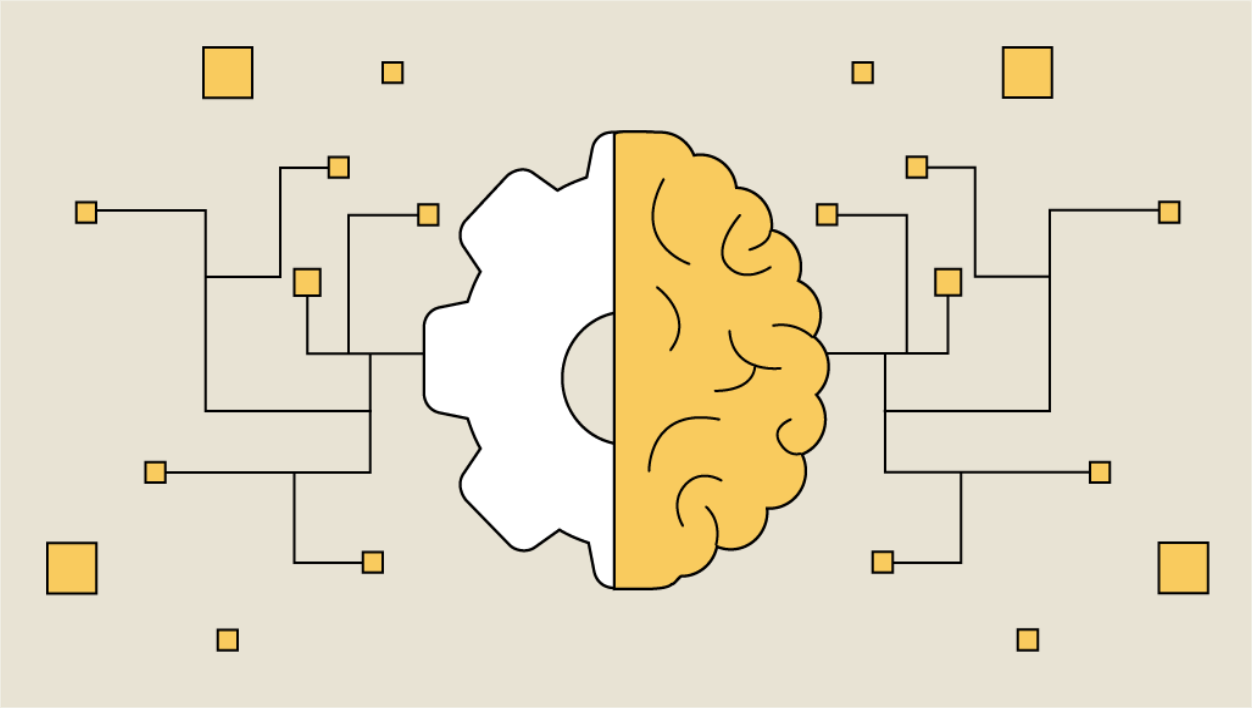
epsi
l'école d'ingénierie informatique

MSPR BLOC 4



- Agouram Yassamine
- Wilson Akuetevi
- Odier Simon
- Mahamat Tahir

TABLE DES MATIERES

I. CONTEXTE

L'entreprise pour laquelle nous avons travaillé dans le cadre de cette MSPR appartient à Jean-Edouard de la Motte Rouge, sa startup est spécialisée dans le conseil en matière de campagnes électorales.

L'objectif est donc d'utiliser l'intelligence artificielle pour prédire les tendances des élections futures tout en se basant sur divers indicateurs (chômage, pib, sécurité ...)

Cette approche innovante lui permettrait d'obtenir un avantage concurrentiel significatif dans son domaine d'activité. Avant d'investir dans une infrastructure et une politique de recherche et développement pour atteindre cet objectif, ainsi que pour solliciter des aides à l'innovation.

I. CHOIX DE LA ZONE GEOGRAPHIQUE

Concernant le choix de la zone géographique, nous avons choisi de nous focaliser sur toute la France, en raison de manque de critères sur une région précise.

II. CHOIX DES CRITERES

Nous avons donc choisi d'étudier la tendance des élections en France de 2002 à 2022, voici les critères choisis :

- CHOMAGE

Le chômage peut fortement influencer les élections en créant un mécontentement économique parmi les électeurs, en changeant leurs priorités politiques vers l'emploi, en favorisant la polarisation, en générant un sentiment d'injustice, en affectant la participation électorale et en incitant les partis politiques à ajuster leurs politiques pour répondre à cette préoccupation centrale.

- INFLATION

L'inflation peut influencer les élections en créant un mécontentement économique parmi les électeurs, les poussant à voter contre le parti au pouvoir et favorisant les candidats qui promettent de lutter contre l'inflation.

- PIB

Le PIB, ou produit intérieur brut, peut avoir un impact significatif sur les élections en influençant la perception des électeurs quant à la performance économique d'un gouvernement. Lorsque le PIB est en croissance robuste, les électeurs ont tendance à être plus enclins à soutenir le parti au pouvoir, car ils associent cette croissance à une gestion économique efficace. En revanche, lorsque le PIB stagne ou diminue, cela peut générer un mécontentement économique et inciter les électeurs à chercher des alternatives politiques, remettant en question les politiques en place. Ainsi, le PIB joue un rôle clé dans la formation des tendances électorales en reflétant la santé économique du pays.

