# Prédiction du Niveau de la Pollution Atmosphérique des Villes

Antoine Ghidini / Laeitia Kamwag / Sebila Doubaeva

Projet d'Application ING2,
Spécialité AI, ESME Sudria

**99% DE POPULATION**

Respirent un air qui dépasse les limites fixées

**IMPACT SUR LA SANTE**

Cardiopathies, cancers du poumon, de système cardi-vasculaire

**7 MILLIONS**

Meurent chaque année suite à la pollution de l'air

# Quelles sont nos motivations?

Amener le projet au stade de l'utilité pour tous : non seulement apprendre, mais aussi partager nos résultats !

**ALGORITHMES CONCRETS**

**DONNEE REELLES**

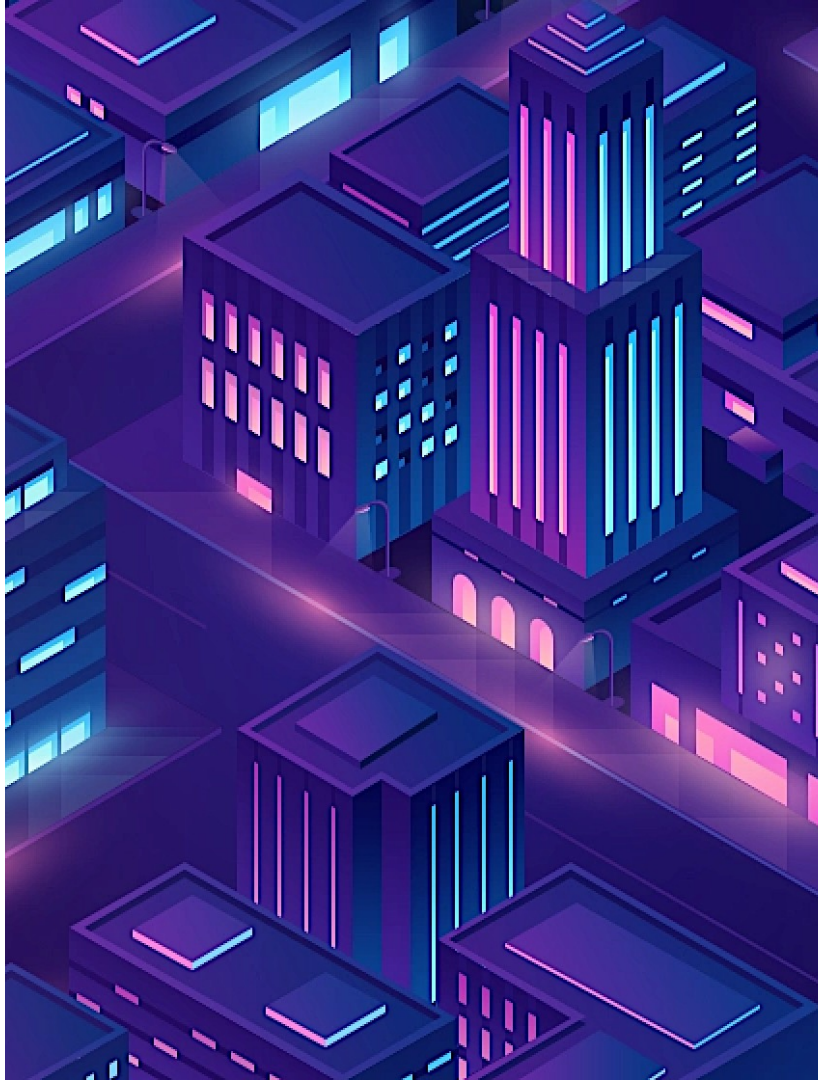
**APPLICATION RESULTANTE**

ESME

# Machine Learning

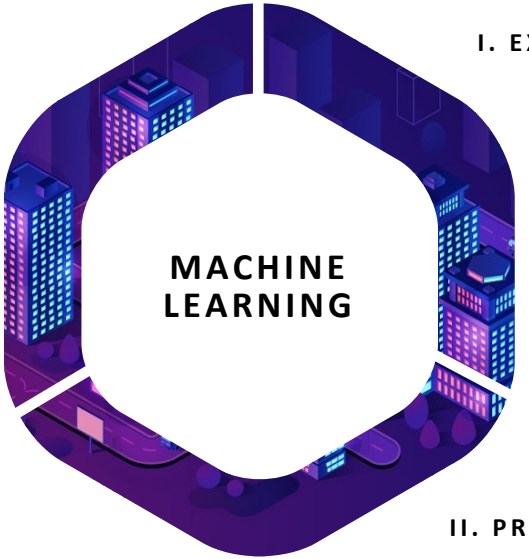
- Numpy
- Matplotlib, Seaborn
- Pandas, Skit Learn

- Jupyter Notebook
- GitHub
- Dash & Heroku



# Méthodologie

Les étapes à suivre dans
un cycle de projet en ML




## I. EXPLORATORY DATA ANALYSIS

- Comprendre les données
- Identifier les erreurs évidentes
- Détecter les anomalies
- Trouver des relations

## II. PREPROCESSING

- Amputation de données
- Validation et nettoyage de données brutes
- Encodage & Fenêtre glissante

