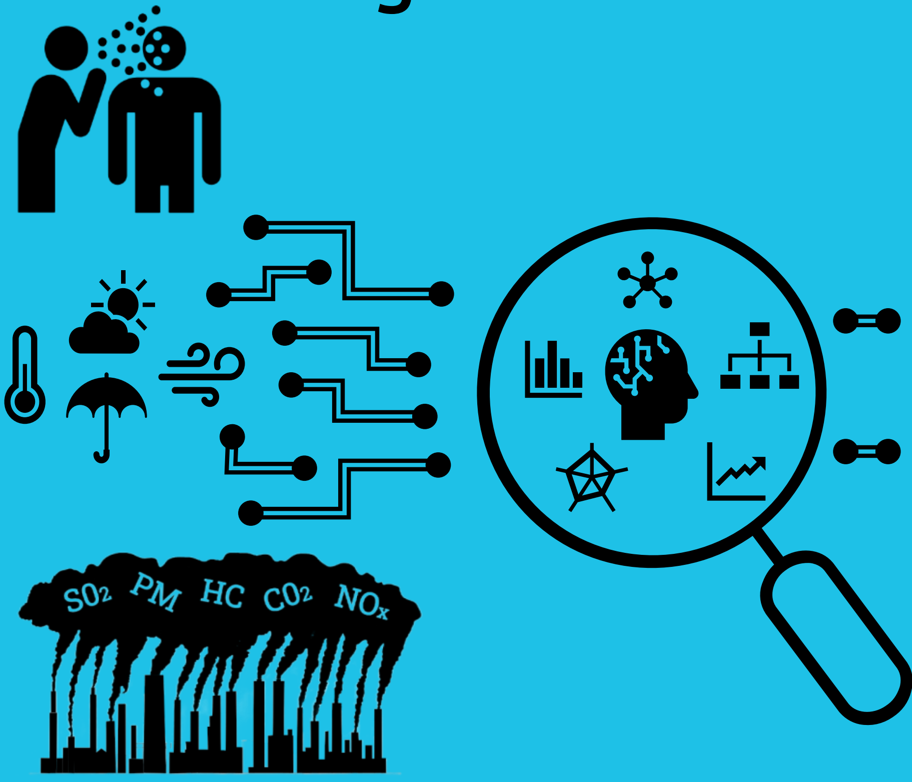
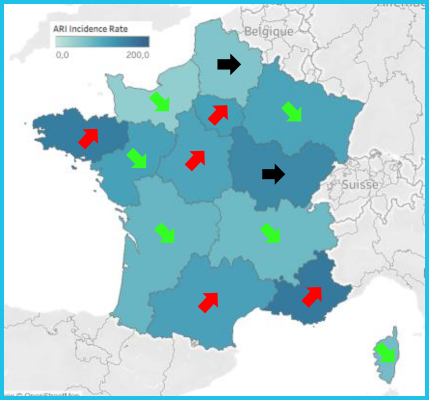
Acute Respiratory Infections Forecast using Machine Learning





Ironhack Data Analyst certification project

Romain Courtois

1. Case Study

Acute **R**espiratory **I**nfections :

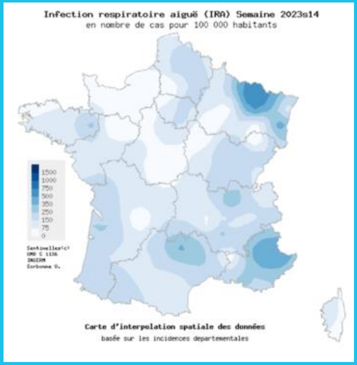
Caused by various respiratory viruses including SARS-CoV-2

Study focused on factors of the disease: pollution, climate and seasons.