III. DEMARCHE SUIVIE

Le process du projet s'est donc déroulé en plusieurs étapes :

- RECHERCHE DATASET DES ELECTIONS PRESIDENTIELLES SUR PLUSIEURS ANNEES

Notre première mission était de chercher plusieurs datasets des élections présidentielles pendant plusieurs années, nous avons décidé de nous focaliser sur les années allant de 2002 à 2022.

- RECHERCHE DATASET D'INDICATEURS SOCIAUX

En parallèle, une autre partie du groupe cherche les critères pouvant impacter les élections présidentielles afin de prédire les tendances des élections.

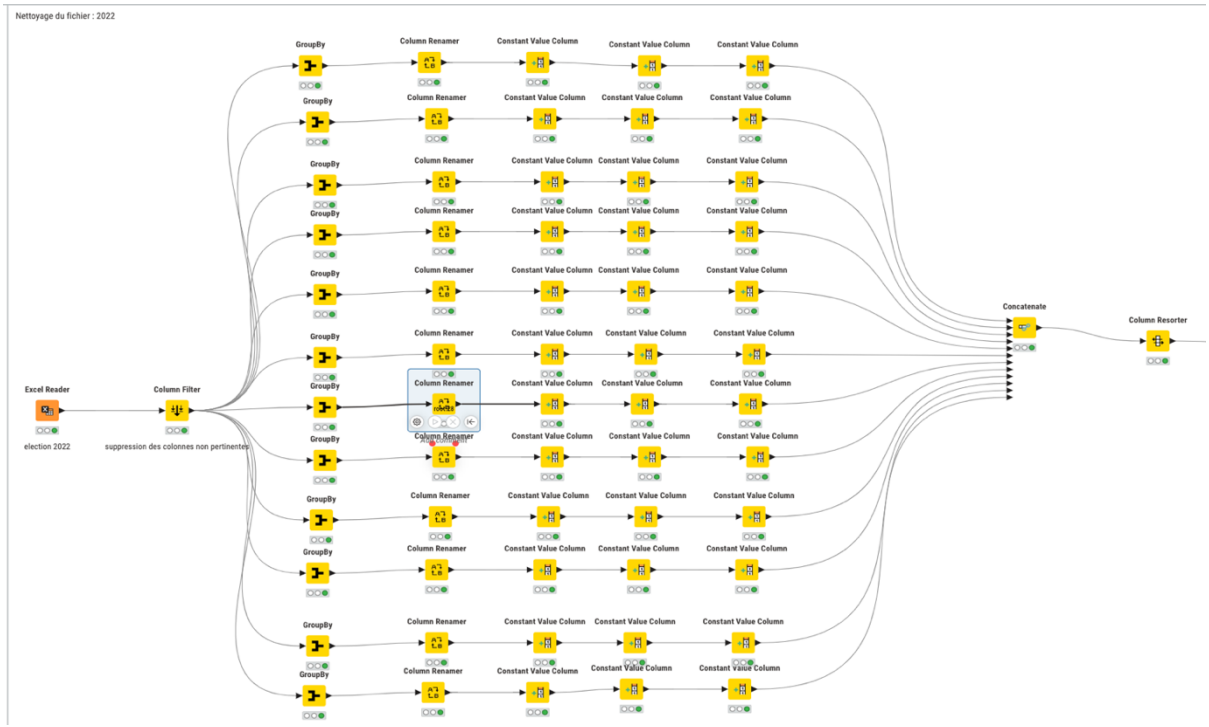
- DICTIONNAIRE

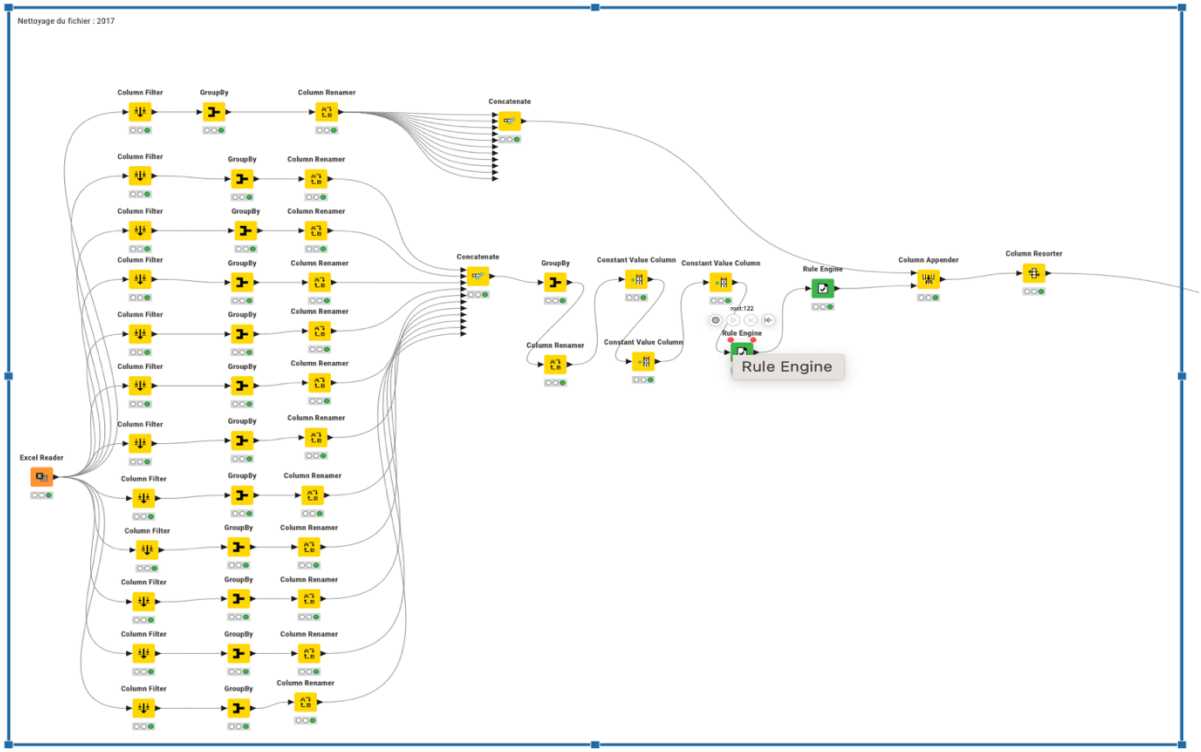
Afin d'avoir une vue globale sur nos datasets (mis à part les datasets des élections parce qu'elles se ressemblent), nous avons réalisé un dictionnaire :