## III. LEARNING

- Decision Tree Regressor
- Random Forest Regressor
- Linear Regression

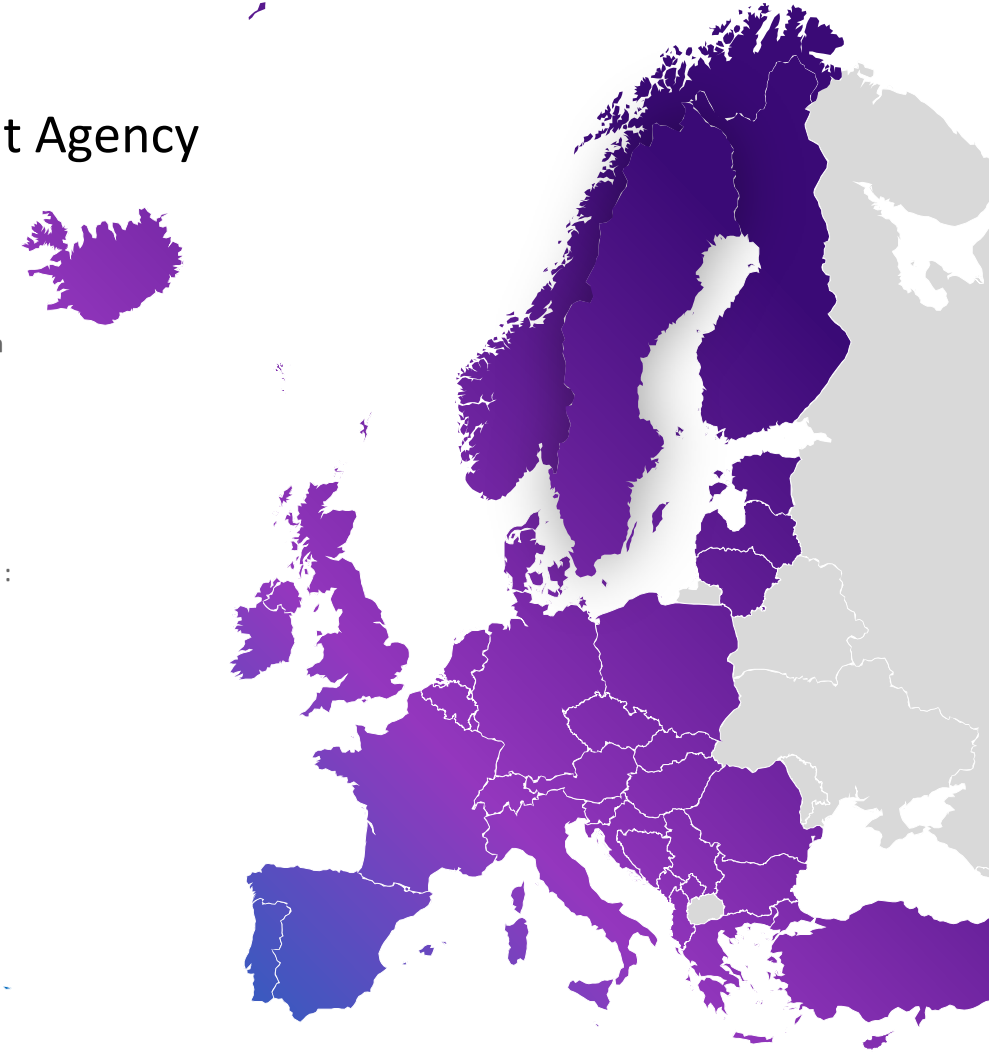
ESME

# European Environment Agency

- Convention sur la pollution atmosphérique à longue distance (LRTAP Convention)

- Inventaire de la Directive sur les Plafond d'Emission Nationaux (NEC)

- Portail Européen des Emissions Industrielles : Directive sur les Emissions Industrielles et du Registre Européen des Rejets et Transferts de Polluants (E-PRTR)



# DataSets

- 2 732 184 (72% ) données d'émissions manquantes pour dataset de LRTAPC

- 95 données d'émissions manquantes pour dataset de E-PRTR
- 224 villes manquantes
- 19 nom d'installations

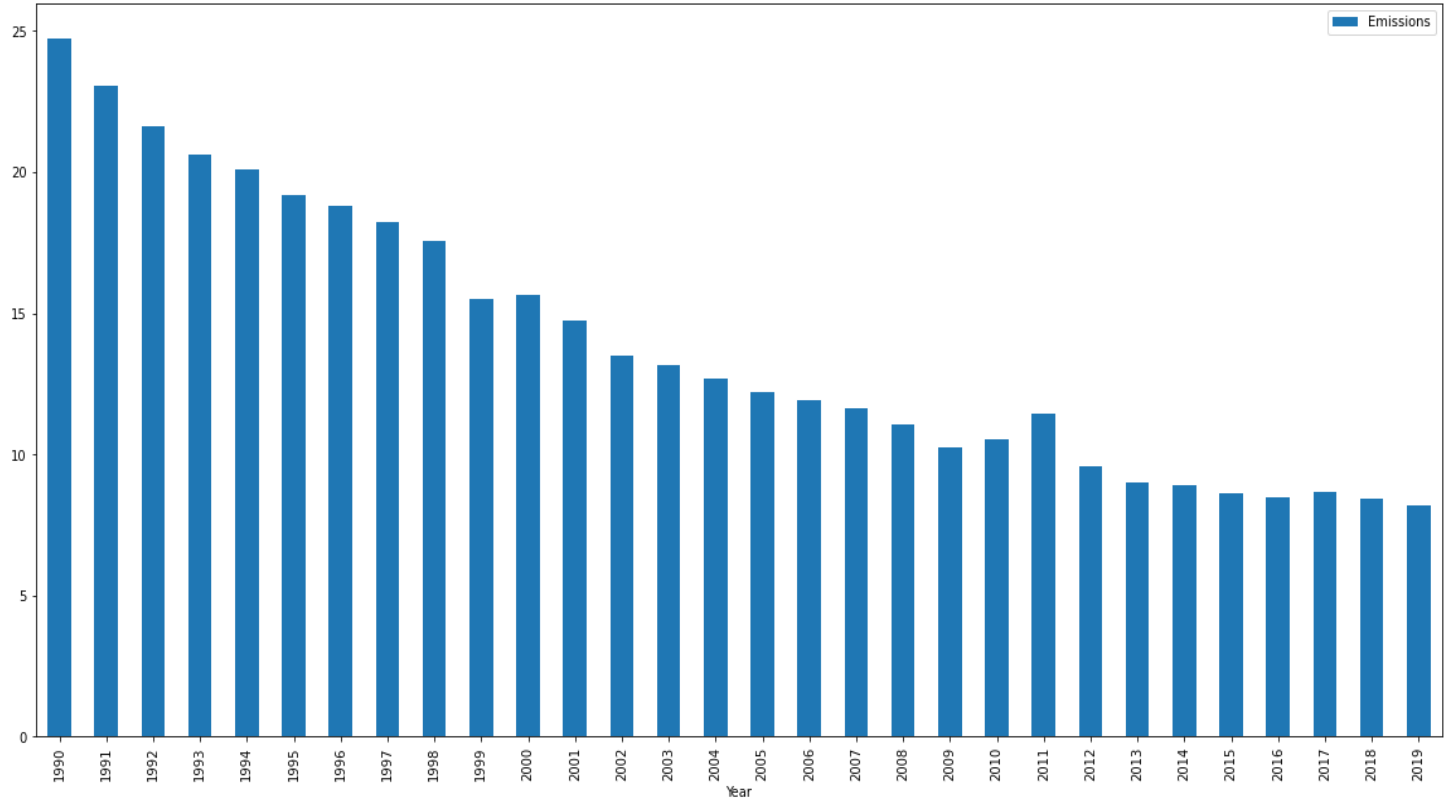
| RangeIndex : 3794700 entries | | |
|---|---|---|
| # | Column | Dtype |
| 0 | Country_Code | object |
| 1 | Country | object |
| 2 | Pollutant_name | object |
| 3 | Format_name | object |
| 4 | Sector_code | object |
| 5 | Year | int64 |
| 6 | Emissions | float64 |
| 7 | Unit | object |
| 8 | Notation | object |
| 9 | VersionId | int64 |
| 10 | Parent_sector_code | object |
| 11 | Sector_name | object |

