

# **MSPR3 « Big Data et Analyse de données »**

**CERTIFICATION PROFESSIONNELLE EXPERT EN INFORMATIQUE  
ET SYSTÈME D'INFORMATION RNCP - RNCP N°35584**

**BLOC E7.3 – Piloter l'informatique décisionnel d'un S.I**

**(Big Data & Business Intelligence)**

**MSc Expert en informatique et système d'information**

**I1 EPSI / 24-25**

**Parties prenantes :**

WEMBE II Serge  
EL BOUZIDI Hudaifa  
LAWANI Cheik Anas Deen  
LADINO Javier



# Sommaire

<b>Sommaire.....</b>	<b>2</b>
<b>INTRODUCTION.....</b>	<b>4</b>
CONTEXTE.....	4
<b>1. Sélection et Justification de la Zone Géographique.....</b>	<b>5</b>
📍 Informations Clés.....	5
<b>2. Justification des critères sélectionnés.....</b>	<b>6</b>
<b>3. Méthodologie.....</b>	<b>7</b>
3.1. Collecte et Nettoyage des Données.....	7
Outils pour la collaboration.....	9
3.2. Analyse Exploratoire et des Visualisations.....	14
3.2.1. Tendances générales observées :.....	15
3.2.2. Analyse des corrélations et principales découvertes.....	17
1. Tendances de la participation électorale.....	18
2. Corrélations avec l'orientation politique.....	18
3. Facteurs socio-économiques.....	18
4. Éducation et vote.....	18
5. Démographie et tendances électorales.....	18
6. Évolution temporelle.....	18
3.2.3. Résultat.....	19
3.3. Modèle Conceptuel des Données.....	19
3.3.1. Structure des Bases de Données.....	19
A. Base de Données des Résultats Électoraux.....	19
B. Base de Données Démographiques et Socioéconomiques.....	19
3.3.2. Schéma Graphique de l'Infrastructure Nécessaire.....	20
3.3.3. Étapes de l'Analyse et du Développement du Projet.....	20
Étape 1 : Collecte et Nettoyage des Données.....	20
Étape 2 : Intégration dans Python et Power BI.....	20
Étape 3 : Développement du Modèle Prédictif.....	21
Étape 4 : Visualisation et Simulation de Scénarios.....	21
Étape 5 : Présentation du Projet et Exemples Pratiques.....	21
3.4. Modèle Prédictif Supervisé.....	21
3.4.1. Choix du modèle prédictif supervisé pour l'implémentation.....	21
3.4.1.1. Préparation des données.....	21
3.4.1.2. Modèles supervisés possibles.....	22
3.4.2. Choix recommandé : Gradient Boosting (XGBoost ou LightGBM).....	23
3.4.3. Évaluation du modèle.....	23

3.4.4. Visualisation des prédictions.....	23
3.4.5. Gradient Boosting (XGBoost ou LightGBM).....	23
3.5. Mise en œuvre du modèle.....	24
3.5.1. Fonctionnement du modèle:.....	24
3.5.2. Visualisation et Prédiction.....	28
<b>3.6. Résultats du Modèle.....</b>	<b>28</b>
3.6.1. Principales conclusions du modèle prédictif:.....	28
3.6.2. Recommandations:.....	28
<b>3.7. Visualisations.....</b>	<b>29</b>

# INTRODUCTION

## CONTEXTE

Dans un contexte où les campagnes électorales deviennent de plus en plus complexes et compétitives, la capacité à anticiper les tendances électorales constitue un avantage stratégique majeur pour les acteurs du conseil politique. C'est dans cette perspective que Jean-Edouard de la Motte Rouge a fondé une start-up spécialisée dans l'analyse et le conseil en stratégie électorale. Son ambition est d'exploiter les avancées de l'intelligence artificielle pour prédire les résultats électoraux à partir d'un ensemble d'indicateurs socio-économiques et politiques.

Avant d'investir massivement dans le développement d'une infrastructure technologique et de solliciter des aides à l'innovation, la start-up souhaite d'abord tester la faisabilité de son approche par le biais d'une **preuve de concept (PoC)**. Cette première expérimentation devra démontrer la pertinence de l'utilisation de modèles prédictifs appliqués à un périmètre géographique restreint.

Pour mener à bien cette étude, plusieurs étapes seront suivies :

1. **Sélection d'un secteur géographique** précis afin de circonscrire l'analyse et d'en assurer la pertinence.
2. **Collecte et sélection de jeux de données** pertinents, incluant des indicateurs comme la sécurité, l'emploi, la vie associative, la population, la dynamique économique et la pauvreté.
3. **Visualisation et exploration des données**, afin d'identifier des tendances et des corrélations significatives entre le contexte socio-économique et les résultats électoraux passés.
4. **Conception et entraînement d'un modèle prédictif supervisé**, permettant d'anticiper les tendances électorales en s'appuyant sur des données historiques.
5. **Représentation graphique des prédictions**, afin d'évaluer la fiabilité du modèle et d'illustrer les projections à 1, 2 et 3 ans.

Ce projet vise à démontrer la viabilité et la robustesse d'un outil prédictif dans le domaine électoral, ouvrant ainsi la voie à une exploitation plus large et à une intégration au sein des services de la start-up.

# 1. Sélection et Justification de la Zone Géographique

- **Zone géographique choisie** : La région Bretagne.

**La Bretagne** est une région située dans le **nord-ouest de la France**, connue pour son **fort héritage celtique**, **ses paysages côtiers spectaculaires** et **sa culture unique**. Elle est bordée par l'**océan Atlantique**, la **Manche**, ainsi que les régions de la **Normandie** et des **Pays de la Loire**.

---

## 📌 Informations Clés

- **Capitale** : Rennes
- **Départements** : Ille-et-Vilaine, Morbihan, Finistère et Côtes-d'Armor
- **Population** : Environ 3,3 millions d'habitants
- **Langues** : Français et breton (langue régionale d'origine celtique)

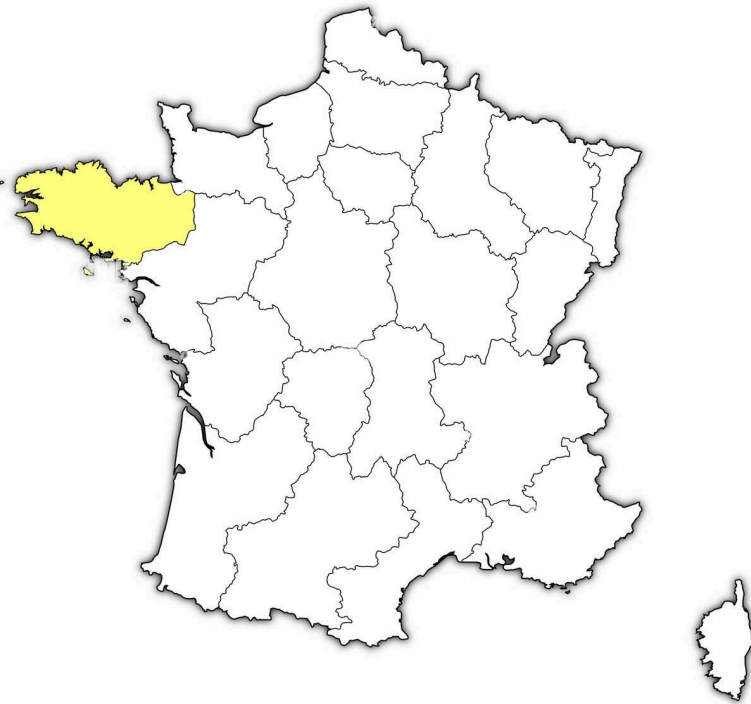


Fig 1. Carte de la région Bretagne, en France.