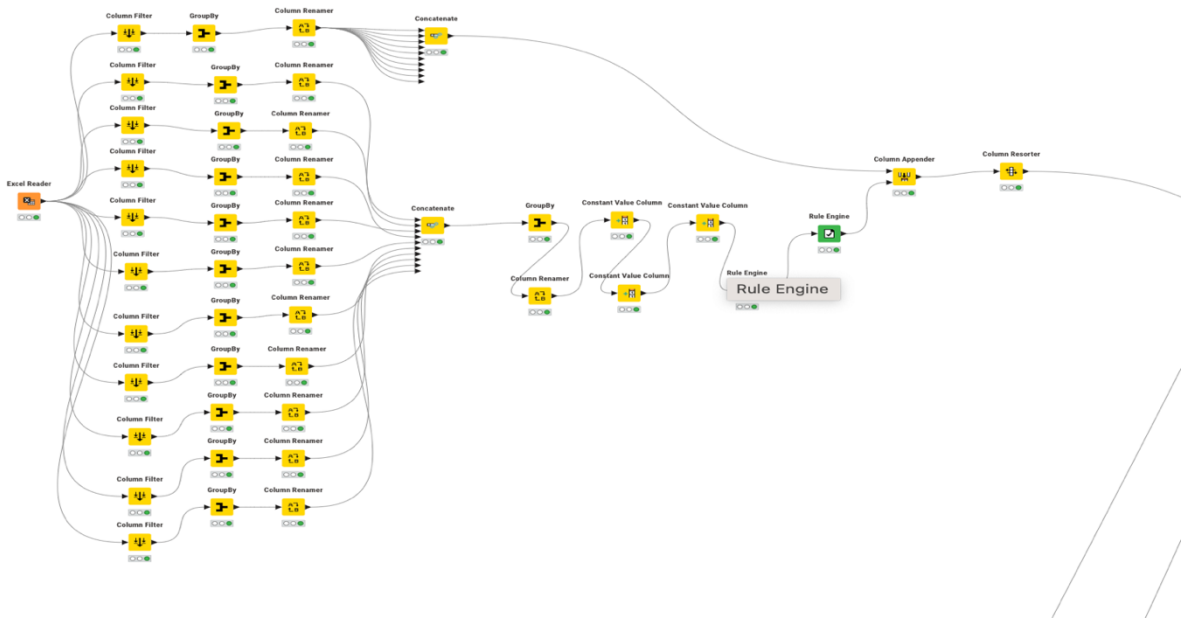
Nom de la table	Nom de la variable	Définition	Type	Unité	Source Web
Chômage	Country Name	Nom du pays	String		https://donnees.banquemondiale.org/indicateur/SL.UEM.TOTL.Z
Chômage	Country Code	Code du Pays	String		https://donnees.banquemondiale.org/indicateur/SL.UEM.TOTL.Z
Chômage	Year	Année	Integer		https://donnees.banquemondiale.org/indicateur/SL.UEM.TOTL.Z
Chômage	Chômage	Taux de chômage	Float		https://donnees.banquemondiale.org/indicateur/SL.UEM.TOTL.Z
Inflation	Country name	Nom du pays	String		https://donnees.banquemondiale.org/indicator/FP.CPI.TOTL.ZG
Inflation	Country Code	Code du Pays	String		https://donnees.banquemondiale.org/indicator/FP.CPI.TOTL.ZG
Inflation	Year	Année	Integer		https://donnees.banquemondiale.org/indicator/FP.CPI.TOTL.ZG
Inflation	Inflation	Taux d'inflation	Float		https://donnees.banquemondiale.org/indicator/FP.CPI.TOTL.ZG
Pib	Country Name	Nom du pays	String		https://donnees.banquemondiale.org/indicator/NY.GDP.PCAP.C
Pib	Country Code	Code du Pays	String		https://donnees.banquemondiale.org/indicator/NY.GDP.PCAP.C
Pib	Year	année	Integer		https://donnees.banquemondiale.org/indicator/NY.GDP.PCAP.C
Pib	PIB	PIB par habitant	Float	$ US	https://donnees.banquemondiale.org/indicator/NY.GDP.PCAP.C