| RangeIndex : 288566 entries | | |
|---|---|---|
| # | Column | Dtype |
| 0 | countryName | object |
| 1 | EPRTRSectorCode | object |
| 2 | eprtrSectorName | object |
| 3 | EPRTRAnnexIMainActivityCode | object |
| 4 | EPRTRAnnexIMainActivityLabel | object |
| 5 | FacilityInspireID | object |
| 6 | facilityName | object |
| 7 | Longitude | float64 |
| 8 | Latitude | float64 |
| 9 | City | object |
| 10 | targetRelease | object |
| 11 | pollutant | object |
| 12 | emissions | float64 |
| 13 | reportingYear | int64 |

ESME

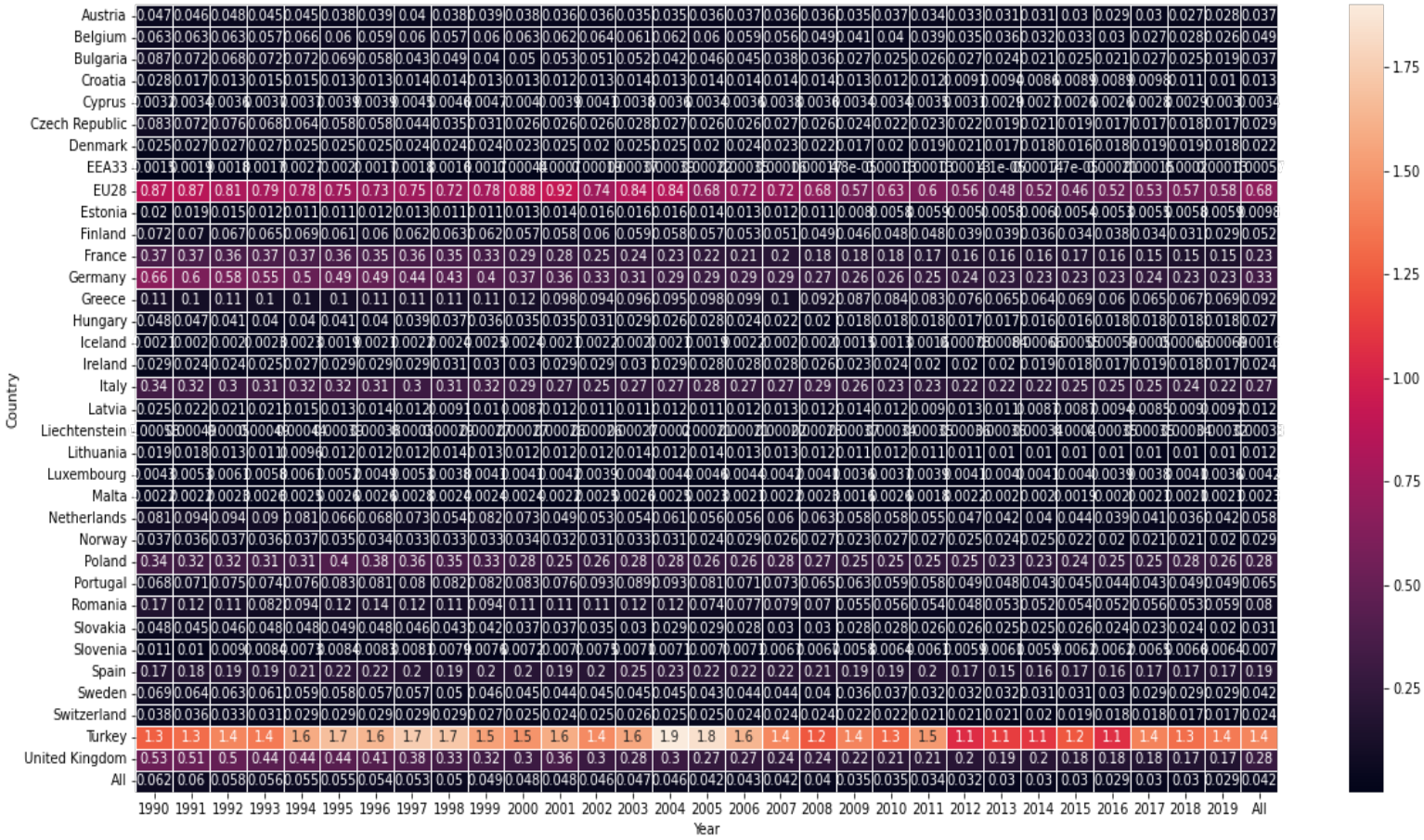
# EDA - Premier Dataset

Graphique d'évolution du niveau des émission de dataset de LRTAP Convention



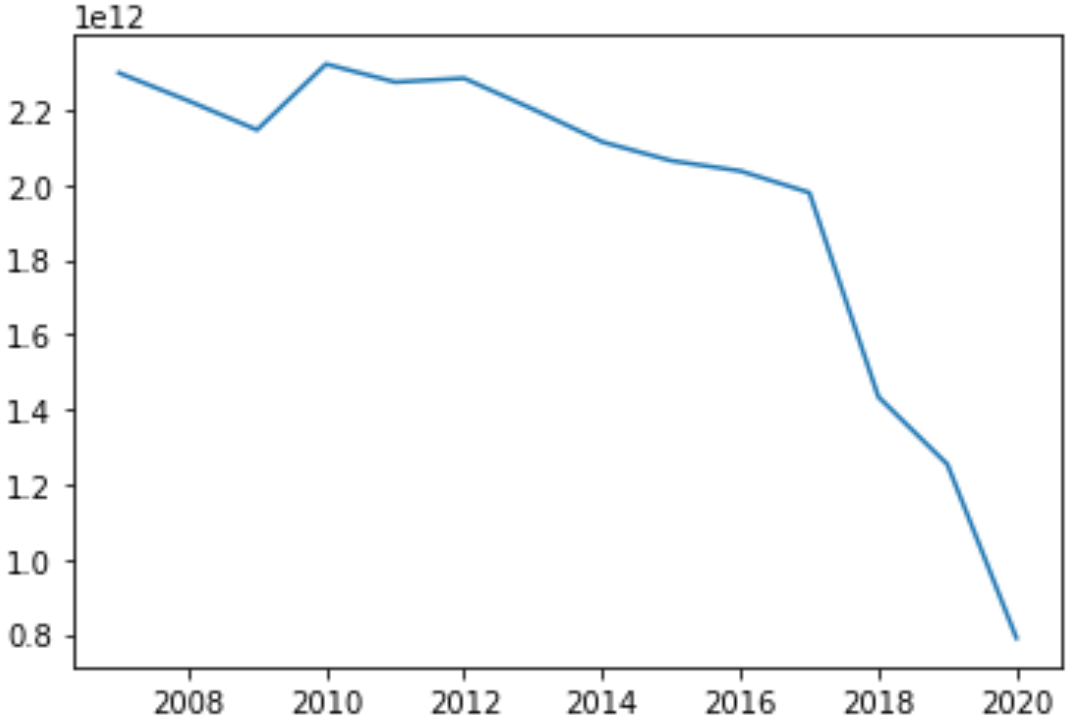
ESME

Heatmap avec Seaborn de dataset de LRTAP Convention



ESME

# EDA - Second Dataset



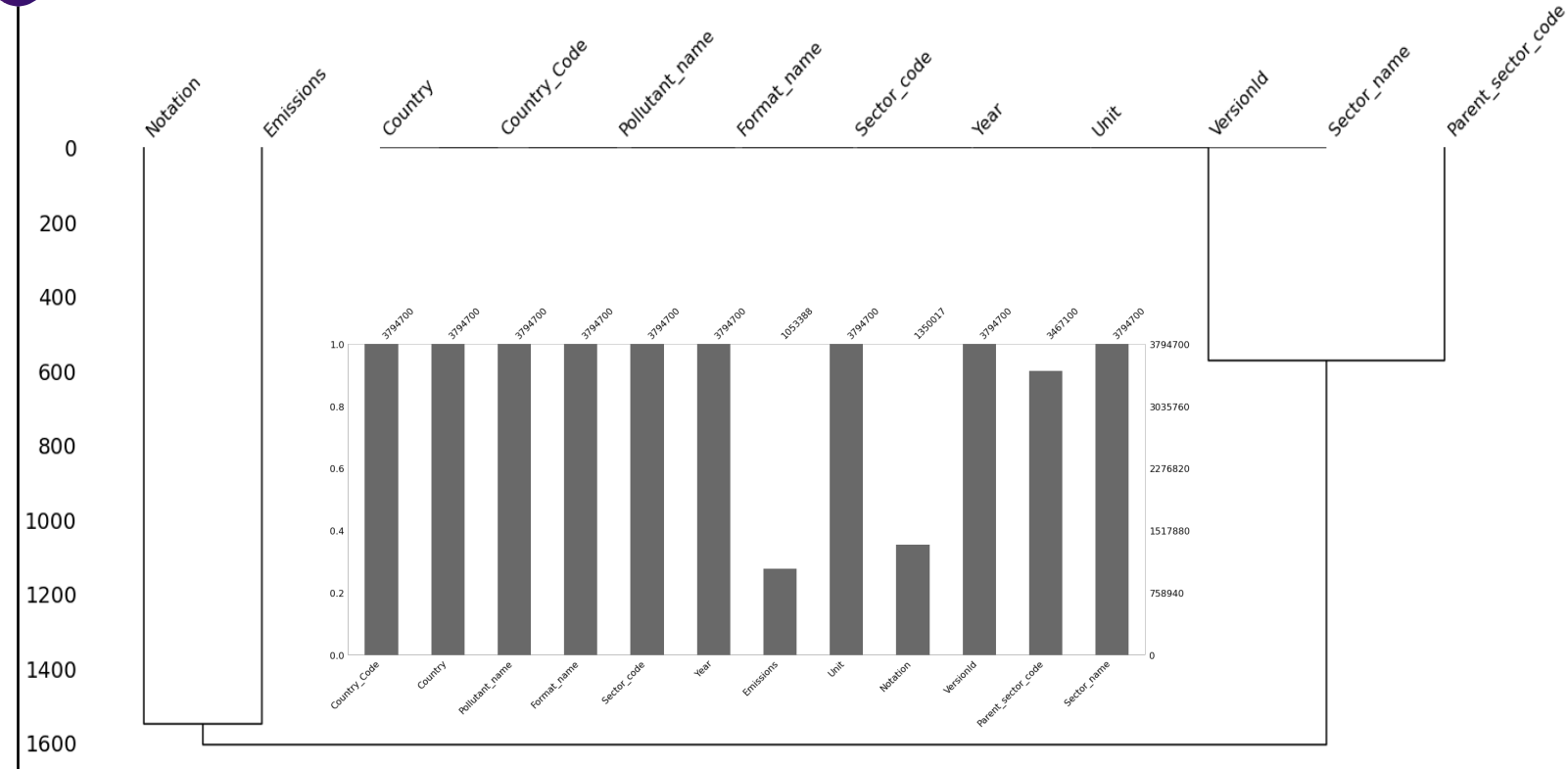

Evolution des émissions en Europe entre 2007 et 2020 pour le dataset de E-PRTR

ESME


1e13
2.5
2.0
1.5
1.0
0.5
0.0
Emissions par polluants en Europe pour le dataset de E-PRTR
Brominated diphenylethers (PBDE)
Chlordecone
1,2,3,4,5,6-hexachlorocyclohexane (HCH)
Lindane
Asbestos
Fluorides (as total F)
Nonylphenol and Nonylphenol ethoxylates
Phenols (as total C)
PCDD + PCDF (dioxins + furans) (as Teq)
Hexachlorobenzene (HCB)
Ethyl benzene
Polychlorinated biphenyls (PCBs)
Trichlorobenzenes (TCB)
Pentachlorophenol (PCP)
Pentachlorobenzene