- **Raisons du choix :**

- La Bretagne est une région économiquement et socialement diversifiée, combinant des zones urbaines (Rennes, Brest) et rurales.
- Données disponibles via des sources officielles comme l'INSEE et les portails gouvernementaux.
- Historique riche en participation électorale, offrant un terrain d'analyse intéressant pour les tendances politiques.

Au cours de cette période, la France a connu des changements politiques majeurs qui ont influencé les résultats électoraux en Bretagne. Voici quelques événements marquants :

- **1965** : Première élection présidentielle au suffrage universel direct, remportée par Charles de Gaulle face à François Mitterrand.
- **1981** : Victoire historique de François Mitterrand (PS), marquant l'alternance politique et le début d'une longue période de domination socialiste en Bretagne.
- **2002** : Percée du Front National avec Jean-Marie Le Pen au second tour contre Jacques Chirac, bien que la Bretagne ait voté massivement contre l'extrême droite.
- **2017 et 2022** : Emmanuel Macron (LREM) l'emporte avec un fort soutien breton, tandis que Marine Le Pen (RN) progresse mais reste minoritaire dans la région.

Ces événements montrent l'évolution politique de la Bretagne, souvent considérée comme une région progressiste et pro-européenne.

## 2. Justification des critères sélectionnés

La sélection des critères utilisés pour la construction du modèle de tendances électorales repose sur des facteurs ayant historiquement un impact significatif sur le comportement électoral. Ces critères permettent d'établir des corrélations entre les dynamiques socio-économiques et les résultats des scrutins passés, tout en offrant une base robuste pour des projections futures.

- **Démographie** : La structure de la population, notamment l'âge moyen et la croissance démographique, influence les préférences électorales, certains groupes d'âge ayant des tendances de vote distinctes.
- **Économie** : Les indicateurs économiques comme le taux de chômage et la création d'entreprises sont souvent des facteurs déterminants du vote, les électeurs étant sensibles aux conditions économiques locales.

- **Sécurité** : Le taux de criminalité peut influencer la perception des électeurs sur les politiques publiques et jouer un rôle clé dans la montée de certaines tendances électorales.
- **Éducation** : Le niveau d'études a un impact sur la perception politique et les préférences idéologiques, ce qui justifie son inclusion dans le modèle.

Ces critères, combinés dans un modèle prédictif, offrent une approche holistique et pertinente pour anticiper les tendances électorales, en intégrant les facteurs les plus influents sur le comportement des électeurs.

## 3. Méthodologie

### 3.1. Collecte et Nettoyage des Données


#### Sources disponibles :

La provenance des données est garantie à la source. Les bases de données ont été téléchargées à partir des portails officiels du gouvernement français. Il est possible de se connecter automatiquement via leurs API, mais le service est très intermittent. C'est pourquoi nous travaillons avec les données localement dans un *Data Lake*.

Critères	Source	Année
<b>Elections</b>		
Élection présidentielle 1965-2012:	<a href="#">Elections présidentielles 1965-2012 - data.gouv.fr</a>	1965-2012
<b>Démographie</b>		
Population	<a href="https://data.bretagne.bzh/explore/dataset/recensement-de-la-population-en-bretagne-evolution-de-la-population">https://data.bretagne.bzh/explore/dataset/recensement-de-la-population-en-bretagne-evolution-de-la-population</a>	1999 - 2021
Âge moyen	<a href="https://www.insee.fr/fr/statistiques/7750177#:~:text=En%202022%2C%20l%27%C3%A2ge%20moyen,2%20ans%20en%2010%20ans).">https://www.insee.fr/fr/statistiques/7750177#:~:text=En%202022%2C%20l%27%C3%A2ge%20moyen,2%20ans%20en%2010%20ans).</a>	1975 - 2022
Densité	<a href="https://www.ined.fr/fr/tout-savoir-population/chiffres/france/structure-population/regions/">https://www.ined.fr/fr/tout-savoir-population/chiffres/france/structure-population/regions/</a>	1995 - 2024
<b>Économie</b>		
Taux de chômage	<a href="https://www.insee.fr/fr/statistiques/serie/001515855#Telechargement">https://www.insee.fr/fr/statistiques/serie/001515855#Telechargement</a>	1982 - 2024
Création d'entreprises	<a href="https://www.insee.fr/fr/statistiques/serie/010756644#Telechargement">https://www.insee.fr/fr/statistiques/serie/010756644#Telechargement</a>	2012-2024
	<a href="https://www.insee.fr/fr/statistiques/serie/001787281#Telechargement">https://www.insee.fr/fr/statistiques/serie/001787281#Telechargement</a>	2012-2024
	<a href="https://www.insee.fr/fr/statistiques/serie/001564353">https://www.insee.fr/fr/statistiques/serie/001564353</a>	2012-2024
<b>Sécurité</b>		
Taux de criminalité par département	<a href="https://www.data.gouv.fr/fr/datasets/bases-statistiques-communale-departementale-et-regionale-de-la-delinquance-enregistree-par-la-police-et-la-gendarmerie-nationales/">https://www.data.gouv.fr/fr/datasets/bases-statistiques-communale-departementale-et-regionale-de-la-delinquance-enregistree-par-la-police-et-la-gendarmerie-nationales/</a>	2016 - 2022
	<a href="https://mobile.interieur.gouv.fr/Media/SSMSI/Files/Donnees-Bilan-statistique-Insecurite-et-delinquance-2023">https://mobile.interieur.gouv.fr/Media/SSMSI/Files/Donnees-Bilan-statistique-Insecurite-et-delinquance-2023</a>	2016 - 2023

Éducation		
Niveau d'études	<a href="https://focus-emploi-formation-bretagne.bzh/population-residente/caracteristiques-de-la-population/departement/22">https://focus-emploi-formation-bretagne.bzh/population-residente/caracteristiques-de-la-population/departement/22</a>	2015 -2021

Nous avons créé une structure de dossiers pour stocker tous les éléments produits dans le projet : données, images, notebooks, code, etc. Il a été hébergé sur **Google Drive** pour travailler de manière collaborative, garantissant ainsi un accès immédiat à tous les membres de l'équipe.