Using **ETL**, **EDA** and **ML**

→ build a model that can predict future **incidence rate**

2. Data Sources


Infection respiratoire aiguë (IRA) Semaine 2023s14
en nombre de cas pour 100 000 habitants
1500
1000
750
500
350
250
150
75
0
Sentinelles(c)
UMR S 1136
INSERM
Sorbonne U.
Carte d'interpolation spatiale des données
basée sur les incidences departementales

A. French epidemiological surveillance:

Weekly ARI incidence rate by regions

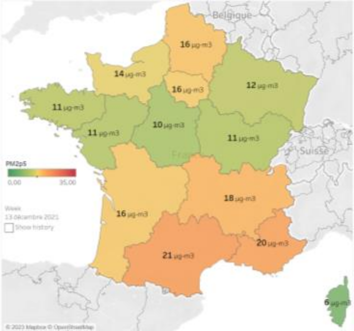
opendatasoft

B. Historical meteorological observation France:

Daily temp., pressure, humidity…

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

data.gouv.fr



C. Concentrations of air pollutants

O3, NOx, SO2, PM10, PM2.5, CO, CH6H6

opendatasoft

3. Extract Transform Load

Extract

API → aggregated data

.csv → raw data

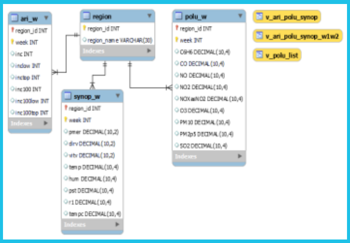
Transform

Python :