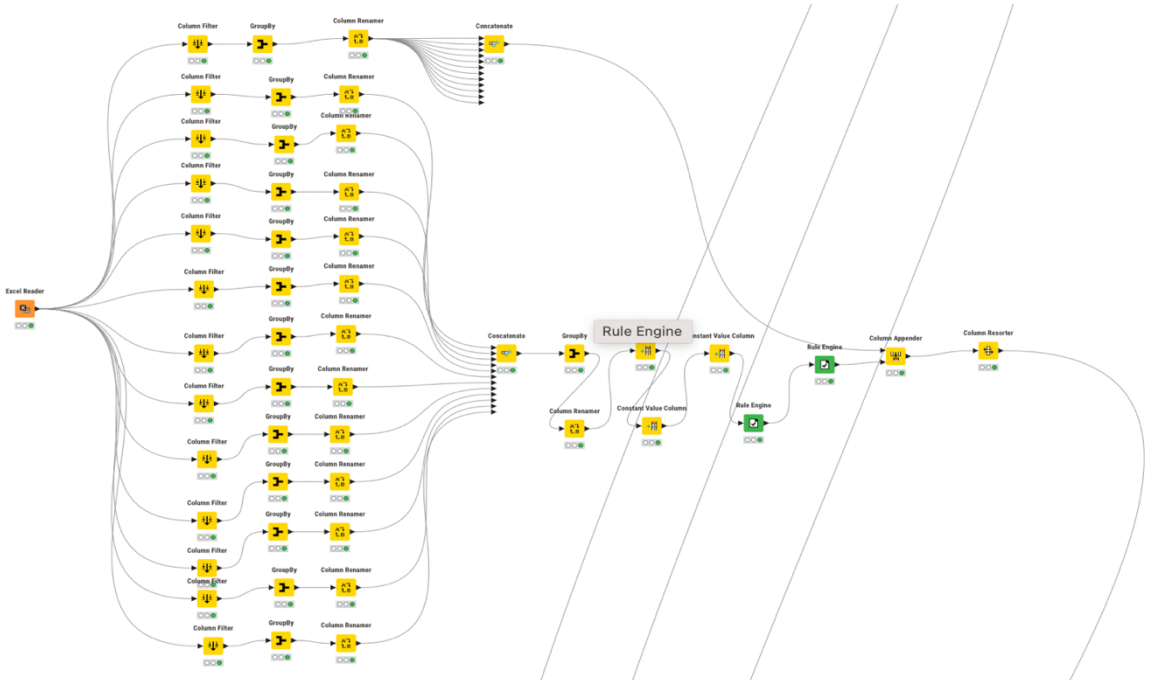
- TRAITEMENT ET NETTOYAGE DES JEUX DE DONNEES

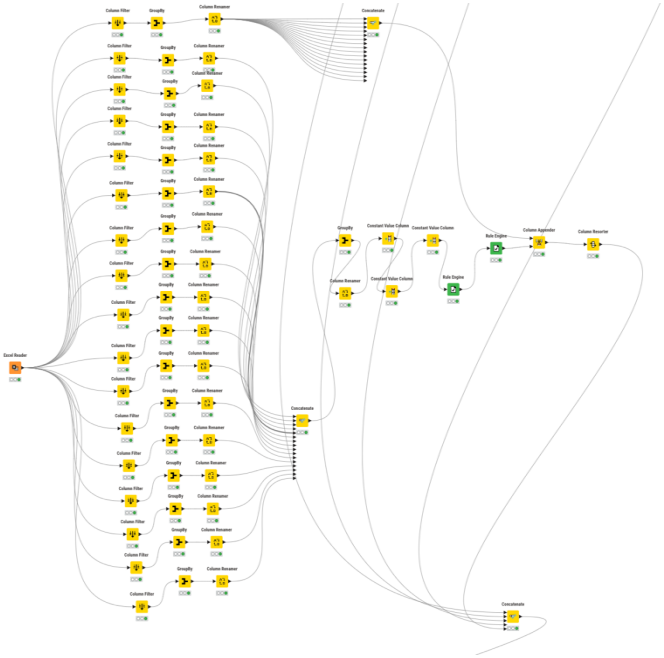
Pour cette partie, nous avons choisi d'utiliser knime pour extraire, transformer et charger les données :






Excel Reader
Column Filter
GroupBy
Column Renamer
Concatenate
Column Appender
Column Resorter
Constant Value Column
Rule Engine