 0.Bases de données
 1.ETL
 2.Analyse Exploratoire et des Visualisations
 3.Modèle Prédictif
 4.Synthèse et de la Présentation
 5.POC

*Structure de données sur Google Drive*

### Traitement des données :

Pour l'ETL (Extraction, Transformation et Chargement) des données, nous utilisons **Google COLAB** comme environnement de développement. L'implémentation est basée sur Python, un langage largement utilisé pour sa rapidité et sa polyvalence dans le traitement des données. L'ensemble du développement et de la documentation est capturé dans des **Jupyter Notebooks (Fichiers.ipynb)**, ce qui nous permet de tester et d'itérer efficacement notre solution. **Jupyter Notebook** est une application web interactive utilisée pour créer et partager des documents informatiques pour l'analyse à l'aide de Python.

Tout le code sera hébergé dans notre dépôt **Github** pour votre référence.



**mspr\_bloce73\_i1\_infra\_2425** (Public)

Pin Unwatch (1) Fork (0) Star (0)

**Set up GitHub Copilot**  
Use GitHub's AI pair programmer to autocomplete suggestions as you code.  
[Get started with GitHub Copilot](#)

**Add collaborators to this repository**  
Search for people using their GitHub username or email address.  
[Invite collaborators](#)

**Quick setup — if you've done this kind of thing before**

[Set up in Desktop](#) or [HTTPS](#) [SSH](#) [https://github.com/javiladino/mspr\\_bloce73\\_i1\\_infra\\_2425.git](https://github.com/javiladino/mspr_bloce73_i1_infra_2425.git)

Get started by [creating a new file](#) or [uploading an existing file](#). We recommend every repository include a [README](#), [LICENSE](#), and [.gitignore](#).

**...or create a new repository on the command line**

```
echo "# mspr_bloce73_i1_infra_2425" >> README.md
git init
git add README.md
git commit -m "first commit"
git branch -M main
git remote add origin https://github.com/javiladino/mspr_bloce73_i1_infra_2425.git
git push -u origin main
```

**...or push an existing repository from the command line**

```
git remote add origin https://github.com/javiladino/mspr_bloce73_i1_infra_2425.git
git branch -M main
git push -u origin main
```

**mspr\_bloc\_3\_v1.ipynb** ☆ ☁

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda

Comandos + Código + Texto

RAM Disco

**Base de Données des Résultats Électoraux**

```
[ ] import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

[ ] def charger_donnees(chemin_fichier):
    """Charge le fichier Excel et retourne un dictionnaire avec les DataFrames de chaque feuille."""
    xls = pd.ExcelFile(chemin_fichier)
    donnees = {feuille: pd.read_excel(xls, feuille) for feuille in xls.sheet_names}
    return donnees

def transformer_donnees(df, niveau='Région'):
    """Transforme les données électorales dans un format structuré."""
    if niveau == 'Région':
        cle_localisation = 'Libellé de la région'
    elif niveau == 'Département':
        cle_localisation = 'Libellé du département'
    else:
        raise ValueError("Niveau non reconnu. Utilisez 'Région' ou 'Département'")

    # Colonnes générales à conserver
    colonnes_generales = ['Année', cle_localisation, 'Inscrits', 'Abstentions',
                          '% Abs/Ins', 'Votants', '% Tot/Ins', 'Blancs et nuls', '% Blancs/Vot']
```

Un fichier **.ipynb** est un fichier Jupyter Notebook qui contient tout le contenu créé avec la session de l'application web Jupyter Notebook ou Google COLAB

[Répertoire sur Github](#)

## Outils pour la collaboration

- **Gestion des tâches** : ClickUP
- **Versionnement du code** : GitHub.
- **Communication** : Microsoft Teams - Discord
- **Documentation partagée** : Google Drive - Markdown

L'ordre de notre ETL était le suivant :

- Installation des bibliothèques nécessaires.

```
[ ] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
```

- Nous avons d'abord travaillé avec les bases de données électorales, puis avec les bases de données démographiques et socio-économiques.

#### ✓ Base de Données des Résultats Électoraux

```
[ ] def charger_donnees(chemin_fichier):
    """Charge le fichier Excel et retourne un dictionnaire avec les DataFrames de chaque feuille."""
    xls = pd.ExcelFile(chemin_fichier)
    donnees = {feuille: pd.read_excel(xls, feuille) for feuille in xls.sheet_names}
    return donnees

def transformer_donnees(df, niveau='Région'):
    """Transforme les données électorales dans un format structuré."""

    if niveau == 'Région':
        cle_localisation = 'Libellé de la région'
    elif niveau == 'Département':
        cle_localisation = 'Libellé du département'
    else:
        raise ValueError("Niveau non reconnu. Utilisez 'Région' ou 'Département'")

    # Colonnes générales à conserver
    colonnes_generales = ['Année', cle_localisation, 'Inscrits', 'Abstentions',
                          '% Abs/Ins', 'Votants', '% Tot/Ins', 'Blancs et nuls', '% Blancs/Vot']

    # Identifier les colonnes des candidats
    candidats = []
    for i in range(16): # Jusqu'à 16 candidats dans certaines élections
        col_sexe = f'Sexe.{i}' if i > 0 else 'Sexe'
        col_nom = f'Nom.{i}' if i > 0 else 'Nom'
        col_prenom = f'Prénom.{i}' if i > 0 else 'Prénom'
        col_voix = f'Voix.{i}' if i > 0 else 'Voix'
        col_voix_ins = f'% Voix/Ins.{i}' if i > 0 else '% Voix/Ins'

        if col_nom in df.columns:
            candidats.append([col_sexe, col_nom, col_prenom, col_voix, col_voix_ins])
```

- Il est nécessaire de séparer les ensembles de données en différents ensembles de données. Cela permet d'améliorer les requêtes et d'optimiser le code.

```
[ ] df_colour_candidates.head()
```

	Sexe	Nom	Prénom	PARTI POLITIQUE	COULEUR POLITIQUE	nom_prenom
0	M	DE VILLIERS	Philippe	Mouvement pour la France (MPF)	Droite conservatrice	Philippe DE VILLIERS
1	M	LE PEN	Jean-Marie	Front National (FN)	Extrême droite	Jean-Marie LE PEN
2	M	CHIRAC	Jacques	Rassemblement pour la République (RPR)	Droite républicaine	Jacques CHIRAC
3	F	LAGUILLER	Arlette	Lutte Ouvrière (LO)	Extrême gauche	Arlette LAGUILLER
4	M	CHEMINADE	Jacques	Solidarité et Progrès	Divers	Jacques CHEMINADE

- Validation du chargement correct de toutes les données. Chaque fichier de données suit le même processus de chargement, de nettoyage et de transformation. Les sources étant différentes, ajustez toujours l'ensemble de données pour la consolidation finale.

```
[ ] df_bretagne_t2.head(12)
```

