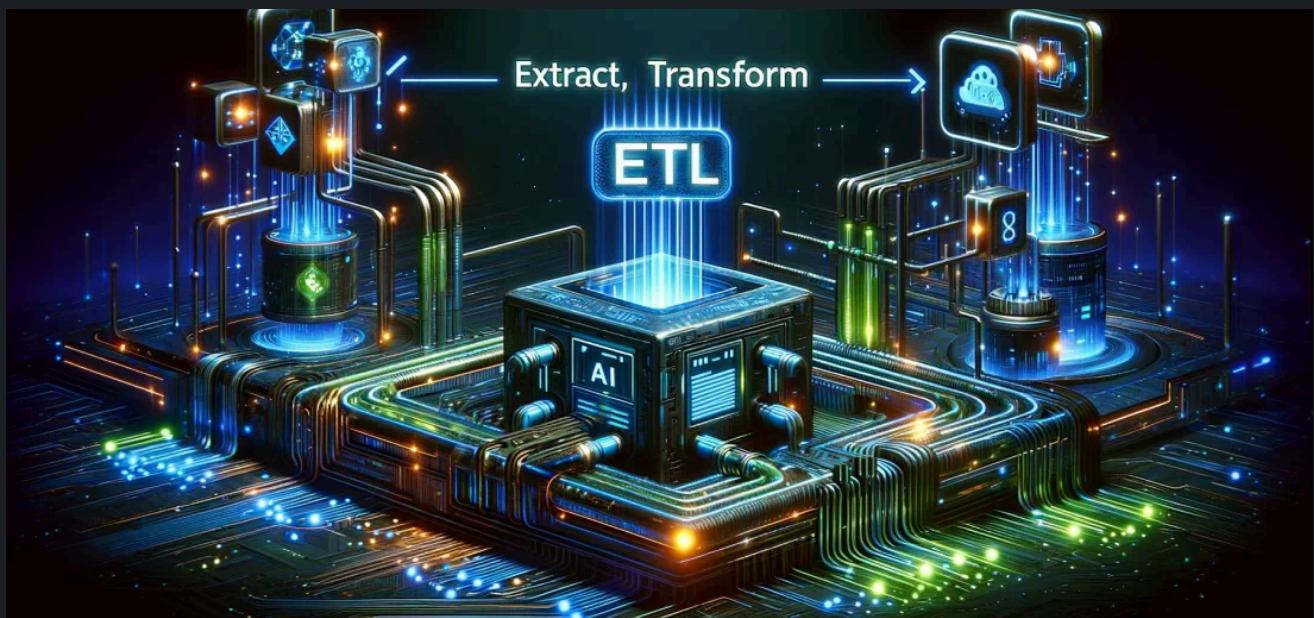




# MSPR EISI BLOC 3

I1 DEV 2 GROUPE 2



SANGMIN SHIM  
JULIEN FLUSIN  
ANTHONY DURET  
QUENTIN LAUNAY

# SOMMAIRE

<b>1) PRÉSENTATION DU PROJET</b>	<b>1</b>
◦ PYTHON	
◦ POSTGRESQL	
◦ POWER BI	
◦ DOCKER	
<b>2) CHOIX DES DONNÉES ET JUSTIFICATION</b>	<b>6</b>
◦ RECHERCHE DES DONNÉES	
◦ SOURCE DES DONNÉES	
◦ RÉCOLTE DES DONNÉES	
◦ NETTOYAGE DES DONNÉES	
◦ TRAITEMENT DES DONNÉES	
◦ STOCKAGE DES DONNÉES	
◦ VISUALISATION DES DONNÉES	
<b>3) UTILISATIONS DES DONNÉES</b>	<b>11</b>
◦ RECHERCHE D'UN MODÈLE PRÉDICTIF	
◦ VISUALISATION DES DONNÉES	
<b>4) CONCLUSION</b>	<b>14</b>
<b>ANNEXE</b>	

# 1) PRÉSENTATION DU PROJET

Le projet de cette MSPR porte sur l'étude de données publiques françaises dans la France entière sur différents facteurs, afin de déterminer des résultats d'élections présidentielles possible en se basant sur différents critères comme :

- L'emploi
- La sécurité
- La population

Grâce à ces différents facteurs, il est possible d'envisager un rapport avec les élections politiques des dernières années et donc d'envisager les résultats des suivants.

C'est dans ce cadre que nous allons récupérer les données et les utiliser afin de pouvoir créer des statistiques et des résultats permettant de comprendre selon quel changement dans notre pays nous pourrons voir les résultats des élections changées.

Afin de préparer ce projet et de le mener à son terme nous avons utilisé plusieurs technologies :

# (1) PYTHON

AIRFLOW : Apache Airflow est une plate-forme de gestion de flux de travail open source. Airflow est écrit en Python et les workflows sont créés via des scripts Python. Airflow est conçu selon le principe de la "configuration as code".

PANDAS : Pandas est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.

FOLIUM : Folium est une bibliothèque pour faciliter la visualisation de données manipulées en Python sur une carte interactive Leaflet. Il permet à la fois l'association de données à une carte pour des visualisations choroplèthes et le passage de visualisations vectorielles/raster/HTML riches en tant que marqueurs sur la carte.