aggregate data by weeks

normalize categories/regions

Load

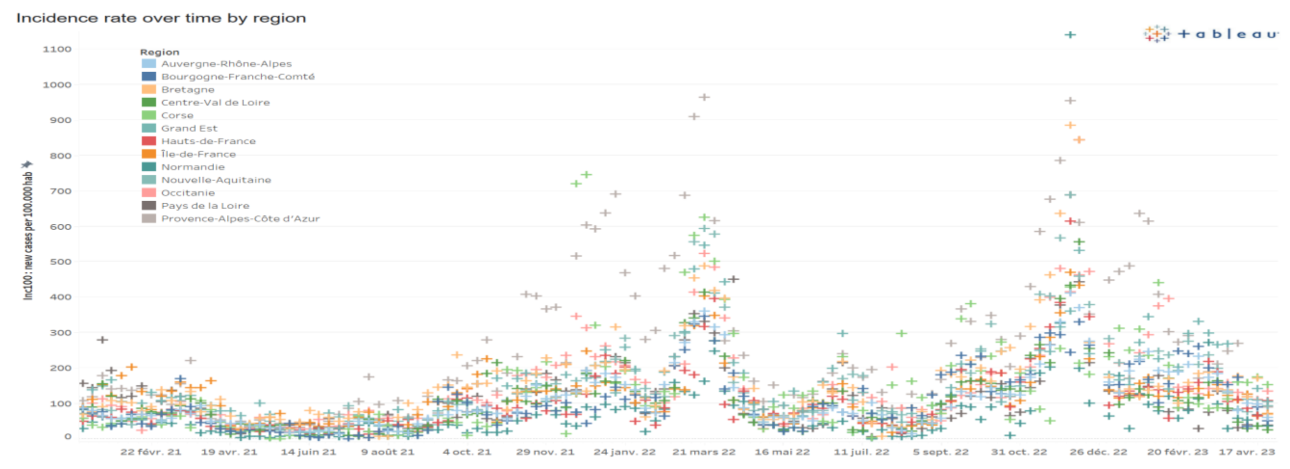


MySQL relational database

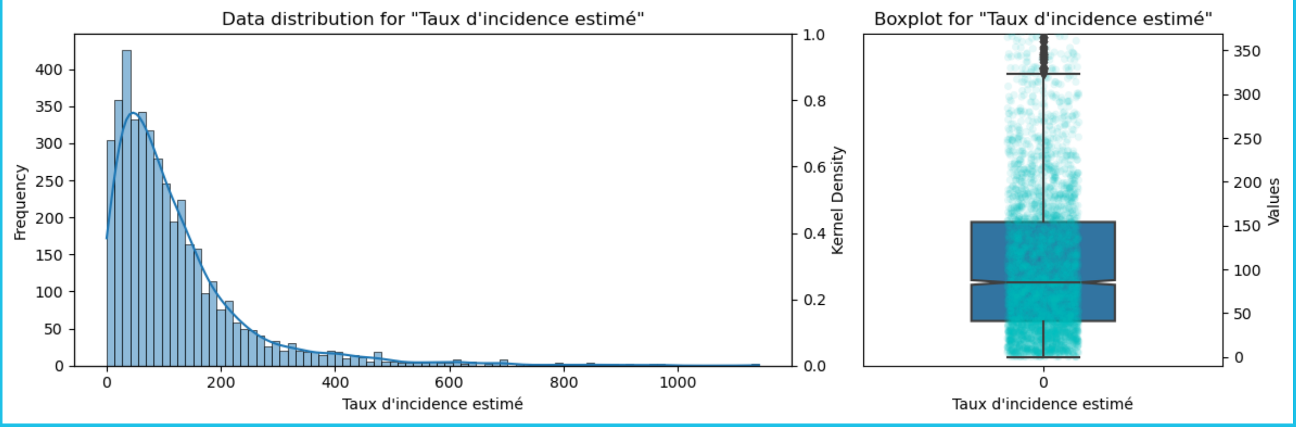
- Staging area
- DataMart

4. Exploratory Data Analysis

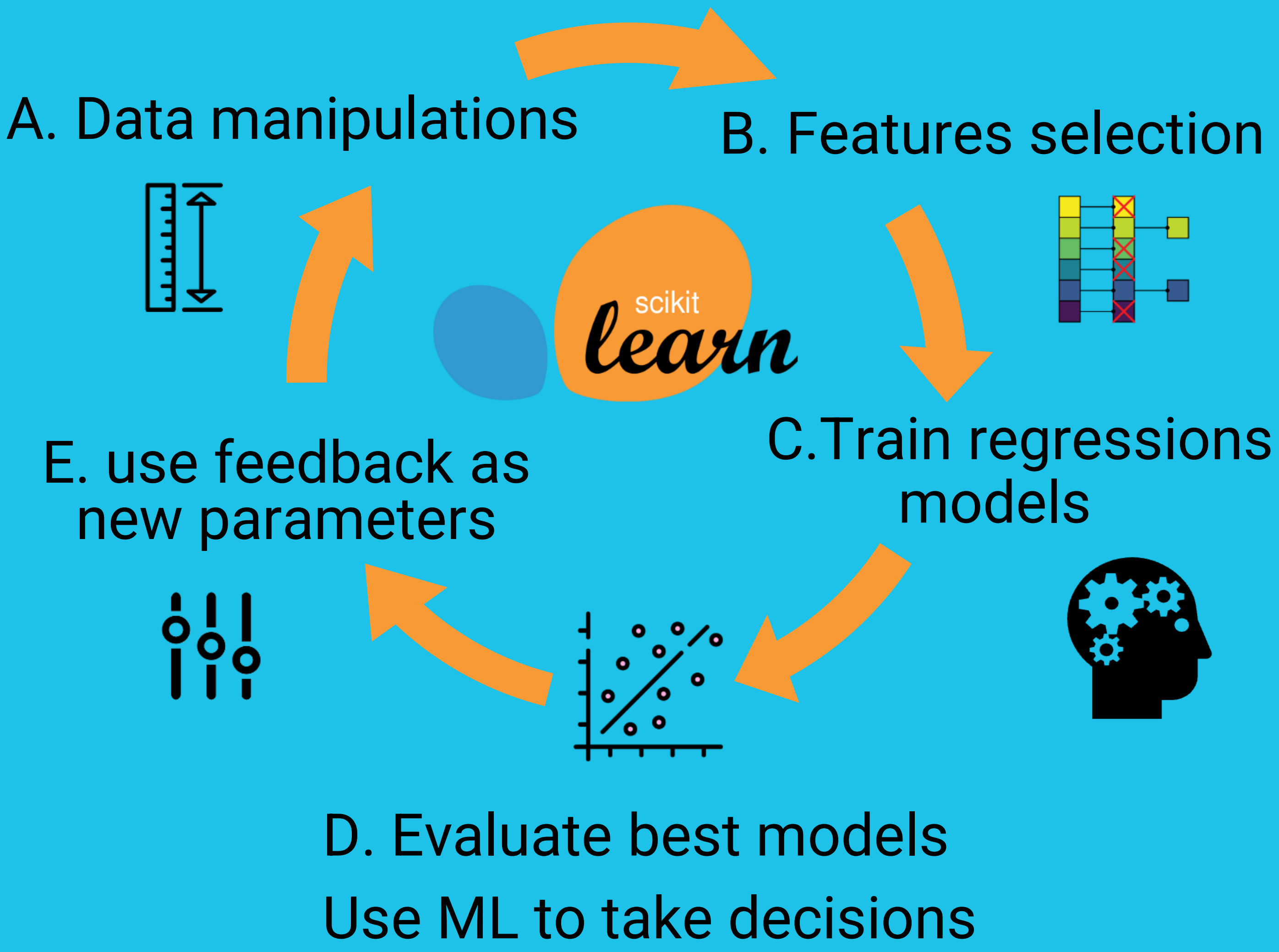
Incidence rate over time



Data distribution



5. Iterative Machine Learning



A. Data manipulations

B. Features selection

C.Train regressions models