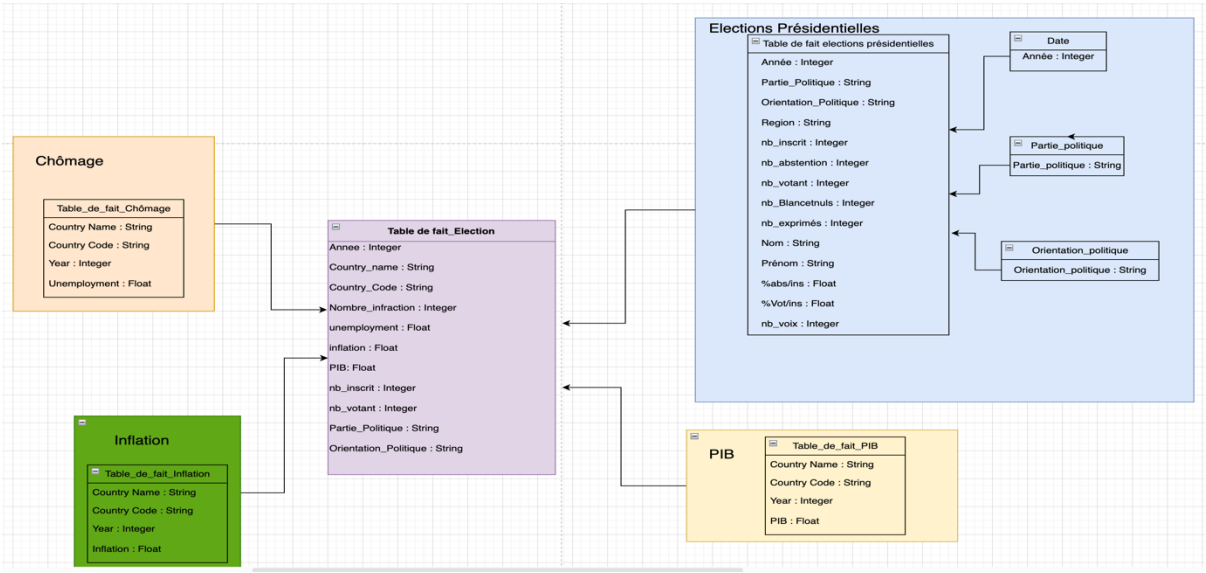

Excel Reader
Column Filter
GroupBy
Column Renamer
Concatenate
Rule Engine



- ÉLABORATION DU MODELE PREDICTIF SUPERVISE

Pour se faire, nous utilisons le langage de programmation python avec des librairies comme scikit learn, pandas et matplotlib

V. MODELE CONCEPTUEL DE DONNEES



IV. MODELES TESTES

- REGRESSION LOGISTIQUE

```
Modèle : Logistic Regression
Précision du modèle : 0.0769
Rapport de classification :
                precision    recall  f1-score   support

        centre       0.00      0.00      0.00         2
 centre droite       0.00      0.00      0.00         1
        droite       0.00      0.00      0.00         3
extreme droite       0.00      0.00      0.00         1
extreme gauche       0.00      0.00      0.00         5
        gauche       0.08      1.00      0.14         1

      accuracy                           0.08        13
     macro avg       0.01      0.17      0.02        13
  weighted avg       0.01      0.08      0.01        13
```

- RANDOM FOREST

```
Modèle : Logistic Regression
Précision du modèle : 0.0769
Rapport de classification :
                precision    recall  f1-score   support

        centre       0.00      0.00      0.00         2
 centre droite       0.00      0.00      0.00         1
        droite       0.00      0.00      0.00         3
extreme droite       0.00      0.00      0.00         1
extreme gauche       0.00      0.00      0.00         5
        gauche       0.08      1.00      0.14         1

      accuracy                           0.08        13
     macro avg       0.01      0.17      0.02        13
  weighted avg       0.01      0.08      0.01        13

================================================
```

- SVM

```
Modèle : SVM
Précision du modèle : 0.0769
Rapport de classification :
                precision    recall  f1-score   support

        centre       0.00      0.00      0.00         2
 centre droite       0.00      0.00      0.00         1
        droite       0.00      0.00      0.00         3
extreme droite       0.00      0.00      0.00         1
extreme gauche       0.00      0.00      0.00         5
        gauche       0.08      1.00      0.14         1

      accuracy                           0.08        13
     macro avg       0.01      0.17      0.02        13
  weighted avg       0.01      0.08      0.01        13

================================================
```

- KNN

```
Modèle : K-Nearest Neighbors
Précision du modèle : 0.0769
Rapport de classification :
                precision    recall  f1-score   support

        centre       0.00      0.00      0.00         2
 centre droite       0.00      0.00      0.00         1
 centre gauche       0.00      0.00      0.00         0
        droite       0.00      0.00      0.00         3
extreme droite       0.00      0.00      0.00         1
extreme gauche       0.33      0.20      0.25         5
        gauche       0.00      0.00      0.00         1

      accuracy                           0.08        13
     macro avg       0.05      0.03      0.04        13
  weighted avg       0.13      0.08      0.10        13
```

Après tous ces tests, nous remarquons que c'est le modèle Random forest qui nous donne la meilleure précision.

V. MODELE CHOISI : RANDOM FOREST

L'utilisation du modèle Random Forest se justifie aussi par :

- PERFORMANCE ROBUSTE

Random Forest est un modèle d'ensemble qui combine plusieurs arbres de décision pour prendre des décisions de classification.

- TRAITEMENT DE CARACTERISTIQUES

Il peut gérer efficacement un grand nombre de caractéristiques et identifier automatiquement les caractéristiques les plus importantes pour la classification. Cela être précieux dans notre cas car nous avons plusieurs variables explicatives comme le chômage, l'inflation et le PIB.

- CAPACITE A GERER DES DONNEES HETEROGENES

Random Forest peut gérer des données de types différents, telles que des données numériques (comme le chômage, l'inflation et le PIB) ainsi que des données catégorielles (comme l'orientation politique).