SIMPLEDBF : Simpledbf est une bibliothèque Python pour convertir des fichiers DBF de base (voir Limitations) en fichiers CSV, Pandas DataFrames, tables SQL ou tables HDF5.

## (2) POSTGRESQL

Automatiser les tâches pour extraire et transformer les données via AIRFLOW nous permet de charger les données depuis le site du gouvernement dans notre base de données. Grâce aux bibliothèques de Python, nous avons réussi à sauvegarder les données en vue de leur visualisation.

The screenshot shows the pgAdmin 4 interface connected to a PostgreSQL database named 'Social\_postgres'. The left sidebar displays the database structure, including 'Databases' (postgres, social\_data), 'Schemas' (public), and various system catalogs and configurations. The main area shows a query window with the following SQL code:

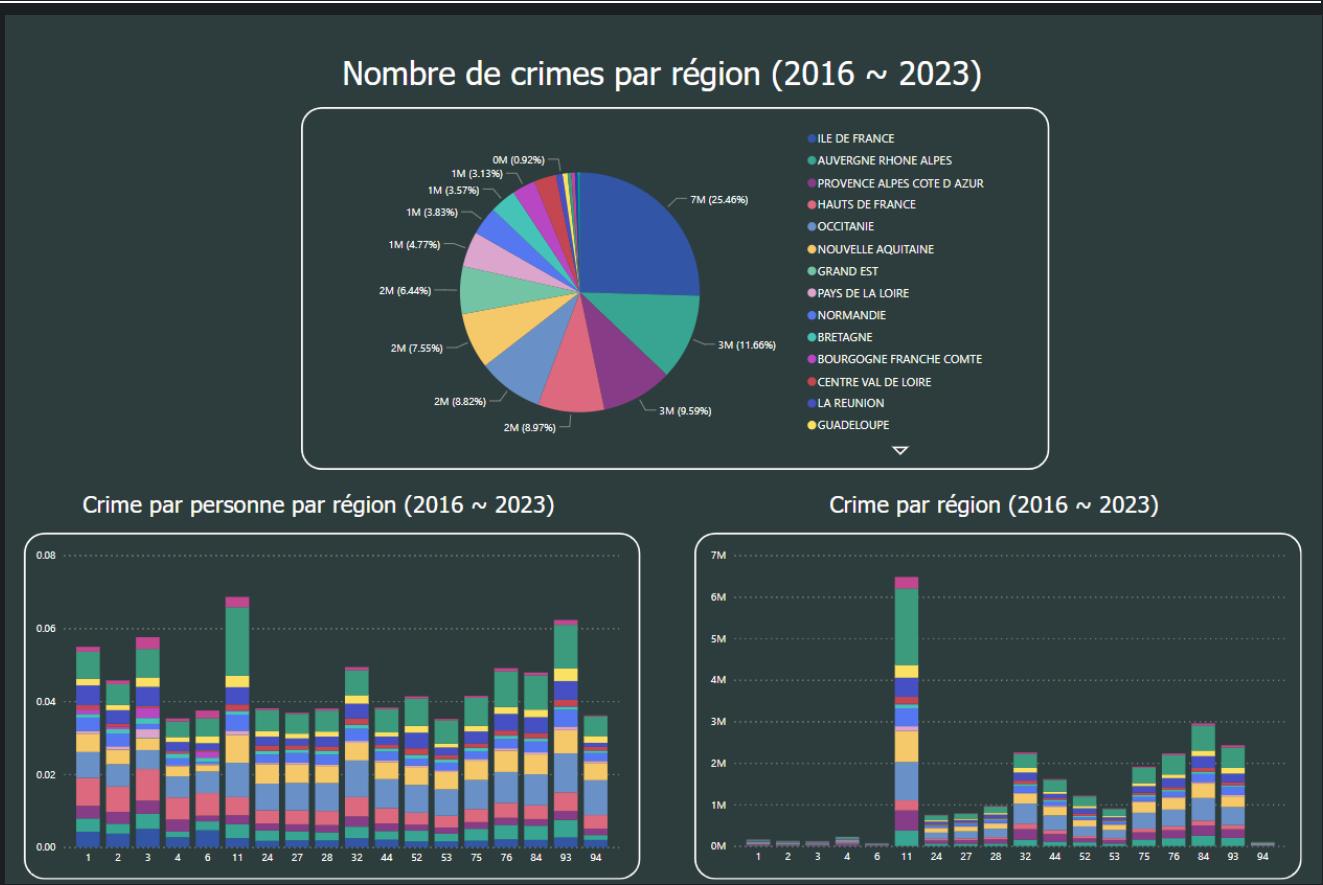
```
1 SELECT adec, mdec, depdec, lieudecr, sexe, anais, depnais, activ, indnat, depdom, tudom, tucom, etamat
2 FROM public.death;
```