	Année	Région	Inscrits	Abstentions	% Abstentions	Votants	% Votants	Blancs et nuls	Blancs/Vot	%	Sexe	Nom	Prénom	Voix	Voix/Inscrits	%
0	2022	Bretagne	2562764	566269	22.09	1996495	77.90	190884	9.56	M	MACRON	Emmanuel	1202202	46.910367		
1	2022	Bretagne	2562764	566269	22.09	1996495	77.90	190884	9.56	F	LE PEN	Marine	603409	23.545243		
2	2017	Bretagne	2453233	498305	20.31	1954928	79.69	228240	8.78	M	MACRON	Emmanuel	1301226	53.041272		
3	2017	Bretagne	2453233	498305	20.31	1954928	79.69	228240	8.78	F	LE PEN	Marine	425462	17.342910		
4	2012	Bretagne	2380266	360356	15.14	2019910	84.86	107827	5.34	M	HOLLANDE	François	1077551	45.270192		
5	2012	Bretagne	2380266	360356	15.14	2019910	84.86	107827	5.34	M	SARKOZY	Nicolas	834532	35.060451		
6	2007	Bretagne	2313685	288486	12.47	2025199	87.53	80887	3.99	M	SARKOZY	Nicolas	921218	39.816051		
7	2007	Bretagne	2313685	288486	12.47	2025199	87.53	80887	3.99	F	ROYAL	Ségolène	1023094	44.219243		
8	2002	Bretagne	2181473	370186	16.97	1811287	83.03	91187	5.03	M	CHIRAC	Jacques	1523388	69.832998		
9	2002	Bretagne	2181473	370186	16.97	1811287	83.03	91187	5.03	M	LE PEN	Jean-Marie	196712	9.017393		
10	1995	Bretagne	2099131	358796	17.09	1740335	82.91	77671	4.46	M	CHIRAC	Jacques	841297	40.078347		
11	1995	Bretagne	2099131	358796	17.09	1740335	82.91	77671	4.46	M	JOSPIN	Lionel	821367	39.128906		

- La création de nouvelles colonnes, la modification des noms de colonnes, le nettoyage des données, la correction des types de données et leur regroupement pour mettre en œuvre l'analyse sont essentiels pour accélérer l'analyse et éviter les erreurs courantes en exploitation.

```
[ ] df_bretagne_t1['nom_prenom'] = df_bretagne_t1['Prénom'] + ' ' + df_bretagne_t1['Nom']
df_bretagne_t2['nom_prenom'] = df_bretagne_t2['Prénom'] + ' ' + df_bretagne_t2['Nom']
df_depts_t1['nom_prenom'] = df_depts_t1['Prénom'] + ' ' + df_depts_t1['Nom']
df_depts_t2['nom_prenom'] = df_depts_t2['Prénom'] + ' ' + df_depts_t2['Nom']

[ ] df_bretagne_t1['exprime'] = df_bretagne_t1['Votants'] - df_bretagne_t1['Blancs et nuls']
df_bretagne_t2['exprime'] = df_bretagne_t2['Votants'] - df_bretagne_t2['Blancs et nuls']
df_depts_t1['exprime'] = df_depts_t1['Votants'] - df_depts_t1['Blancs et nuls']
df_depts_t2['exprime'] = df_depts_t2['Votants'] - df_depts_t2['Blancs et nuls']

[ ] df_bretagne_t1['%_voix_obtenu'] = (df_bretagne_t1['Voix'] - df_bretagne_t1['exprime']) * 100
df_bretagne_t2['%_voix_obtenu'] = (df_bretagne_t2['Voix'] - df_bretagne_t2['exprime']) * 100
df_depts_t1['%_voix_obtenu'] = (df_depts_t1['Voix'] - df_depts_t1['exprime']) * 100
df_depts_t2['%_voix_obtenu'] = (df_depts_t2['Voix'] - df_depts_t2['exprime']) * 100

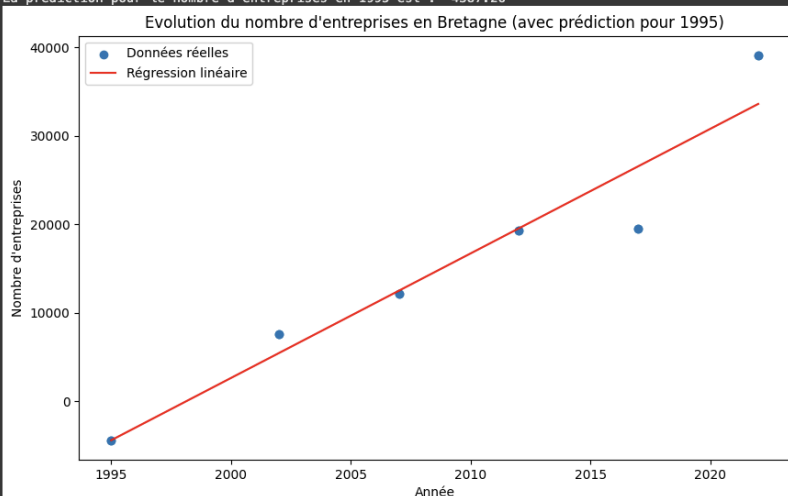
# Merge with the color information
df_bretagne_t1_c = pd.merge(df_bretagne_t1, df_colour_candidates, on='nom_prenom', how='left')
df_bretagne_t2_c = pd.merge(df_bretagne_t2, df_colour_candidates, on='nom_prenom', how='left')
df_depts_t1_c = pd.merge(df_depts_t1, df_colour_candidates, on='nom_prenom', how='left')
df_depts_t2_c = pd.merge(df_depts_t2, df_colour_candidates, on='nom_prenom', how='left')

df_bretagne_t2_c.head(12)
```

- La recherche de données est un processus long et incertain. Parfois, nous ne savons pas si les données que nous allons trouver seront utiles. Dans notre cas, il était impossible d'obtenir les valeurs exactes de certaines années et de certains critères. Nous avons donc eu recours à la régression linéaire, qui est un algorithme d'apprentissage supervisé utilisé dans l'apprentissage automatique et les statistiques. Cet algorithme permet de prédire des valeurs en fonction de leur tendance et de leur comportement.

```
# Créer un graphique avec Matplotlib
plt.figure(figsize=(10, 6))
plt.scatter(df_merged_bretagne_filtre['Année'], df_merged_bretagne_filtre['nombre_entreprises'], label='Données réelles')
plt.plot(df_merged_bretagne_filtre['Année'], model.predict(df_merged_bretagne_filtre[['Année']]), color='red', label='Régression linéaire')
plt.xlabel('Année')
plt.ylabel('Nombre d\'entreprises')
plt.title('Evolution du nombre d\'entreprises en Bretagne (avec prédiction pour 1995)')
plt.legend()
plt.show()
```

La prédiction pour le nombre d'entreprises en 1995 est : -4387.26



- En fin de compte, nous cherchons à tout unifier dans un seul cadre de données qui nous permette de commencer l'analyse exploratoire des données et de passer ensuite à notre modèle prédictif.

```
[ ] df_merged_bretagne_filtre.head()
```