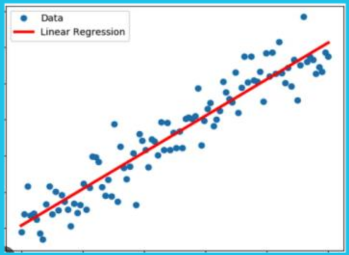
D. Evaluate best models

Use ML to take decisions

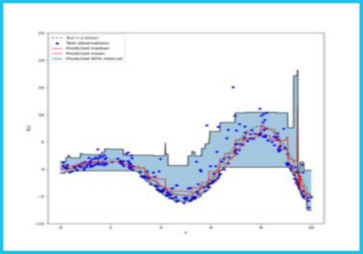
E. use feedback as new parameters

6. Regression models used

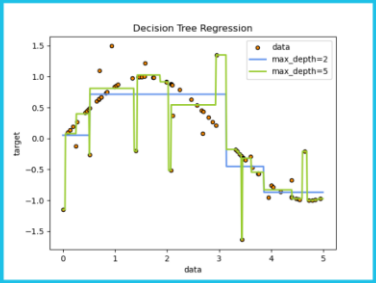
- Linear Regression



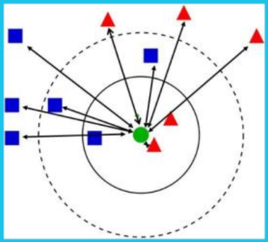
- Gradient Boosting



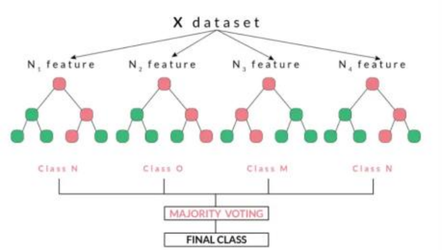
- Decision Tree



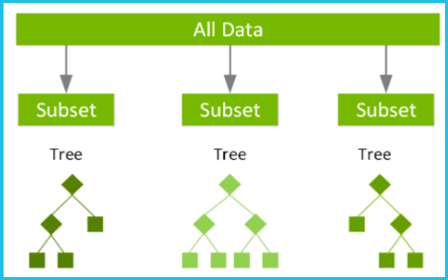
- K Neighbors