Below the query window is a 'Data Output' tab showing the results of the query. The results are as follows:

	adec bigint	mdec bigint	depdec text	lieudecr text	sexe bigint	anais bigint	depnais text	activ bigint	indnat bigint	depdom text	tudom bigint	tucom text	etamat bigint
1	2017	12 01	AUT	1	1968	99		2	2	99	9 [null]		2
2	2017	5 01	HMR	2	1940	99		1	2	99	9 [null]		4
3	2017	12 01	AUT	2	1928	99		1	1	99	9 [null]		3
4	2017	12 01	AUT	1	1945	99		1	2	99	9 [null]		4
5	2017	7 01	AUT	1	1959	99		3	2	99	9 [null]		1
6	2017	1 01	AUT	1	1949	99		1	2	99	9 [null]		1
7	2017	6 01	AUT	1	1958	99		3	2	99	9 [null]		2
8	2017	7 01	AUT	1	1971	99		2	2	99	9 [null]		2
9	2017	2 01	HMR	2	1927	99		1	2	99	9 [null]		3
10	2017	7 01	AUT	2	1951	99		1	2	99	9 [null]		2
11	2017	5 01	LOG	1	1921	99		1	2	99	9 [null]		2
12	2017	9 01	AUT	1	1958	99		2	2	99	9 [null]		4
13	2017	1 01	EHC	1	1955	99		3	2	99	9 [null]		2
14	2017	10 01	EHC	2	1937	99		1	1	99	9 [null]		3
15	2017	2 02	AUT	2	1947	99		1	2	99	9 [null]		2
16	2017	1 02	AUT	1	1951	99		2	2	99	9 [null]		2
17	2017	5 02	AUT	1	1949	99		1	2	99	9 [null]		2
18	2017	3 02	EHC	2	1920	99		1	2	99	9 [null]		2

### (3) POWER BI

POWER BI permet de traiter et de visualiser les données de manière variée, notamment les crimes par région, les naissances, les décès et les DPAE (Déclarations Préalables à l'Embauche), entre autres. Grâce à POWER BI, nous avons réussi à créer un tableau de bord sur lequel ces données sont présentées de manière claire et accessible.



## (4) DOCKER

Avec Docker, nous avons réussi à créer des conteneurs qui exécutent les fonctions essentielles d'AIRFLOW ainsi que des bases de données pour stocker les données après le processus de tâche.

Name	Image	Status	CPU (%)	Port(s)	Last started	Actions
<input checked="" type="checkbox"/> airflow		Running (7/8)	6.08%		30 minutes ago	
<input type="checkbox"/> airflow-worker-1 8980524d148d	<a href="#">airflow-airflow-worker</a>	Running	0.01%		30 minutes ago	
<input type="checkbox"/> airflow-webserver-1 919c2e3615d8	<a href="#">airflow-airflow-webserver</a>	Running	0.07%	<a href="#">8080:8080</a>	30 minutes ago	
<input type="checkbox"/> airflow-triggerer-1 47118f6be3e2	<a href="#">airflow-airflow-triggerer</a>	Running	0.83%		30 minutes ago	
<input type="checkbox"/> airflow-scheduler-1 bbd69584cd8d	<a href="#">airflow-airflow-scheduler</a>	Running	1.47%		30 minutes ago	
<input type="checkbox"/> airflow-init-1 be6f35864ba4	<a href="#">airflow-airflow-init</a>	Exited	0%		30 minutes ago	
<input type="checkbox"/> redis-1 816b593dd24f	<a href="#">redis:latest</a>	Running	0.16%		30 minutes ago	
<input type="checkbox"/> social_postgres-1 5caf1c059d5c	<a href="#">postgres:15.3</a>	Running	2.66%	<a href="#">5432:5432</a>	30 minutes ago	
<input type="checkbox"/> postgres-1 c8012c5537c4	<a href="#">postgres:13</a>	Running	0.88%		30 minutes ago	

## 2) CHOIX DES DONNÉES ET JUSTIFICATIONS

### (1) RECHERCHE DES DONNÉES

Grâce au document donné nous avions quelques sources de données proposées, mais avant de les récolter et d'analyser les sources possibles nous avions pour but de décider à l'avance de quelle type de données nous pourrions avoir besoin. Notre but étant de prédire les possibles résultats des élections présidentielle nous avons donc cherché principalement des données ou les élections ont eu lieu, puis nous avons cherché les données relatives au paysage français et qui pourrait influencer leur vote, sécurité, emploi, naissances / décès (population) ...

### (2) SOURCE DES DONNÉES

Pour notre source de données nous nous sommes servis de différents tableaux de données disponibles sur les sites du gouvernement (voir annexe). Nous avons donc pu récolter les données de plusieurs années où ont eu lieu des élections (2012,2017, 2022). Nous nous sommes basés sur la France entière pour avoir la plus grande base de données possible.

## (3) RÉCOLTE DES DONNÉES

Pour récolter nos données nous avons donc cherché sur les sites du gouvernement différents jeux de données au format CSV dans les meilleurs cas, sinon au format dBASE (.dbf) dans les autres cas, nous avons ensuite analysé ce qu'ils contenaient afin de s'assurer que les données soit consistante.

Les données étant récupérées par le gouvernement français nous avons considéré que la base des données était de qualité, il fallait donc que nous gardions cette qualité.

