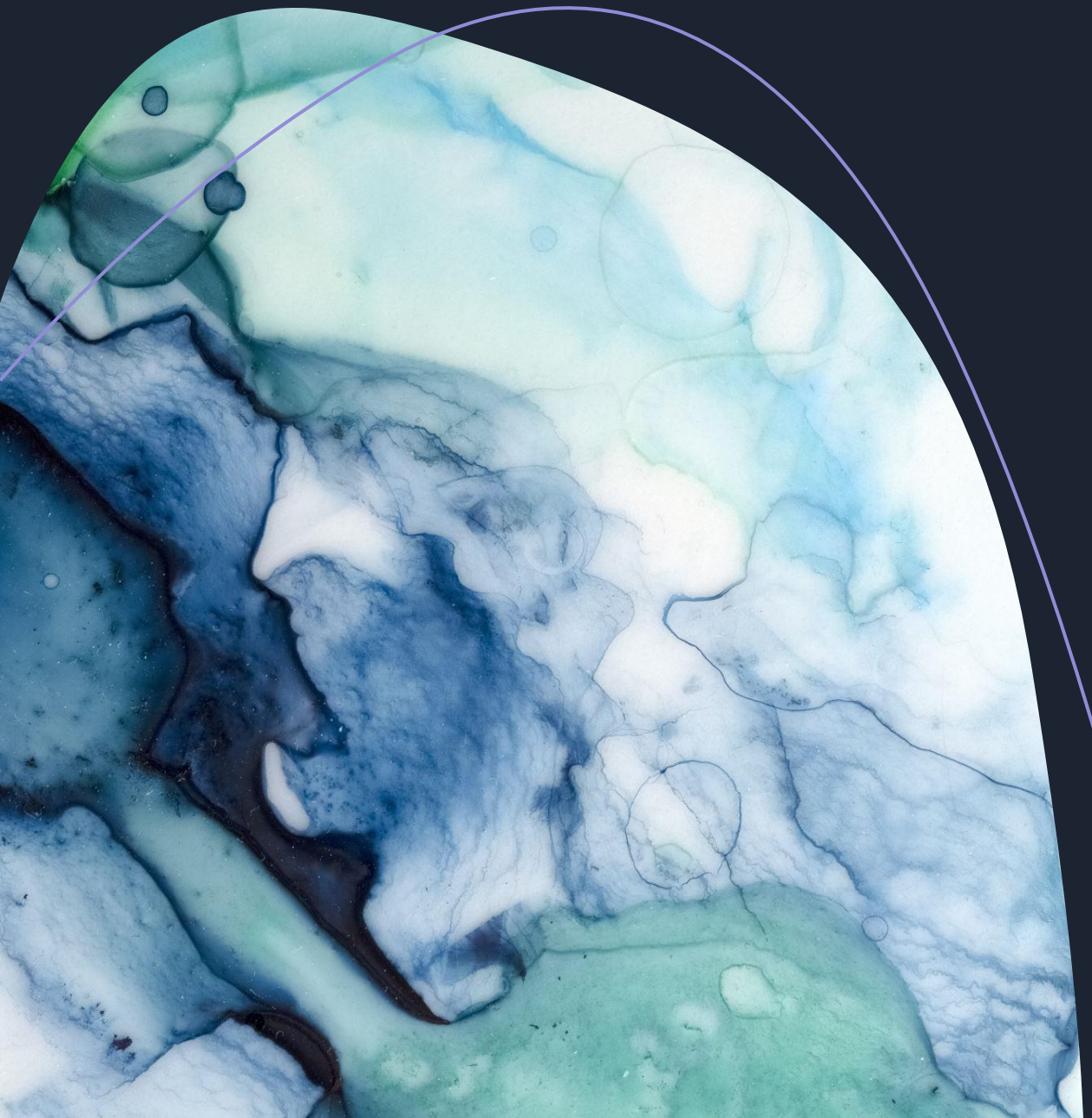

Schneider Electric Hackathon

Data Science Challenge

Luis Carranza

EDA

	train1.csv + train2.csv	train1.csv + train2.csv + json1 + json2 + json3
N° rows	37,127	65,628
Important features	EPRTRAnnexIMainActivityLabel = 0.586 eprtrSectorName = 0.502 countryName = 0.111	EPRTRAnnexIMainActivityLabel = 0.363 eprtrSectorName = 0.304 countryName = 0.066

Data analysis before merge datasets ✓

Feature ranking based on Corr matrix ✓

PDFs automation data retrieval ✓

ML Experiments - LightGBM

train1.csv + train2.csv	train1.csv + train2.csv + json1 + json2 + json3
f1_macro: 0.648 (0.007)	f1_macro: 0.470 (0.004)

CSV + JSON decision chosen

Less precise

More robust

Stack

- Anaconda for the Jupiter Notebook and DS libraries
- LightGBM for fast ML experimentation
- Pymupdf for PDF text extraction