	Année	Votants	%_voix_obtenus	nom_prenom	PARTI POLITIQUE	COULEUR POLITIQUE	0 à 19 ans	20 à 39 ans	40 à 59 ans	60 à 74 ans	75 ans et plus	Total	taux_chomage	taux_pour_mille	nombre_entreprises	Baccalauréat professionnel	Baccalauréat technologique	Baccalauréat général
13	1995	1740335.0	82.91	Jacques CHIRAC	Rassemblement pour la République (RPR)	Droite républicaine	735549.0	805794.0	665278.0	449124.0	184935.0	2840680.0	8.325	1.287348	-4387.26	7.9	17.6	37.2
14	1995	1740335.0	82.91	Lionel JOSPIN	Parti Socialiste (PS)	Gauche	735549.0	805794.0	665278.0	449124.0	184935.0	2840680.0	8.325	1.287348	-4387.26	7.9	17.6	37.2
21	2002	1811287.0	83.03	Jacques CHIRAC	Rassemblement pour la République (RPR)	Droite républicaine	738672.0	783368.0	777030.0	432726.0	249969.0	2981765.0	6.375	1.548055	7618.00	11.5	17.7	32.4
22	2002	1811287.0	83.03	Jean-Marie LE PEN	Front National (FN)	Extrême droite	738672.0	783368.0	777030.0	432726.0	249969.0	2981765.0	6.375	1.548055	7618.00	11.5	17.7	32.4
27	2007	2025199.0	87.53	Nicolas SARKOZY	Union pour un Mouvement Populaire (UMP)	Droite républicaine	762553.0	771058.0	856770.0	436130.0	293477.0	3120288.0	6.525	1.734274	12146.00	12.6	16.4	33.7

Après ce processus ETL, nous sécurisons nos données normalisées et propres dans une base de données SQL. Dans ce cas, nous utiliserons Postgresql et SQLAlchemy, une bibliothèque Python qui s'y connecte directement pour effectuer des requêtes de toutes sortes. Nous obtenons ainsi des données optimisées que nous pouvons utiliser dans des systèmes tels que Power BI, pour des tableaux de bord et des visualisations interactives à l'intention du client ou d'autres utilisateurs.

```

Enregister Dataset SQL

[1] pip install mysql-connector-python

Collecting mysql-connector-python
  Downloading mysql_connector_python-9.2.0-cp311-cp311-manylinux_2_28_x86_64.whl.metadata (6.0 kB)
  Downloading mysql_connector_python-9.2.0-cp311-cp311-manylinux_2_28_x86_64.whl (34.0 MB)
    34.0/34.0 MB 8.2 MB/s eta 0:00:00
Installing collected packages: mysql-connector-python
Successfully installed mysql-connector-python-9.2.0

import mysql.connector
import pandas as pd
from sqlalchemy import create_engine
from datetime import datetime

pd.set_option('mode.chained_assignment', None)

# Charger le dataset
nom_fichier = "df_merged_bretagne_filtre.csv"
df = pd.read_csv(nom_fichier)

# Connexion à la base de données MySQL (mspr1)
connexion = mysql.connector.connect(
    host="localhost",
    user="root",
    password="NouvelleMotDePasseSecurisee",
    database="bretagne_data")

# Création du moteur SQLAlchemy
moteur = create_engine("mysql+mysqlconnector://root:NouvelleMotDePasseSecurisee@localhost/bretagne_data")

# Sauvegarde du DataFrame dans la base de données
table_nom = "elections"
df.to_sql(table_nom, moteur, if_exists='replace', index=False)

print(f"Le dataset a été sauvegardé avec succès dans la table '{table_nom}'.")

# Fermeture de la connexion
connexion.close()

```

## 3.2. Analyse Exploratoire et des Visualisations

Les données couvrent les élections présidentielles en Bretagne de 1995 à 2022, avec des informations sur :

- Les résultats électoraux (votants, pourcentages)
- Les candidats et leurs affiliations politiques
- La démographie de la région (répartition par âge)
- Les indicateurs socio-économiques (taux de chômage, nombre d'entreprises)
- Les niveaux d'éducation (pourcentages des différents types de baccalauréat)

Dans cette analyse de statistiques descriptives, nous examinons plusieurs aspects des données électorales, économiques et démographiques. Nous commençons par une analyse générale des variables numériques et catégorielles afin d'obtenir un aperçu global des données.

```
# Statistiques descriptives pour les variables numériques
print(data.describe())

# Statistiques descriptives pour les variables catégorielles
print(data.describe(include=['object']))

# Analyse de la distribution des votes par orientation politique
print(data.groupby('COULEUR POLITIQUE')['%_voix_obtenu'].describe())

# Analyse de la participation électorale par année
print(data.groupby('Année')['%_Votants'].describe())

# Analyse du taux de chômage par orientation politique
print(data.groupby('COULEUR POLITIQUE')['taux_chomage'].describe())

# Analyse du niveau d'éducation par orientation politique
print(data.groupby('COULEUR POLITIQUE')['Baccalauréat professionnel', 'Baccalauréat technologique', 'Baccalauréat général'].describe())

# Analyse de la structure démographique par orientation politique
print(data.groupby('COULEUR POLITIQUE')['% 0-19 ans', '% 20-39 ans', '% 40-59 ans', '% 60-74 ans', '% 75+ ans'].describe())
```

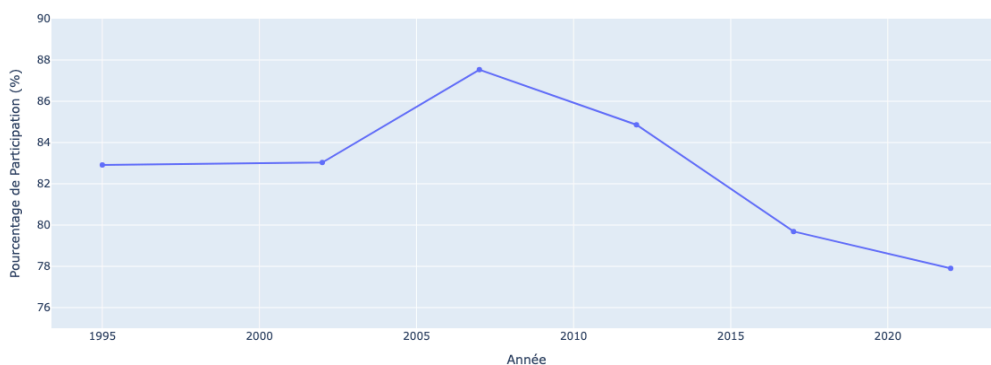
	Année	Votants	%_Votants	%_voix_obtenu	0 à 19 ans \
count	12.000000	1.200000e+01	12.000000	12.000000	12.000000
mean	2009.166667	1.924692e+06	82.653333	50.000000	764147.833333
std	9.446821	1.144840e+05	3.309340	21.116329	21672.939285
min	1995.000000	1.740335e+06	77.900000	11.440000	735549.000000
25%	2002.000000	1.811287e+06	79.690000	41.092500	738672.000000
50%	2009.500000	1.975712e+06	82.970000	50.000000	769482.500000
75%	2017.000000	2.019910e+06	84.860000	58.907500	781997.000000
max	2022.000000	2.025199e+06	87.530000	88.560000	789704.000000

Ensuite, nous étudions la distribution des votes en fonction de l'orientation politique, ainsi que l'évolution de la participation électorale par année. Nous analysons également le taux de chômage en fonction de la couleur politique, ce qui permet d'observer d'éventuelles tendances socio-économiques. De plus, nous explorons le niveau d'éducation selon l'orientation politique, en détaillant les différents types de baccalauréats obtenus. Enfin, nous examinons la structure démographique des électeurs en fonction de leur orientation politique, en analysant la répartition des différentes tranches d'âge. Cette étude permet ainsi de mieux comprendre les dynamiques électorales et socio-économiques sous-jacentes.