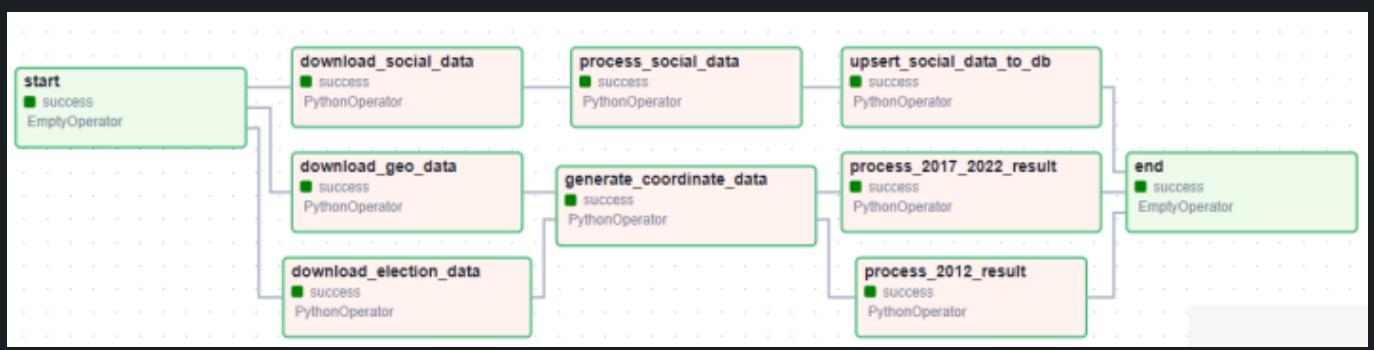
De plus, nos données ne contiennent aucune donnée sensible, et sont anonymes, dans le cadre du RGPD nous nous devons juste d'indiquer ce à quoi servent nos données.

## (4) NETTOYAGE DES DONNÉES

Afin de nettoyer les données, nous avons utilisé la bibliothèque pandas, qui nous permet d'analyser et de nettoyer facilement les données telles que les élections présidentielles, les chiffres du crime, etc. Avant de stocker ou de visualiser les données, cette méthode est primordiale pour garantir et améliorer la qualité des données.

## (5) TRAITEMENT DES DONNÉES

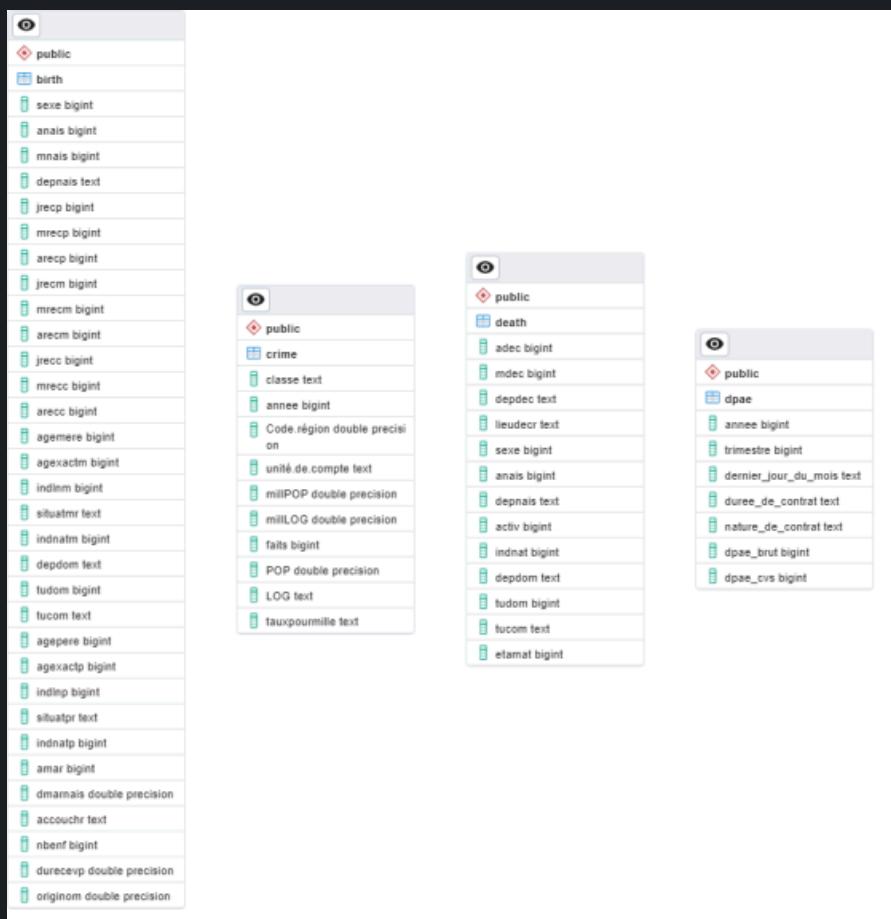
Comme le processus ETL, avec AIRFLOW, nous avons réussi à automatiser toutes les tâches de données juste avant la visualisation. Cela comprend le téléchargement des données via des requêtes HTTP, le nettoyage des données, la fusion de deux ensembles de données différents, la production des fichiers HTML affichant les résultats des élections présidentielles en 2012, 2017 et 2022, ainsi que le stockage des données sociales qui seront visualisées via POWER BI.