- REDUCTION DE LA VARIANCE

En agrégeant les résultats de plusieurs arbres de décision, Random Forest a tendance à réduire la variance des prédictions, ce qui en fait un modèle plus stable.

- FACILITE D'UTILISATION

Il est relativement simple à mettre en œuvre grâce à des bibliothèques telles que scikit-learn en Python. De plus, il nécessite moins de réglages d'hyperparamètres que certains autres modèles complexes.

- INTERPRETABILITE

Bien que Random Forest ne soit pas aussi interprétable qu'un seul arbre de décision, il peut encore fournir des informations sur l'importance des caractéristiques, ce qui peut aider à comprendre quelles caractéristiques influencent le plus les prédictions.

VI. PERFORMANCE DU MODELE RANDOM FOREST

```
Précision du modèle : 0.15384615384615385
Rapport de classification :
                precision    recall  f1-score   support

        centre       0.00      0.00      0.00         2
 centre droite       0.00      0.00      0.00         1
        droite       0.25      0.33      0.29         3
extreme droite       0.00      0.00      0.00         1
extreme gauche       0.33      0.20      0.25         5
        gauche       0.00      0.00      0.00         1

      accuracy                           0.15        13
     macro avg       0.10      0.09      0.09        13
  weighted avg       0.19      0.15      0.16        13

Prédiction de l'orientation politique :  extreme gauche
```

On a une précision de 15%, ce qui est très faible car on n'a pas assez de données pour mieux entrainer nos modèles

VII. VISUALISATIONS

QUALITE DE DONNEES DES ELECTIONS PRESIDENTIELLES :

Une étape très importante avant de visualiser les données est d'améliorer la qualité des données.

Comme nous récupérons des données partant de 2002 à 2022, la manière de formuler les datasets change d'une année à une autre.

Pour bien traiter cette problématique, le but était d'avoir un format standard dans tous les datasets ce qui va nous permettre de bien traiter les donner.

Pour ceci, nous nous sommes servis de Power Query qui fournit plusieurs fonctionnalités pour faciliter cette tâche.

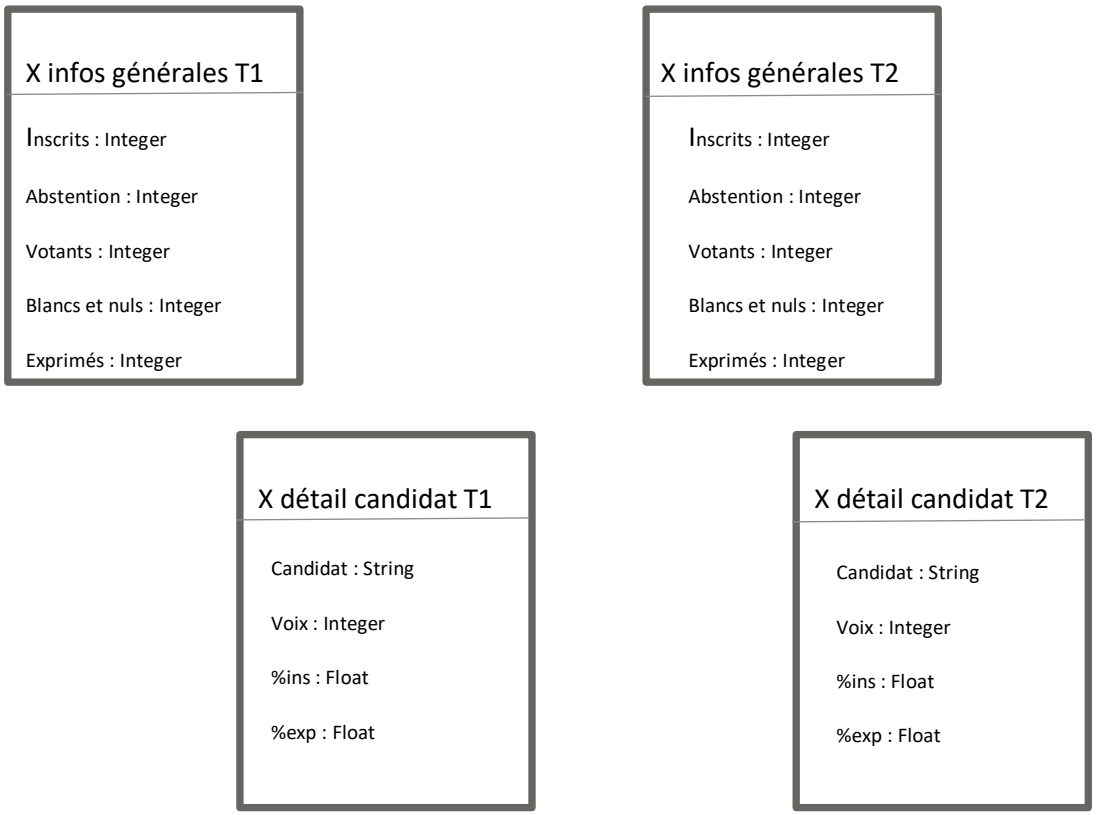
Nous avons formulé donc les données en sorte que nous ayons pour chaque année X, les tables suivantes :

