Schneider Electric Hackathon
Data Science

Paso1: Recolectar DataSet

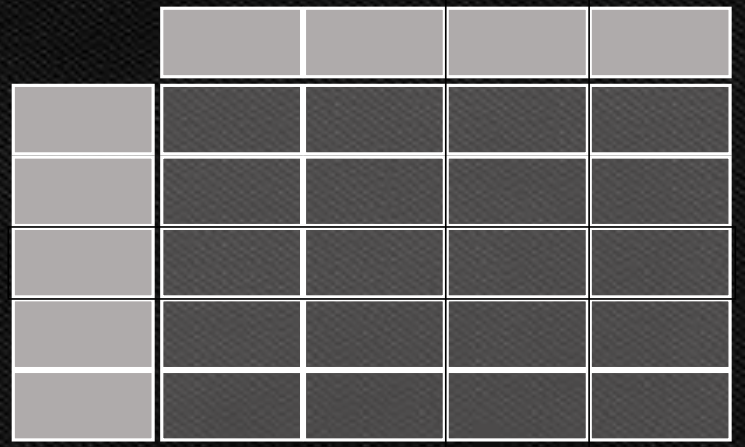
- train1.csv
- train2.csv

- train3.json
- train4.json
- train5.json

x 82

- pdfs81515.pdf
- pdfs81516.pdf
- …
- Pdfs81596.pdf

DataFrame



"Esta parte del proceso consiste en juntar todo los Datasets a una tabla en pandas con las variables de Interes para la clasificación"

Paso2: Limpiar dataset

	countryName	eprtrSectorName	EPRTRAnnexIMainActivityLabel	FacilityInspireID	facilityName	City	targetR
0	Germany	Mineral industry	Installations for the production of cement cli...	https://registry.gdi-de.org/id/de.ni.mu/062217...	Holcim (Deutschland) GmbH Werk Höver	Sehnde	
1	Italy	Mineral industry	Installations for the production of cement cli...	IT.CAED/240602021.FACILITY	Stabilimento di Tavernola Bergamasca	TAVERNOLA BERGAMASCA	
2	Spain	Waste and wastewater management	Landfills (excluding landfills of inert waste ...	ES.CAED/001966000.FACILITY	COMPLEJO MEDIOAMBIENTAL DE ZURITA	PUERTO DEL ROSARIO	
3	Czechia	Energy sector	Thermal power stations and other combustion in...	CZ.MZP.U422/CZ34736841.FACILITY	Elektrárny Prunéřov	Kadaň	
4	Finland	Waste and wastewater management	Urban waste-water treatment plants	http://paikkatiedot.fi/so/1002031/pf/Productio...	TAMPEREEN VESI LIIKELAITOS, VIINIKANLAHDEN	Tampere	

	reportingYear	pollutant	MONTH	DAY	max_wind_speed	avg_wind_speed	min_wind_speed	max_temp	avg_temp	min_temp	...	eprtrSectorN
0	0.666667	1	0.818182	0.703704	0.641688	0.626823	0.699173	0.249436	0.314744	0.366280	...	
1	1.000000	0	0.727273	0.740741	0.848447	0.872242	0.711089	0.357326	0.448394	0.463555	...	
2	1.000000	2	0.090909	0.111111	0.532941	0.645728	0.546712	0.193219	0.283346	0.322292	...	
3	0.333333	0	0.636364	0.185185	0.493206	0.714689	0.562020	0.586043	0.559031	0.595002	...	
4	1.000000	2	1.000000	0.777778	0.732404	0.912700	0.703303	0.619338	0.606565	0.630816	...	

5 rows × 21 columns

“En esta parte del proceso se trabajó el desbalanceo de datos con SMOTE También se normalizaron las variables cuantitativas, además se aplicó un One-Hot-Encoding a las variables categóricas.”

Paso3: Entrenamiento del modelo

```python
modelos = {
    "XGB": XGBClassifier(),
    "LGB": LGBMClassifier(),
    "MLP": MLPClassifier(max_iter=5000),
    "RFC": RandomForestClassifier(),
    "ABC": AdaBoostClassifier(),
    "DTC": DecisionTreeClassifier(),
    "KNC": KNeighborsClassifier()
}

for name, model in modelos.items():
  model.fit(X_train, y_train)
  y_pred = model.predict(X_test)
  f1 = f1_score(y_test, y_pred, average="macro")
  print(f"F1 {name:<5}: {f1:.3f}")
```

```
F1 XGB  : 0.615
F1 LGB  : 0.613
F1 MLP  : 0.578
F1 RFC  : 0.681
F1 ABC  : 0.586
F1 DTC  : 0.609
F1 KNC  : 0.611
```

“De los 7 modelos entrenados,
Es RandomForestClassifier el
Que obtiene un F1 score mejor
Al resto.”

Paso4: Entrenamiento del modelo

“Este modelo de Random Forest puede clasificar categorias de contaminación (NOX, CO2, CH4) hasta con un 68% de certeza

https://github.com/cesarmsq/NUWEproject/