## (6) STOCKAGE DES DONNÉES

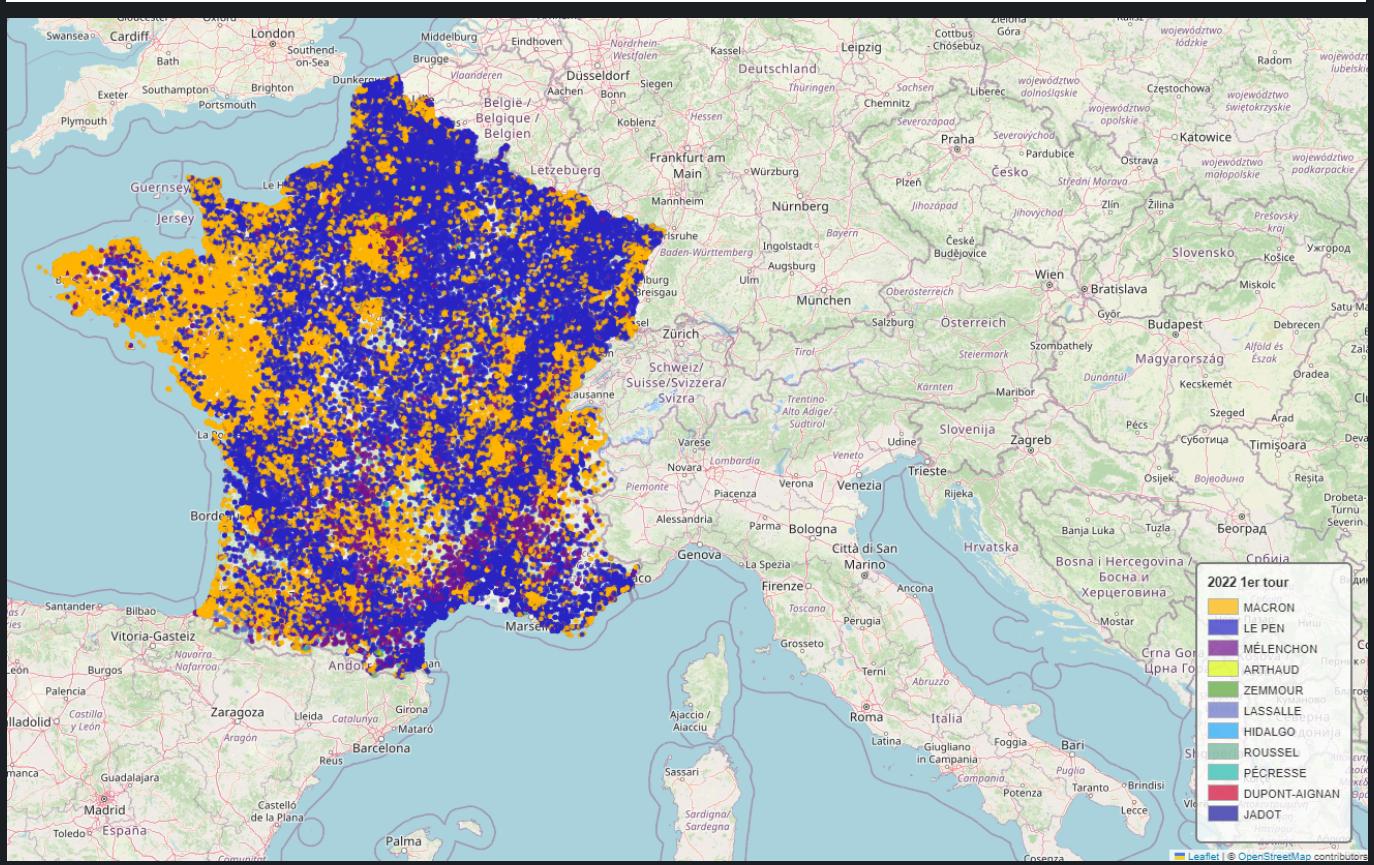
Les données sont stockées dans notre base de données, créée à partir de l'image PostgreSQL de Docker, et comprennent des données sociales telles que les naissances, les décès, les crimes et les DPAE. Nous n'avions pas besoin de relation entre nos tables, il n'y a donc pas de MCD ou de MLD.

Voici l'ERD de notre base de données :



## (7) VISUALISATION DES DONNÉES

Selon les données que nous avons recueillies sous différents formats (CSV, DBF), cela nous a permis de visualiser les résultats des élections de 2012, 2017 et 2022 pour les deux tours. De plus, grâce à Power BI et à nos affichages qui seront ci-dessous, nous avons pu générer des graphiques concernant les crimes dans différentes régions, les naissances et les décès dans ces mêmes régions, ainsi que les données DPAE.



# 3) UTILISATION DES DONNÉES

## (1) RECHERCHE D'UN MODÈLE PRÉDICTIF

Afin de compléter les données manquantes nous avons voulu créer un modèle de prédiction linéaire et régressif, le but étant de pouvoir en utilisant les données que nous avions récupéré qui n'étaient pas toujours complète, réussir à déterminer les résultats des élections précédents mais aussi essayer de prédire les résultats du premier et du second tour des élections de 2027.