X infos générales T1,

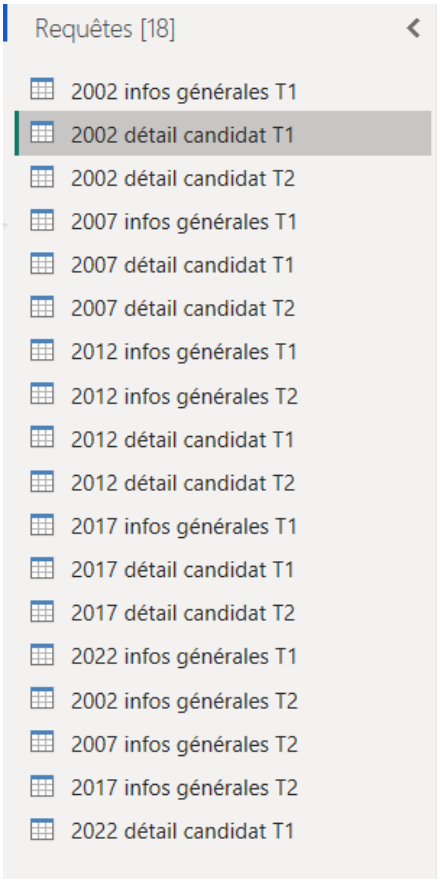
X infos générales T2,

X détail candidat T1,

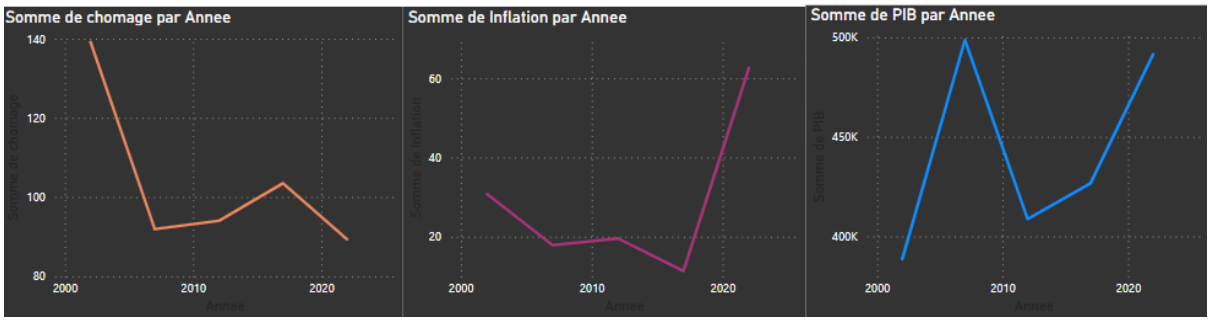
X détail candidat T2.



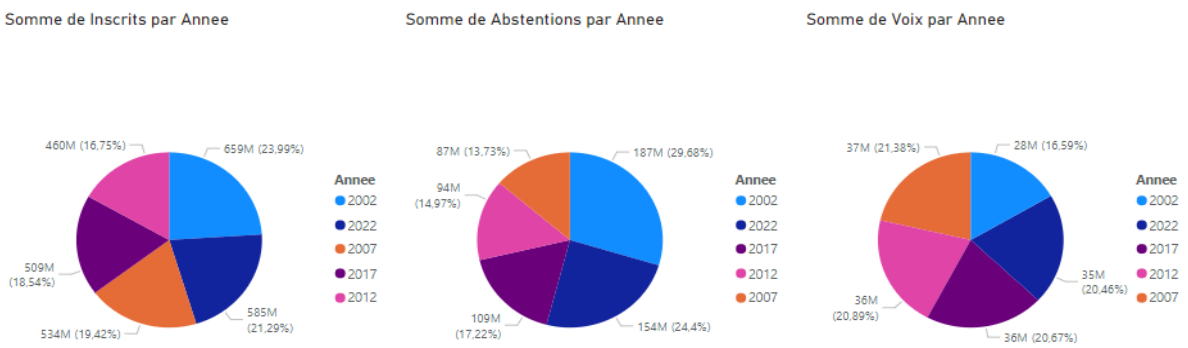
Dans cette étapes-là plusieurs modifications ont été faites grâce au langage M (suppression d'erreurs, pivot des colonnes, modification de types …) pour finalement avoir les tables suivantes avec des données standards et correctes :



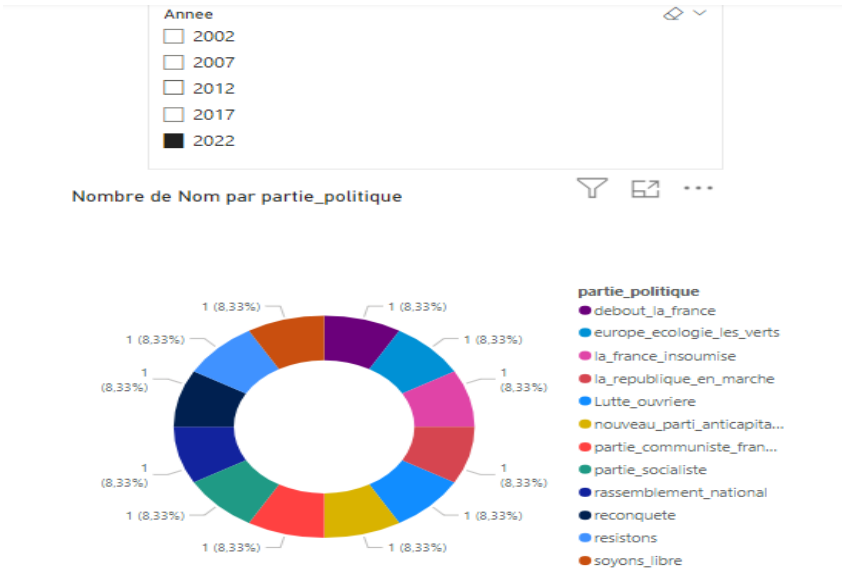
VISUELS POWER BI :