Aldrin
Benzo(g,h,i)perylene
1,1,2,2-tetrachloroethane (TETRACHLOROETHANE-1,1,2,2)
Di-(2-ethyl hexyl) phthalate (DEHP)
Anthracene
Halons
Total organic carbon(as total C or COD/3) (TOC)
Cadmium and compounds (as Cd)
Fine particulate matter (PM2.5)
Sulphur hexafluoride (SF6)
Xylenes
Mercury and compounds (as Hg)
Arsenic and compounds (as As)
Ethylene oxide
1,1,1-trichloroethane (TCE-1,1,1)
Tetrachloromethane (TCM)
Polycyclic aromatic hydrocarbons (PAHs)
Chlorofluorocarbons (CFCs)
Toluene
Chromium and compounds (as Cr)
Copper and compounds (as Cu)
Naphthalene
Trichloromethane
Nickel and compounds (as Ni)
Hydrogen cyanide (HCN)
Lead and compounds (as Pb)
Trichloroethylene (TRI)
Tetrachloroethylene
Hydrochlorofluorocarbons (HCFCs)
Vinyl chloride
Zinc and compounds (as Zn)
1,2-dichloroethane (DCE-1,2)
Hydro-fluorocarbons (HFCS)
Dichloromethane (DCM)
Benzene
Fluorine and inorganic compounds (as HF)
Perfluorocarbons (PFCs)
Chlorine and inorganic compounds (as HCl)
Nitrous oxide (N2O)
Particulate matter (PM10)
Ammonia (NH3)
Non-methane volatile organic compounds (NMVOC)
Methane (CH4)
Nitrogen oxides (NOX)
Sulphur oxides (SOX)
Carbon monoxide (CO)
Carbon dioxide (CO2) excluding biomass
Carbon dioxide (CO2)

ESME

# Data Amputation




Le dendrogramme du regroupement hiérarchique des colonnes ainsi que barplot de nullité par colonne

# Data Cleaning I

**UNITE EN KG**

gigagramme