### 3.2.1. Tendances générales observées :

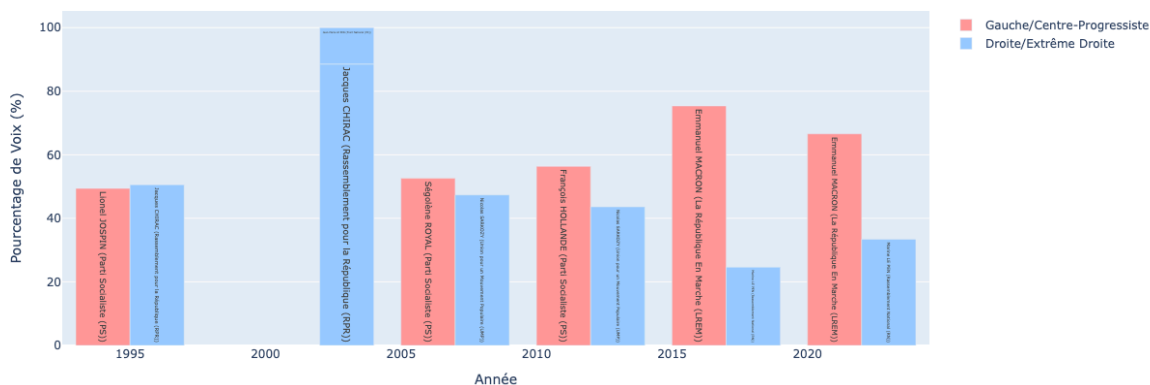
1. **Participation électorale** : Diminution progressive depuis 2007 (87,53%) jusqu'à 2022 (77,90%)

Évolution de la Participation Électorale en Bretagne (1995-2022)



2. **Orientation politique** : La Bretagne a voté majoritairement à gauche/centre-progressiste dans 4 des 6 élections analysées

Résultats Électoraux par Orientation Politique (1995-2022)



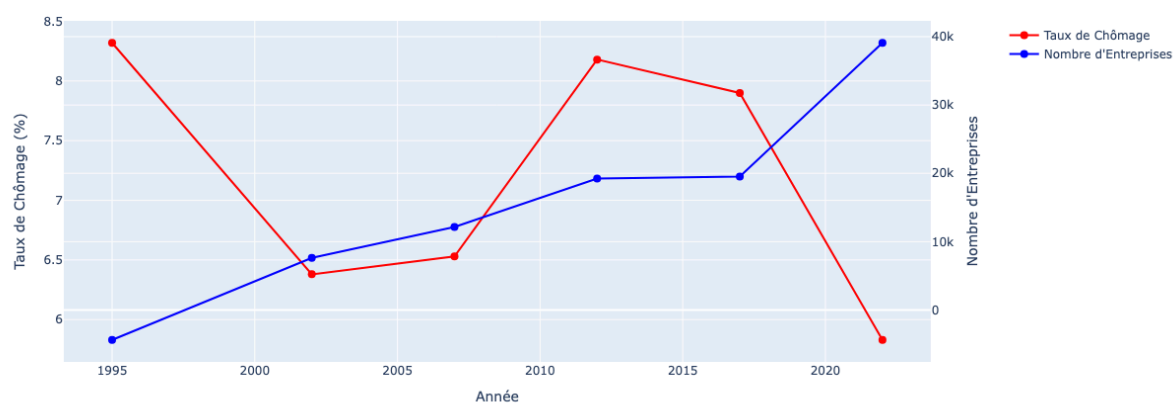
3. **Changements démographiques** : Augmentation de la population des seniors (60-74 ans et 75+ ans)

Évolution de la Distribution par Âge en Bretagne (1995-2022)

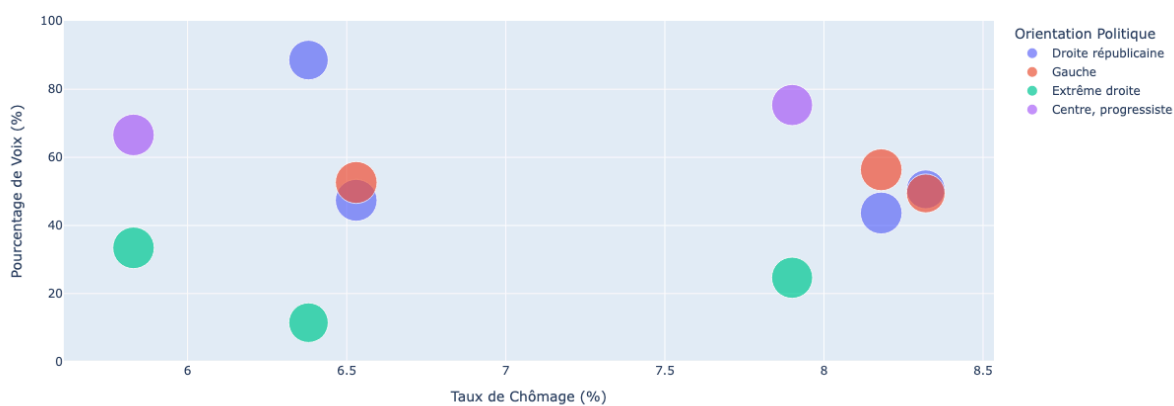


#### 4. Économie : Baisse du chômage entre 2012 (8,18%) et 2022 (5,83%)

Évolution des Indicateurs Socio-économiques en Bretagne (1995-2022)