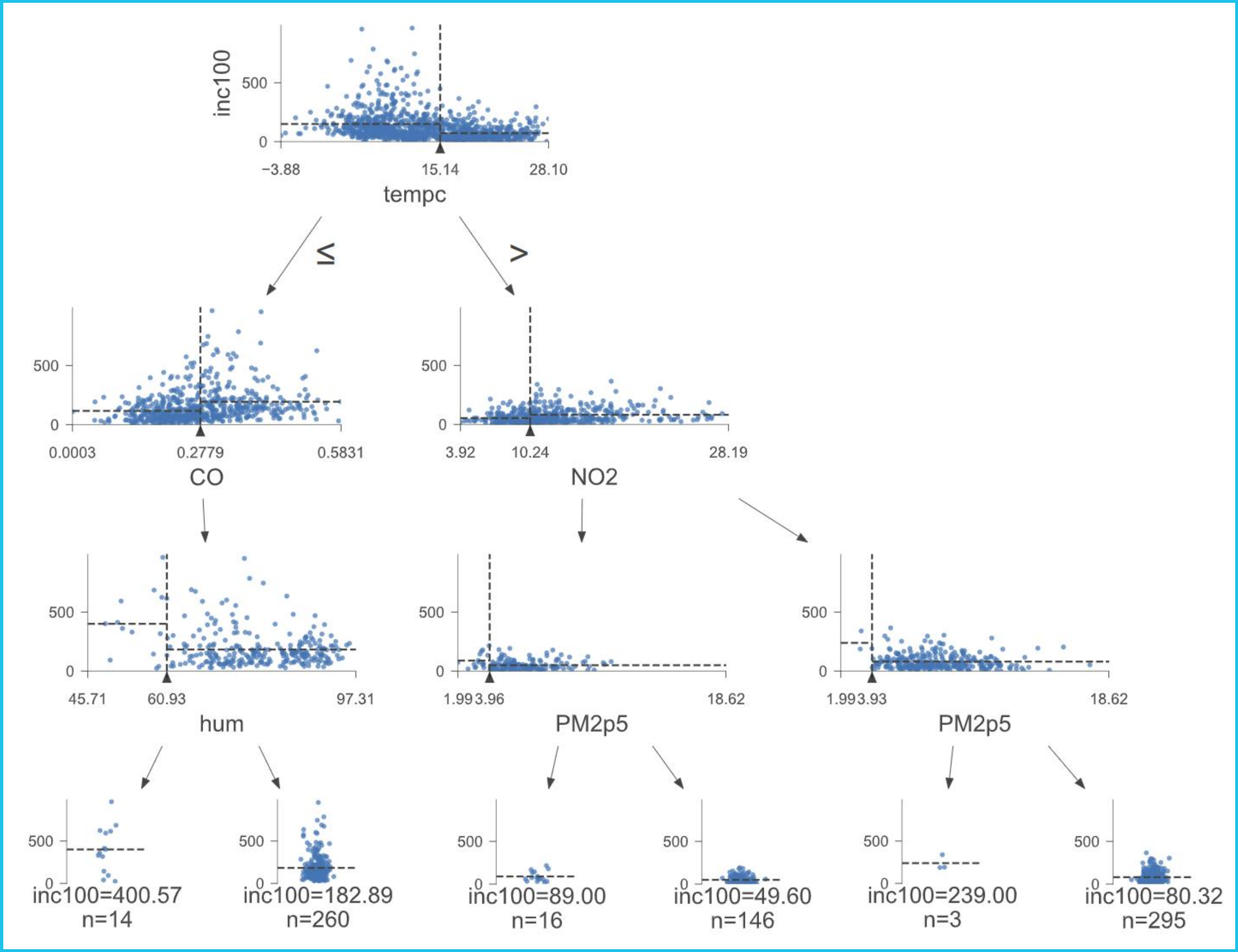
- Random Forest



- XGBoost



7. Focus on XGBoost tree regression



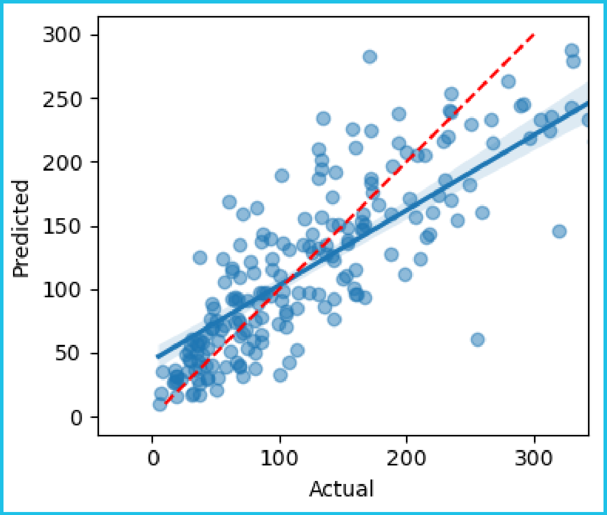
CO and humidity is used when temp < 15°c

NO2 and PM2.5 when >15°c

8. Conclusion

Best prediction score : Gradient Boosting

37% (mape)



With more data, it improves over time.

I will publish the prediction dashboard when score is < 10%.

Follow me on :

github.com/romaincrt/ARIForecast