milligramme
kilogramme
gramme

**SECTOR CODE**

+ de 200 valeurs uniques

- Energy production and distribution
- Energy use in industry
- Non road transport
- Commercial, institutional. and households
- Industrial processes and product use
- National total for the entire territory
- Agriculture
- Waste
- Other

**POLLUANTS**

Nombre important de devisions

- CO
- NH3
- NMVOC
- Nox
- PM10
- PM2.5
- Sox
- TSP

ESME

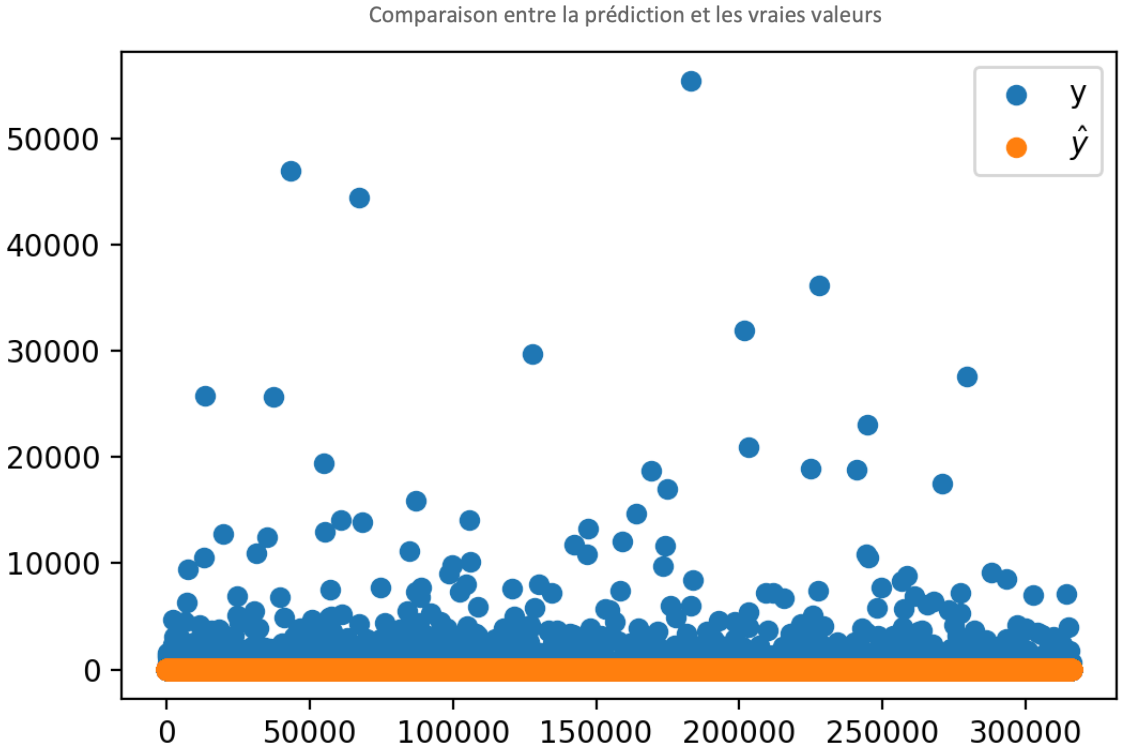
# Data Cleaning II

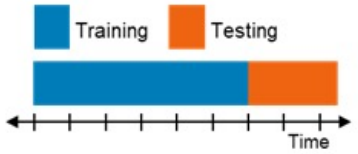
| RangeIndex : 288566 entries | | |
|---|---|---|
| # | Column | Dtype |
| 0 | countryName | object |
| 1 | EPRTRAnnexIMainActivityCode | object |
| 2 | Longitude | float64 |
| 3 | Latitude | float64 |
| 4 | pollutant | object |
| 5 | emissions | float64 |
| 6 | reportingYear | int64 |

- Suppression des colonnes
- Ajout des villes

| RangeIndex : 288566 entries | | |
|---|---|---|
| # | Column | Dtype |
| 0 | countryName | object |
| 1 | EPRTRSectorCode | object |
| 2 | eprtrSectorName | object |
| 3 | EPRTRAnnexIMainActivityCode | object |
| 4 | EPRTRAnnexIMainActivityLabel | object |
| 5 | FacilityInspireID | object |
| 6 | facilityName | object |
| 7 | Longitude | float64 |
| 8 | Latitude | float64 |
| 9 | City | object |
| 10 | targetRelease | object |
| 11 | pollutant | object |
| 12 | emissions | float64 |
| 13 | reportingYear | int64 |