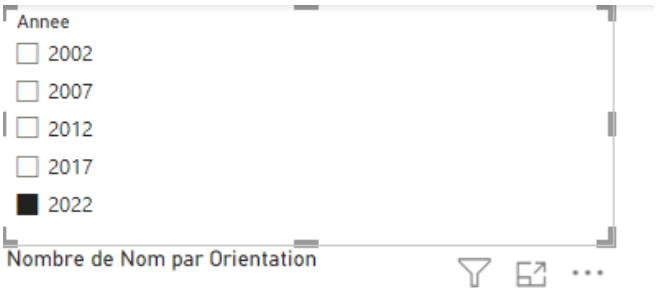
- **Evolution du chômage entre 2002 à 2022 :** Le premier visuel représente l'évolution du chômage, qui était très élevée en 2002 puis a chuté en 2007. La variation est légère juste après.
- **Evolution de l'inflation entre 2002 à 2022 :** Ce visuel illustre l'évolution de l'inflation, on remarque que les variations ne sont pas très remarquables aux premières années, et une augmentation importante à partir de 2017.
- **Evolution du PIB entre 2002 à 2022 :** Ce graphe montre la variation du PIB au fil des années, on remarque un pique en 2007 et une grande diminution en 2012. La courbe augmente après 2012 pour regagner le pique en 2022.

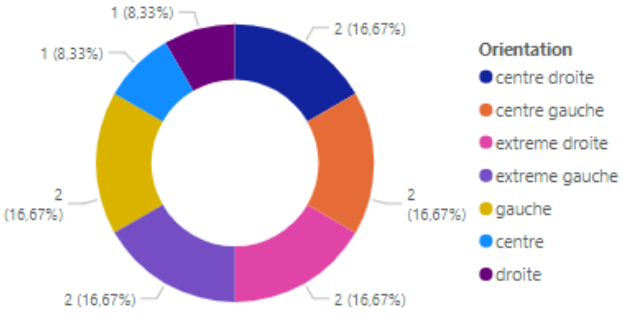


- Les trois visuels représentent respectivement des camemberts de nombre d'inscrits, abstentions et voix des années 2002, 2007, 2012, 2017 et 2022



- Ce graphique en anneau représente les parties politiques présentes selon l'année choisie sur le filtre en dessus.





- Ce visuel représente les orientations des candidats selon l'année choisie sur le filtre en dessus, on remarque qu'il y a une domination du centre droite, centre gauche, extrême droite, extrême gauche et gauche.

Nom	Prénom
ARTHAUD	Nathalie
ASSELINEAU	François
CHEMINADE	Jacques
DUPONT-AIGNAN	Nicolas
FILLON	François
HAMON	Benoît
LASSALLE	Jean
LE PEN	Marine
MACRON	Emmanuel
MÉLENCHON	Jean-Luc
POUTOU	Philippe

- Ce tableau nous permet de voir la liste des candidats selon l'année choisie.

VIII. COMPETENCES ACQUISES

- La capacité à recueillir les besoins en données des différentes directions métiers de l'entreprise, afin d'établir une vue structurée de l'ensemble des données du système d'information et de partager la stratégie globale en matière de données avec le comité de direction
- La compétence à concevoir une stratégie big data, de la collecte au traitement des données, en accord avec les orientations stratégiques convenues avec le comité de direction. Cette stratégie vise à aider l'entreprise à mieux comprendre ses clients et à développer de nouveaux services.
- L'aptitude à structurer les sources de données de manière à produire des résultats exploitables (data visualisation) pour alimenter les outils de prise de décision et présenter les résultats de manière compréhensible, facilitant ainsi la prise de décision au sein des différentes directions métiers
- L'aptitude à assurer la qualité des données en utilisant des outils de gestion de la qualité des données pour garantir leur exactitude, cohérence, synchronisation et traçabilité, répondant ainsi aux exigences d'accessibilité des utilisateurs métiers.

IX. CONTRAINTES

Nous avons rencontré plusieurs problèmes au cours de notre avancement sur ce projet, voici quelques-uns :

- Premièrement le manque de datasets sur une région précise, les données en liaison avec les élections sont facilement trouvables mais pas les facteurs surtout sur des années précises nous étions très limités.
- Le format des jeux de données diffère d'une année à une autre, ce qui a causé une grande perte de temps pour unifier.

X. CONCLUSION

En conclusion, ce projet nous a permis de démontrer la faisabilité de la prédiction de l'orientation politique à partir de données économiques, mais il reste des opportunités pour améliorer la précision en explorant davantage de variables et en ajustant les hyperparamètres du modèle.

Ce projet illustre également l'importance de la collecte des données, de leur prétraitement grâce à un ETL et de la sélection de modèles dans le domaine de l'apprentissage automatique pour résoudre des problèmes de classification.