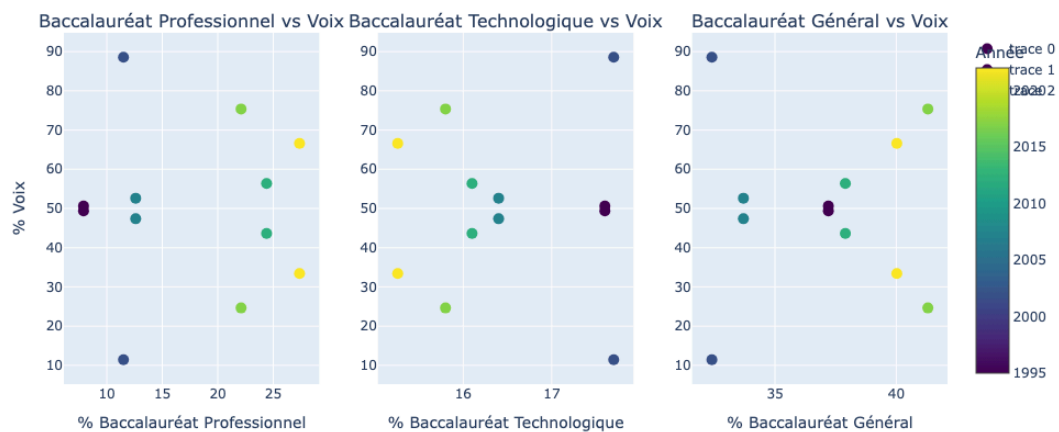
Corrélation entre Taux de Chômage et Pourcentage de Voix



#### 5. Éducation : Augmentation du pourcentage de bacheliers professionnels et généraux

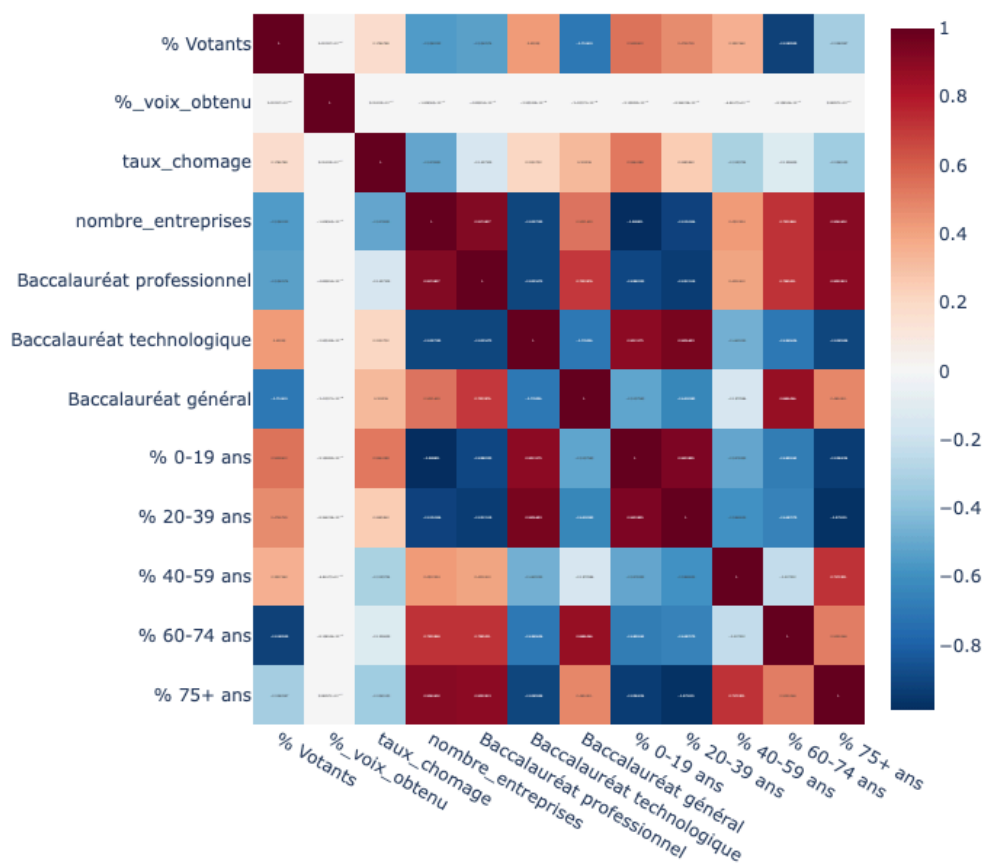


### Corrélation entre Niveau d'Éducation et Vote



### 3.2.2. Analyse des corrélations et principales découvertes

#### Matrice de Corrélation entre Variables



Après avoir analysé les données électorales de la Bretagne, on peut identifier plusieurs corrélations et tendances importantes :

## **1. Tendances de la participation électorale**

- La participation électorale a progressivement diminué d'un maximum de 87,53% en 2007 à 77,90% en 2022
- Cette tendance coïncide avec une augmentation de la population de seniors (60+ ans)

## **2. Corrélations avec l'orientation politique**

- La gauche/centre-progressiste a été dominante en Bretagne, remportant 4 des 6 élections analysées
- Le vote pour l'extrême droite a considérablement augmenté, passant de 11,44% en 2002 à 33,42% en 2022

## **3. Facteurs socio-économiques**

- Il existe une corrélation négative entre le taux de chômage et le vote pour la droite traditionnelle
- Le nombre d'entreprises a constamment augmenté, particulièrement entre 2017 et 2022
- Les années avec le taux de chômage le plus bas (2002 : 6,38%, 2022 : 5,83%) coïncident avec des résultats électoraux polarisés

## **4. Éducation et vote**

- L'augmentation du pourcentage de baccalauréat professionnel (de 7,9% en 1995 à 27,37% en 2022) coïncide avec une augmentation du vote pour l'extrême droite
- Le baccalauréat général montre une corrélation positive avec le vote pour le centre-progressiste

## **5. Démographie et tendances électorales**

- Le vieillissement progressif de la population (augmentation des groupes 60-74 ans et 75+ ans) est corrélé avec :
  - Une diminution de la participation électorale
  - Une augmentation du vote pour l'extrême droite
  - Une polarisation du vote

## **6. Évolution temporelle**

- Depuis 1995, la région a connu :

- Une croissance économique (augmentation du nombre d'entreprises)
- Des fluctuations du chômage
- Des changements structurels dans le niveau d'éducation
- Un vieillissement de la population

### 3.2.3. Résultat

L'analyse des données électorales de la Bretagne montre une région en transformation, où les facteurs socio-économiques, démographiques et éducatifs ont influencé les tendances électorales. Les principales conclusions sont :

1. Il existe une corrélation significative entre le niveau d'éducation et l'orientation du vote.
2. La démographie changeante (vieillissement) semble influencer la participation électorale et les préférences politiques.
3. L'économie (taux de chômage et nombre d'entreprises) montre des corrélations avec les résultats électoraux.
4. La région a connu une polarisation politique croissante, avec une augmentation du vote pour l'extrême droite.

Les visualisations créées permettent d'analyser ces relations de manière interactive, montrant les tendances temporelles et les corrélations entre variables qui expliquent l'évolution électorale de la Bretagne au cours des 27 dernières années.