# Fenêtre Glissante

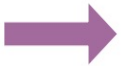



- train [2007] test [2008]
- train [2007, 2008] test [2009]
- train [2007…2019] test [2020]

# Encodage

| country |
|---|
| Austria |
| Belgium |
| Bulgaria |
| Croatia |



| country |
|---|
| 0 |
| 1 |
| 2 |
| 3 |

- Encodage avec OrdinarEncoder() de Sklearn
- Encodage avec One Hot Coder

| color |
|---|
| red |
| red |
| yellow |
| green |
| yellow |

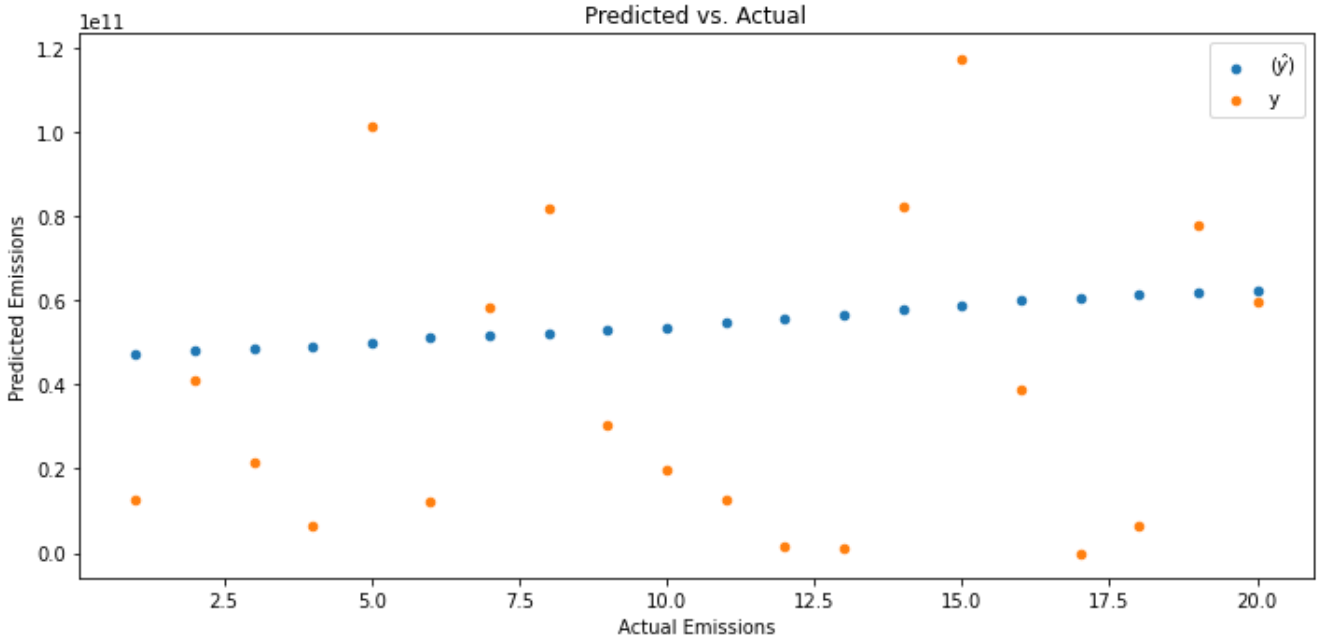
| red | yellow | green |
|---|---|---|
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |

ESME

# Linear Regression

- R^2 = -0.174
- RMSE = 385972242

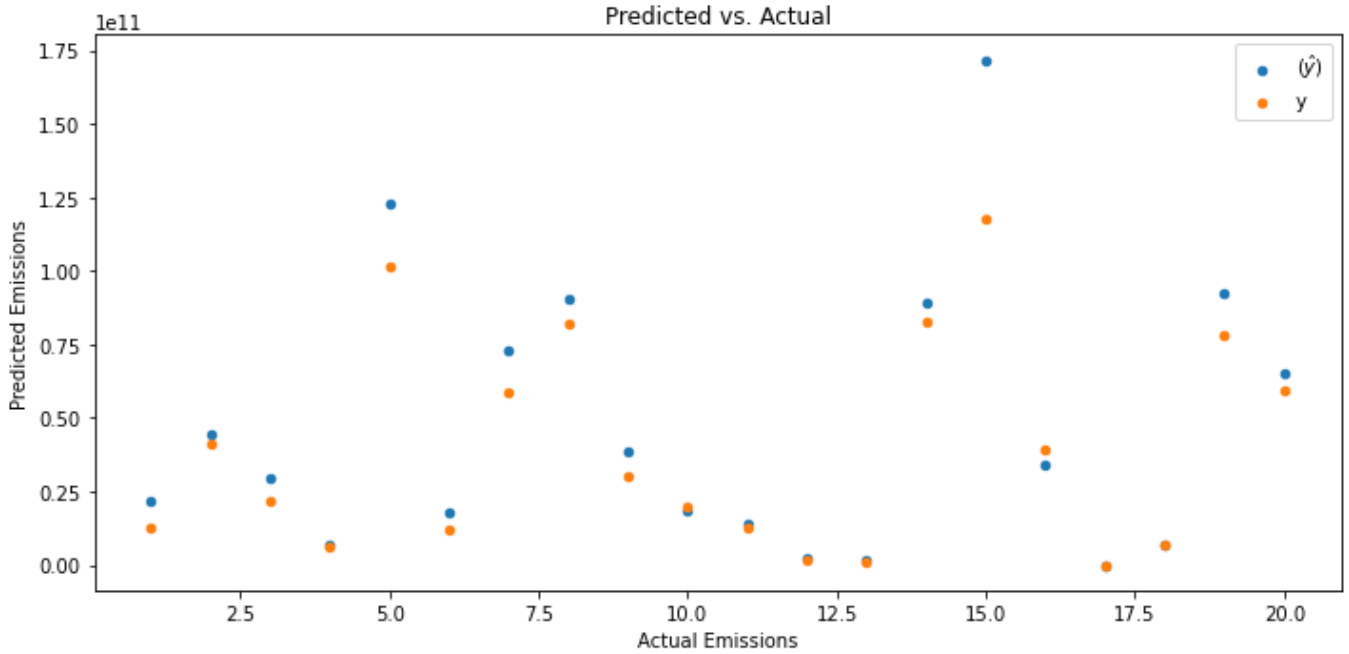
Prediction des émissions pour 2020 pour le CO2



