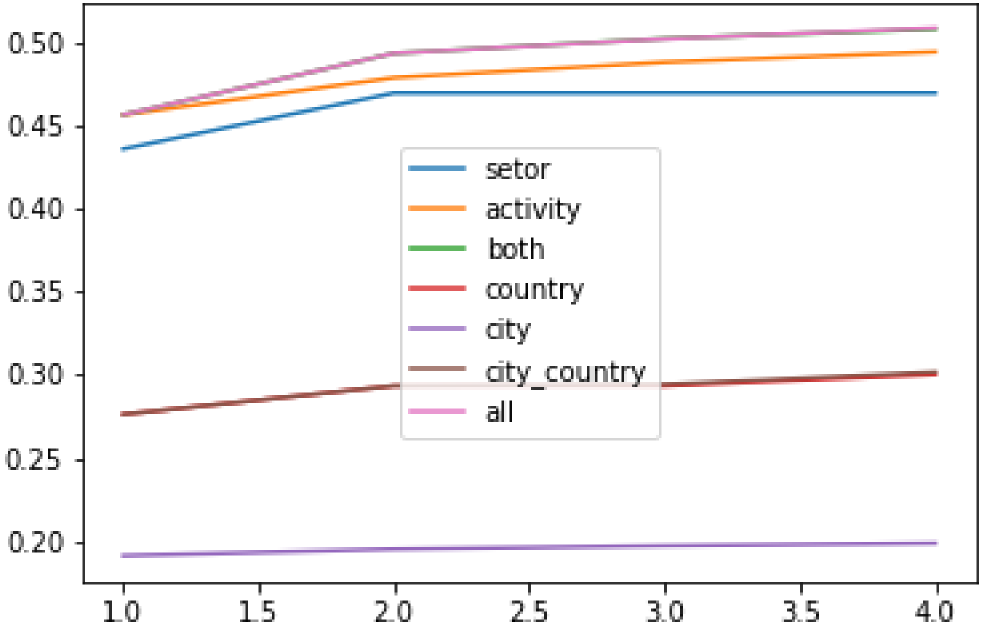
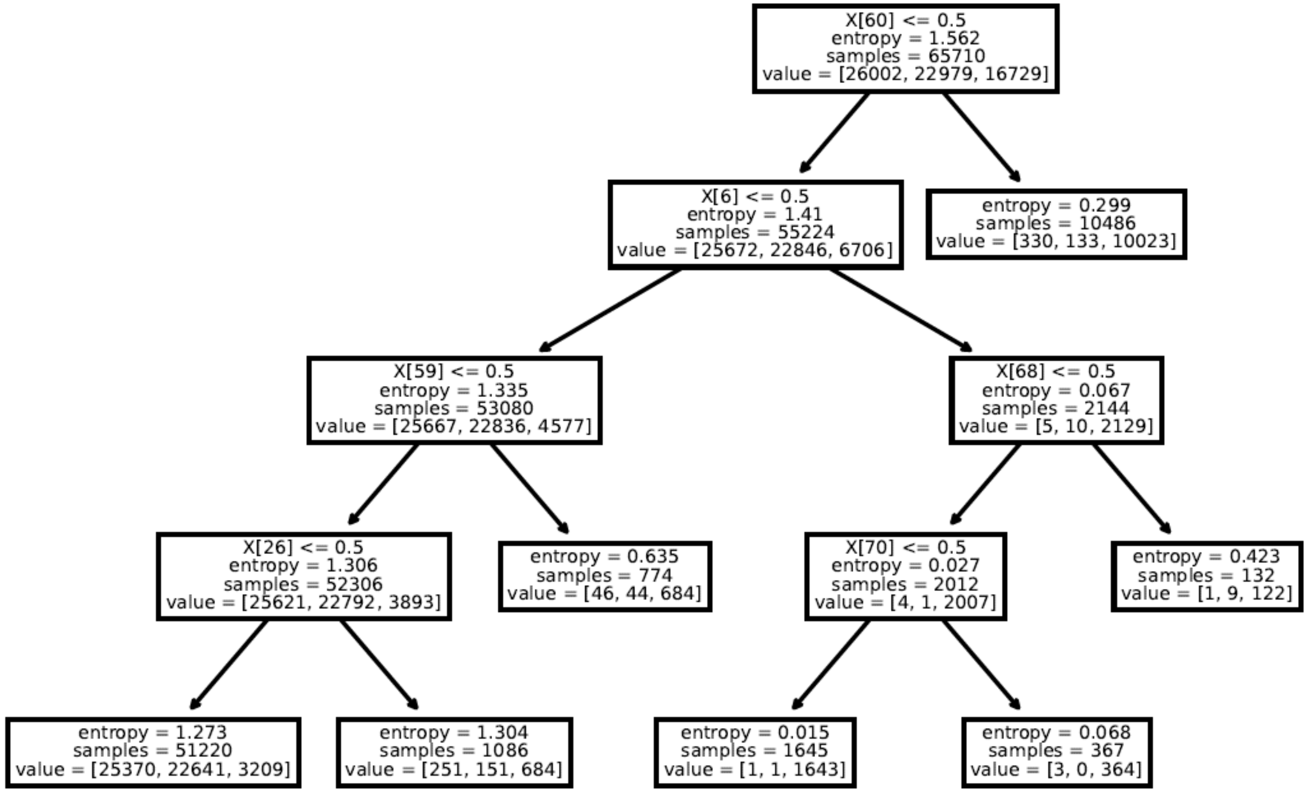


Fig 1. – Training with different fields

Results

- This model have a F1 score of 0.5077352830457011.


X[60] <= 0.5
entropy = 1.562
samples = 65710
value = [26002, 22979, 16729]
X[6] <= 0.5
entropy = 1.41
samples = 55224
value = [25672, 22846, 6706]
entropy = 0.299
samples = 10486
value = [330, 133, 10023]
X[59] <= 0.5
entropy = 1.335
samples = 53080
value = [25667, 22836, 4577]
X[68] <= 0.5
entropy = 0.067
samples = 2144
value = [5, 10, 2129]
X[26] <= 0.5
entropy = 1.306
samples = 52306
value = [25621, 22792, 3893]
entropy = 0.635
samples = 774
value = [46, 44, 684]
X[70] <= 0.5
entropy = 0.027
samples = 2012
value = [4, 1, 2007]
entropy = 0.423
samples = 132
value = [1, 9, 122]
entropy = 1.273
samples = 51220
value = [25370, 22641, 3209]
entropy = 1.304
samples = 1086
value = [251, 151, 684]
entropy = 0.015
samples = 1645
value = [1, 1, 1643]
entropy = 0.068
samples = 367
value = [3, 0, 364]