Pour cela nous nous sommes d'abord documentés sur les méthodes pour créer un modèle prédictif de données. en regardant différents cours, des questions sur différents forum ou même à des modèles d'IA en ligne.

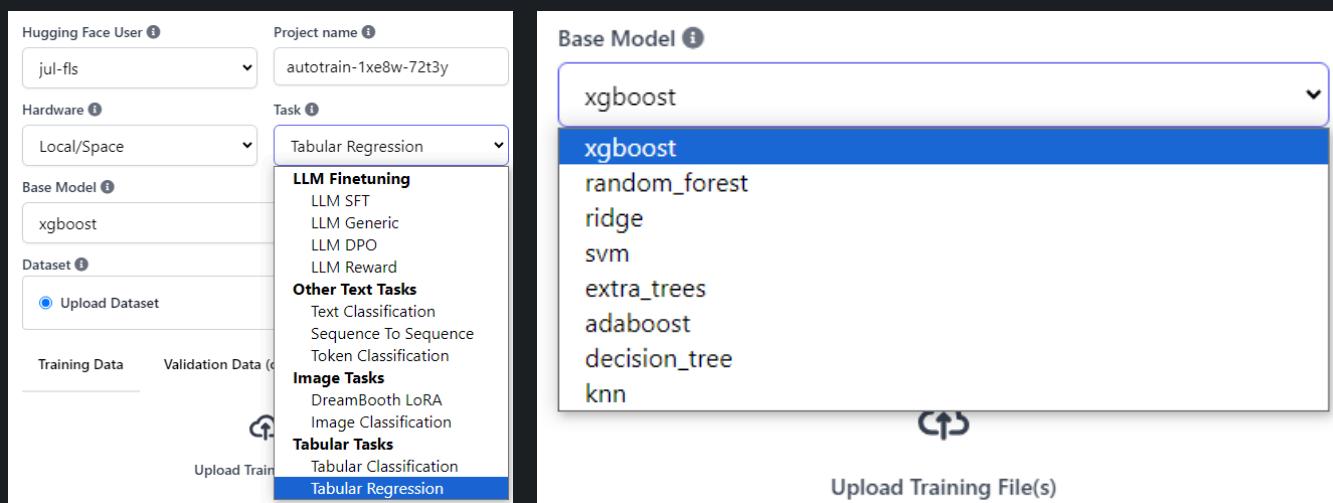
Le but était de faire de l'apprentissage supervisée, donc de créer un algorithme que nous devons entraîner sur un jeu de données choisi afin qu'il puisse s'adapter à de nouvelles données une fois des jeux de test ont été passé et jugé acceptable

Devant notre manque de réussite nous nous sommes retournés vers le site Hugging Face, une immense communauté de chercheur et de développeur spécialisé dans l'IA et le développement et la création de modèle de données et de set de données pour les entraîner.

Nous avons étudié ce site et essayer de créer un modèle sur ce dernier mais devant notre manque de connaissance malgré un résultat que nous avons pu obtenir nous avons décidé de nous occuper de la visualisation et d'assurer des données de qualités afin d'expliciter bien mieux les possibilités de nos données et leur qualités.

Notre projet visait à prédire l'issue des élections présidentielles en France à l'aide d'analyses de régression sur des données sociétales. Nous avons expérimenté divers modèles de machine learning, comme XGBoost et Random Forest, pour traiter nos données en CSV. Malgré de nombreuses tentatives et ajustements de modèles, les résultats concluants nous échappent encore. Cependant, chaque essai nous apporte de nouvelles connaissances et renforce notre détermination à trouver la solution. Notre persévérance illustre notre engagement dans la recherche et l'apprentissage, des qualités que nous valorisons autant que les résultats.