# Decision Tree Regressor

- R^2 = 0.8338
- RMSE = 14540912642

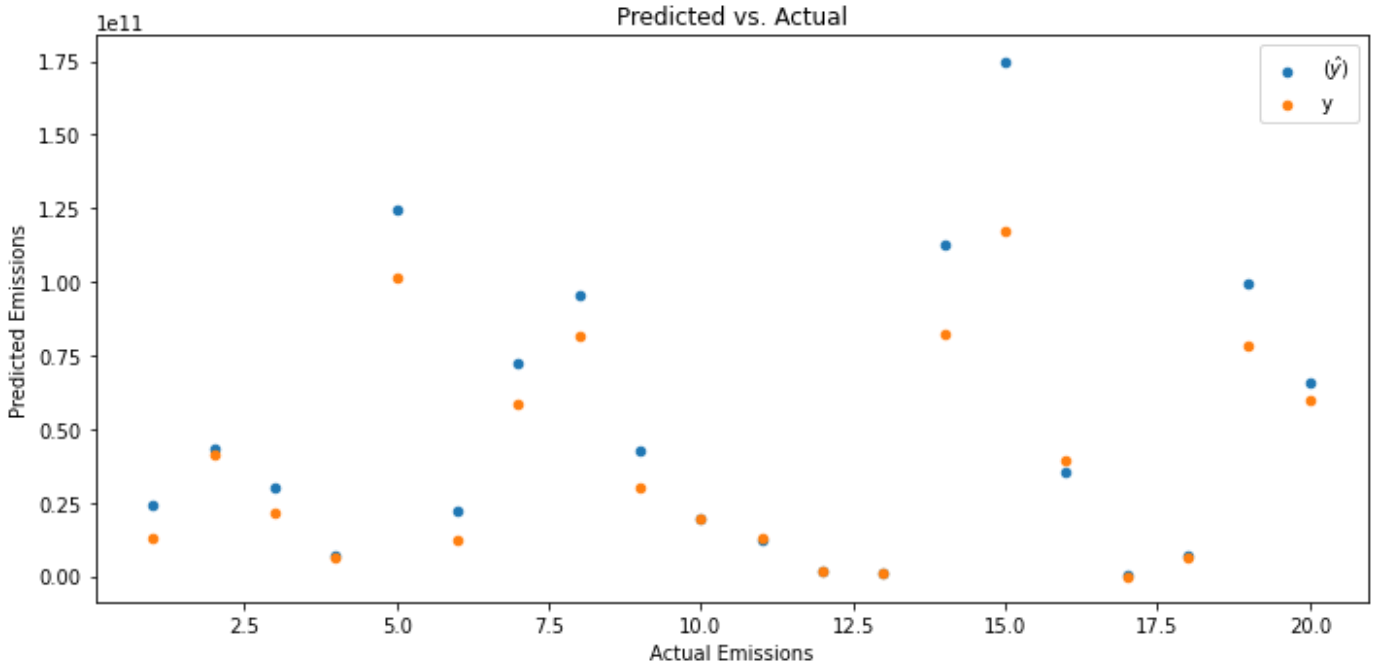
Prediction des émissions pour 2020 pour le CO2 avec random_state=1



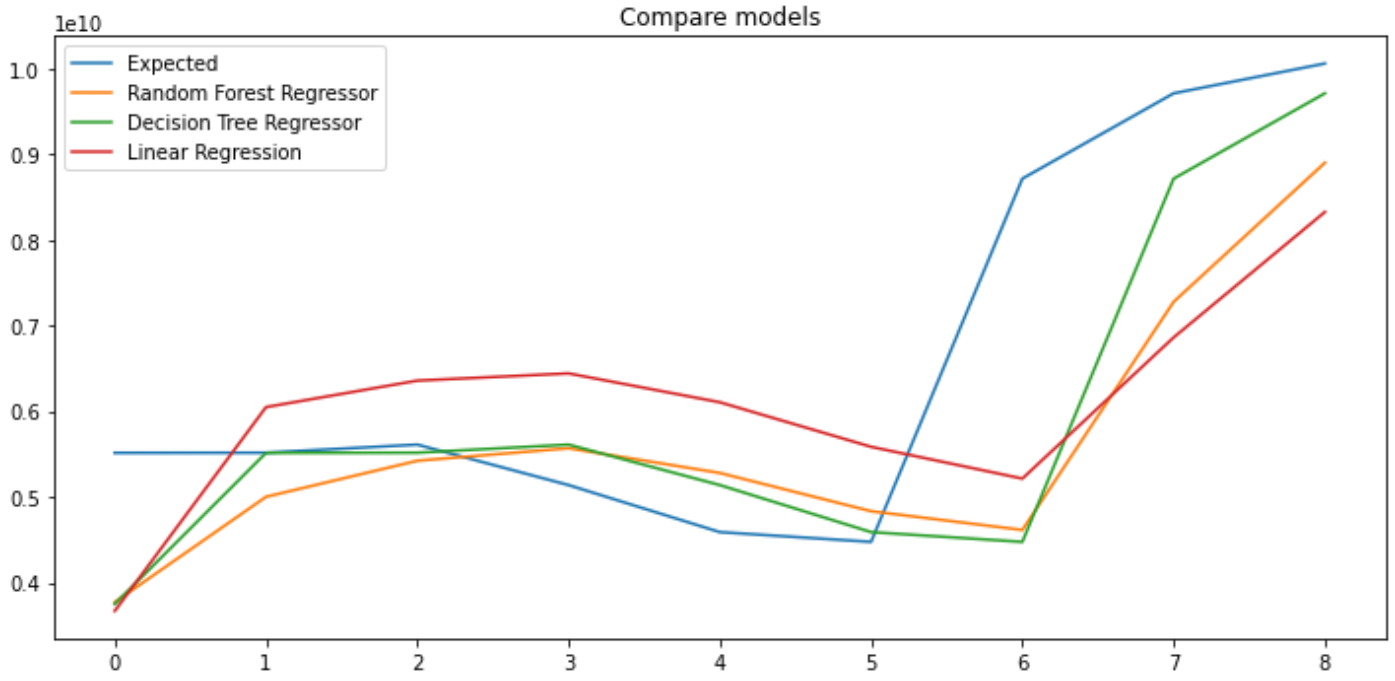
# Random Forest Regressor

- R^2 = 0.7613
- RMSE = 17429321368

Prediction des émissions pour 2020 pour le CO2 avec n_estimatorns =1000, n_jobs=-1, random_state=1