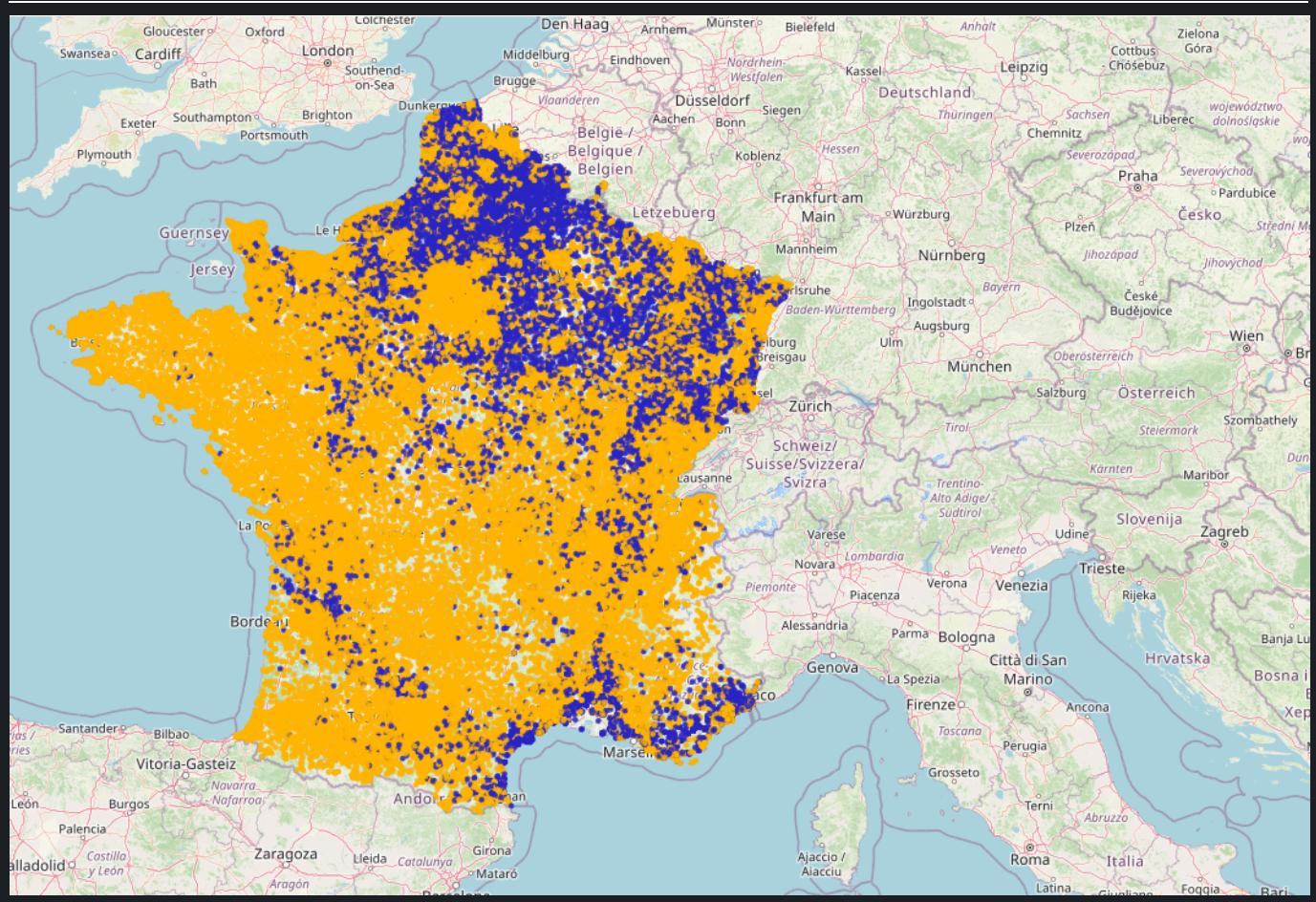
Nous n'avons donc pas d'accuracy, c'est-à-dire que comme nous n'avons aucune donnée prédite, nous ne pouvons pas savoir combien de ces dites données prédites sont correctes et donc déterminer l'accuracy de notre modèle ainsi, nous ne savons pas quelle donnée de notre base n'est pas plus ou moins corrélée avec le résultat direct des élections.



## (2) VISUALISATION DES DONNÉES

Pour la visualisation des données nous utilisons POWER BI connecté à la base de données que nous avons créé plutôt afin de stocker les données.

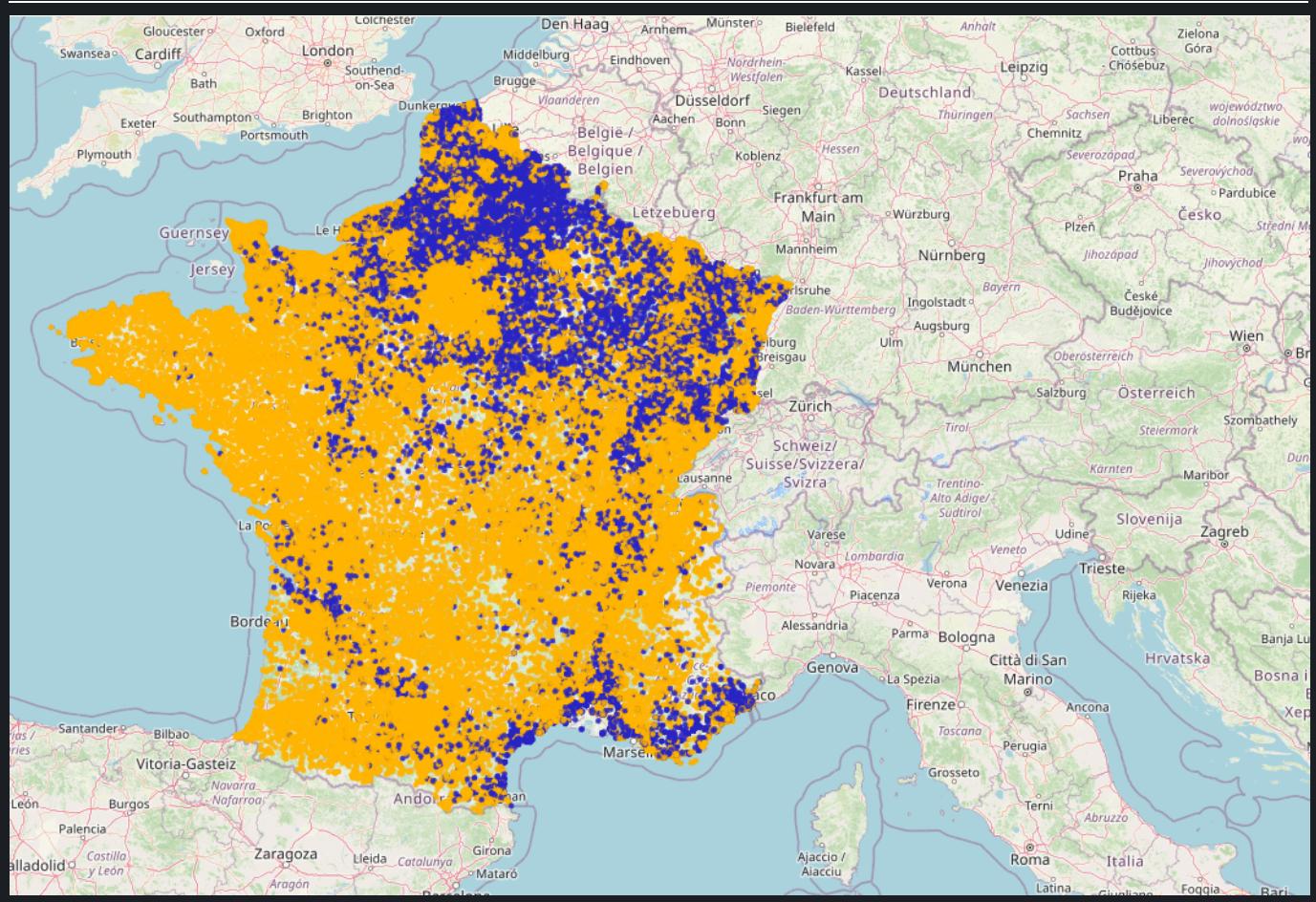
Nous utilisons aussi grâce à Folium et Branca des méthodes afin de pouvoir voir la carte de la France avec tous les votes enregistrés, une couleur étant associée à chaque partie pour lequel les gens ont voté.