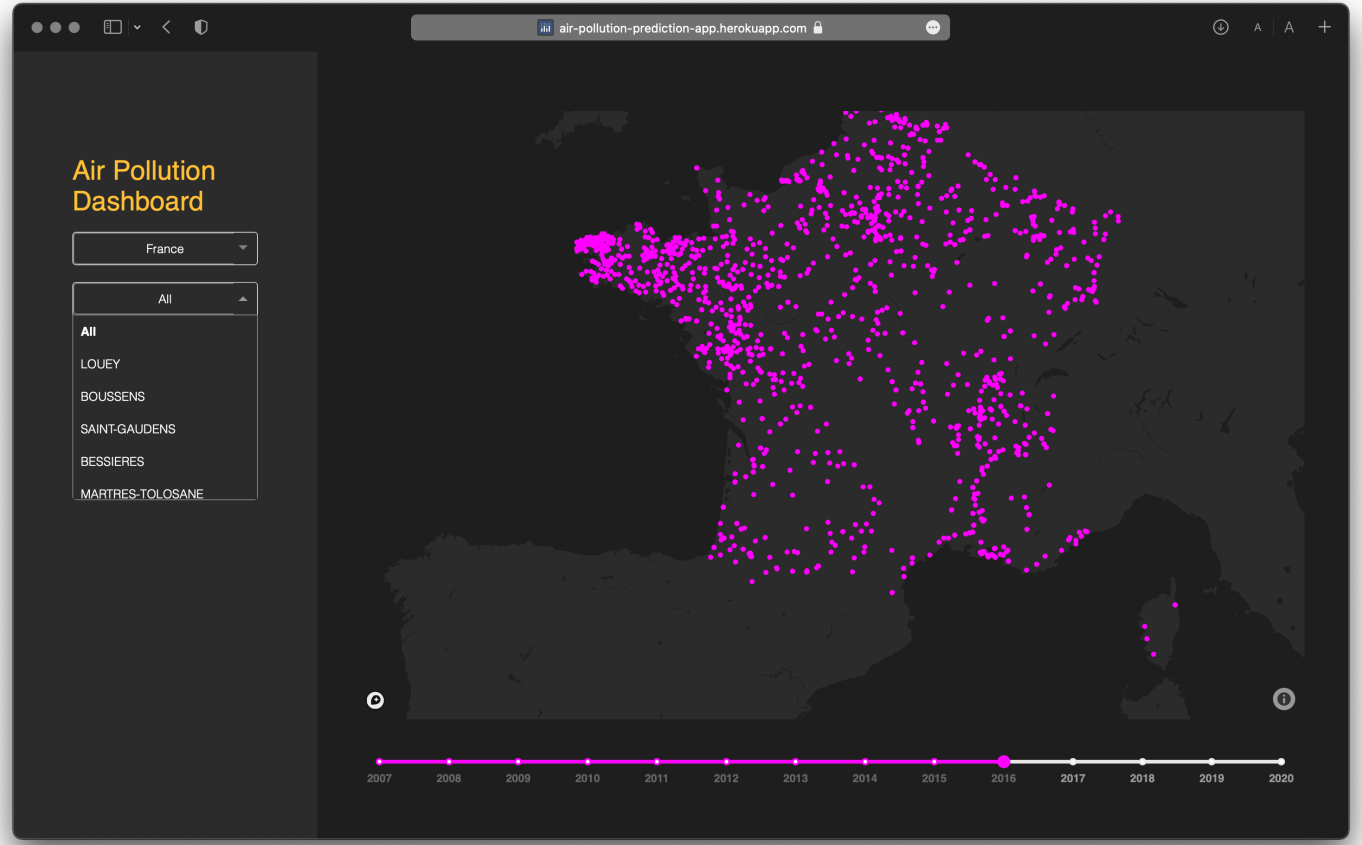


# Comparer trois modèles



ESME

# Dashboard



ESME

# Analyse et Interpretation

- Decision Tree Regressor est plus rapide
- apprend et s'adapte vite
- lien emissions – sector code vue en amputation

```
Results for DecisionTreeRegressor
R² for train: [2007] test: [2008] : 0.997
R² for train: [2007, 2008] test: [2009] : 0.966
R² for train: [2007, 2008, 2009] test: [2010] : 0.983
R² for train: [2007, 2008, 2009, 2010] test: [2011] : 0.997
R² for train: [2007, 2008, 2009, 2010, 2011] test: [2012] : 0.995
R² for train: [2007, 2008, 2009, 2010, 2011, 2012] test: [2013] : 0.994
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013] test: [2014] : 0.989
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014] test: [2015] : 0.995
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015] test: [2016] : 0.992
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016] test: [2017] : 0.987
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017] test: [2018] : 0.988
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018] test: [2019] : 0.930
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019] test: [2020] : 0.834
```

- combine plusieurs abres
- appretissage lent
- adaptation lente aux changments

```
Results for RandomForestRegressor
R² for train: [2007] test: [2008] : 0.826
R² for train: [2007, 2008] test: [2009] : 0.918
R² for train: [2007, 2008, 2009] test: [2010] : 0.952
R² for train: [2007, 2008, 2009, 2010] test: [2011] : 0.990
R² for train: [2007, 2008, 2009, 2010, 2011] test: [2012] : 0.990
R² for train: [2007, 2008, 2009, 2010, 2011, 2012] test: [2013] : 0.995
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013] test: [2014] : 0.987
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014] test: [2015] : 0.991
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015] test: [2016] : 0.987
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016] test: [2017] : 0.985
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017] test: [2018] : 0.987
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018] test: [2019] : 0.921
R² for train: [2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019] test: [2020] : 0.761
```

# Conclusion et Perspectives

**AJOUTER DES DONNEES**

Remplissage des données, recherche des nouvelles

**RECHERCHE DE CAUSALITE**

Utilisation des statistique pour prédiction avancée

**ACCES AUX RESULTATS**

Mie à jour des analyses, amélioration de l'application


Merci
pour votre attention!