## (2) VISUALISATION DES DONNÉES

Pour la visualisation des données nous utilisons POWER BI connecté à la base de données que nous avons créé plutôt afin de stocker les données.

Nous utilisons aussi grâce à Folium et Branca des méthodes afin de pouvoir voir la carte de la France avec tous les votes enregistrés, une couleur étant associée à chaque partie pour lequel les gens ont voté.



## 4) CONCLUSION

Nous sommes capables de démontrer un lien entre nos données et les élections, il existe bel et bien un moyen de prédire ceci mais pour cela il nous faudrait plus de temps afin d'étudier des solutions d'IA et pouvoir appliquer la meilleure possible sur notre sujet.

Grâce à nos données il serait possible d'entraîner un modèle en ligne afin de fournir des prédictions sur ce qui pourrait se passer en 2027, voir avec plus de temps pour se documenter et apprendre, allez créer notre propre modèle qui sera capable de prendre les données des futures élections afin de les incorporer et de s'entraîner dessus pour donner des prédictions bien plus efficaces et juste sur les élections suivantes.

# ANNEXE

1. AirFlow : <https://airflow.apache.org/>

2. Hugging Face : <https://huggingface.co/>

3. Données du gouvernement :

a. <https://www.data.gouv.fr/fr/datasets/bases-statistiques-communale-departementale-et-regionale-de-la-delinquance-enregistree-par-la-police-et-la-gendarmerie-nationales/>

b. <https://www.data.gouv.fr/fr/datasets/declarations-prealables-a-lembauche-mensuelles-de-plus-dun-mois-france-entiere/#resources-panel>

c. <https://www.insee.fr/fr/statistiques/3596198>

d. <https://www.data.gouv.fr/fr/pages/donnees-des-elections/>

e. <https://www.data.gouv.fr/fr/pages/donnees-securite/>

f. [https://www.data.gouv.fr/fr/pages/donnees\\_emploi/](https://www.data.gouv.fr/fr/pages/donnees_emploi/)

g. [https://www.data.gouv.fr/fr/organizations/institut-national-de-la-statistique-et-des-etudes-economiques-insee/?datasets\\_page=7#organization-datasets](https://www.data.gouv.fr/fr/organizations/institut-national-de-la-statistique-et-des-etudes-economiques-insee/?datasets_page=7#organization-datasets)

4. Bibliothèque de Python:

a. Folium : <https://datascientest.com/folium-tout-savoir>

b. Pandas : <https://pandas.pydata.org/>

c. Simpledbf : <https://pypi.org/project/simpledbf>

d. Aiofiles : <https://pypi.org/project/aiofiles/>

e. AioHttp : <https://docs.aiohttp.org/en/stable/>

f. SQLAlchemy : <https://www.sqlalchemy.org/>

g. Branca : <https://pypi.org/project/branca/>

h. Numpy : <https://numpy.org/>

5. Power-BI : <https://learn.microsoft.com/fr-fr/power-bi/>

6. Le repository du projet : [https://github.com/Sangmin-SHIM/AIRFLOW/tree/MSPR\\_I1](https://github.com/Sangmin-SHIM/AIRFLOW/tree/MSPR_I1)