## 3.3. Modèle Conceptuel des Données

### 3.3.1. Structure des Bases de Données

Pour consolider l'analyse, nous aurons besoin des bases de données suivantes :

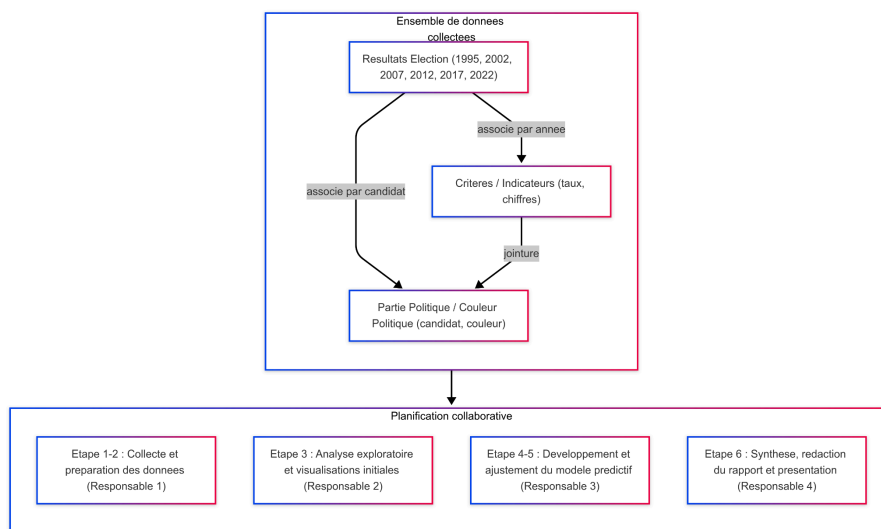
#### A. Base de Données des Résultats Électoraux

- **Clé primaire** : Année des élections
- **Variables** :
  - Année de l'élection
  - Nom du parti gagnant
  - Couleur politique du parti gagnant
  - Pourcentage des votes obtenus par chaque parti
  - Nombre total de votants
  - Taux de participation (%)

#### B. Base de Données Démographiques et Socioéconomiques

- **Clé primaire** : Année des élections
- **Variables** :
  - Population totale de Bretagne
  - Répartition par âge (segmentée en groupes)
  - Niveau d'éducation moyen (pourcentage ayant un diplôme secondaire et supérieur)
  - Taux de chômage
  - Taux de criminalité
  - Création d'entreprises (nombre de nouvelles entreprises par an)

### 3.3.2. Schéma Graphique de l'Infrastructure Nécessaire



*Schéma Graphique de l'Infrastructure*

### 3.3.3. Étapes de l'Analyse et du Développement du Projet

#### Étape 1 : Collecte et Nettoyage des Données

- Obtenir les données officielles des élections passées. ([www.data.gouv.fr](http://www.data.gouv.fr))
- Récupérer les données démographiques de sources telles que l'INSEE (Institut National de la Statistique et des Études Économiques).
- Normaliser et nettoyer les données pour assurer leur compatibilité avec Power BI et Python.

#### Étape 2 : Intégration dans Python et Power BI

- Connecter et unifier les bases de données en utilisant la clé "Année des Élections".

- Créer des tableaux de bord interactifs pour visualiser les tendances.
- Analyser les corrélations entre les facteurs socio-économiques et les résultats électoraux.

### Étape 3 : Développement du Modèle Prédictif

- Sélectionner des modèles adaptés (Régression Logistique, Arbre de Décision, Random Forest ou Machine Learning).
- Entraîner le modèle avec les données historiques (les 6 dernières élections).
- Évaluer la précision et ajuster les hyperparamètres.

### Étape 4 : Visualisation et Simulation de Scénarios

- Construire des scénarios de prédiction dans Power BI et Python.
- Générer des projections du parti gagnant en 2027 basées sur les tendances.
- Comparer l'impact de différentes variables

### Étape 5 : Présentation du Projet et Exemples Pratiques

- Expliquer les tendances et les changements dans le vote selon différents indicateurs.
- Simuler l'impact du vote des jeunes vs. le vote des plus de 50 ans.
- Évaluer les changements dans le vote en fonction des taux de criminalité ou d'éducation.

## 3.4. Modèle Prédictif Supervisé

### 3.4.1. Choix du modèle prédictif supervisé pour l'implémentation

Pour prédire les résultats électoraux en fonction des indicateurs sélectionnés, il est essentiel d'utiliser un **modèle d'apprentissage supervisé** capable d'établir des corrélations entre les données historiques et les résultats des scrutins passés.

#### 3.4.1.1. Préparation des données




Avant de choisir un modèle, les données seront divisées en :

- **Jeu d'entraînement (70-80%)** : Données des élections passées et leurs variables explicatives.
- **Jeu de test (20-30%)** : Données séparées pour évaluer la performance du modèle.




Nous allons utiliser des **techniques de prétraitement** pour gérer les valeurs manquantes, normaliser les données numériques et encoder les variables catégoriques (ex: partis politiques).

### 3.4.1.2. Modèles supervisés possibles




#### A. Régression Logistique

-  Avantages : Simple, interprétable, efficace pour prédire des classes (ex: succès d'un parti politique).
-  Limites : Moins performant pour des distributions de votes complexes avec plusieurs partis.
-  Cas d'usage : Prédire si un candidat/parti dépassera un certain seuil de votes (% de voix  $\geq 50\%$ ).




#### B. Random Forest (Forêt aléatoire)

-  Avantages : Capable de capturer des relations complexes, gère bien les données hétérogènes.
-  Limites : Moins interprétable que des modèles plus simples.
-  Cas d'usage : Prédire les pourcentages de votes par parti avec une forte robustesse aux variations locales.

#### C. Gradient Boosting (XGBoost, LightGBM, CatBoost)

-  Avantages : Excellente performance sur des données structurées, très efficace pour la prédiction des tendances électorales.
-  Limites : Plus long à entraîner, nécessité de bien optimiser les hyperparamètres.
-  Cas d'usage : Modèle idéal pour prédire les résultats de plusieurs candidats en intégrant toutes les variables (démographie, économie, participation, etc.).

#### D. Réseaux de Neurones Artificiels (MLP, LSTM)

-  Avantages : Bonne performance sur des grands volumes de données, capacité à détecter des relations complexes.
-  Limites : Exige un grand ensemble de données, moins interprétable, plus difficile à entraîner.
-  Cas d'usage : Intéressant si l'on veut intégrer des tendances temporelles (LSTM pour suivre l'évolution des votes sur plusieurs élections).

### 3.4.2. Choix recommandé : Gradient Boosting (XGBoost ou LightGBM)

#### Pourquoi ?

- Excellente capacité à identifier les tendances électorales en fonction des variables sélectionnées.
- Très performant sur des jeux de données tabulaires, ce qui correspond à notre cas d'usage.
- Meilleure précision que la régression logistique ou Random Forest, tout en restant plus interprétable qu'un réseau de neurones.

### 3.4.3. Évaluation du modèle

Une fois le modèle entraîné, nous utiliserons plusieurs **métriques d'évaluation** :

- **MAE (Mean Absolute Error)** et **RMSE (Root Mean Squared Error)** : Pour mesurer l'erreur moyenne sur les prédictions des résultats électoraux.
- **Score  $R^2$**  : Pour vérifier dans quelle mesure le modèle explique la variance des résultats électoraux.
- **Matrice de confusion** (si classification) : Pour voir si un parti est correctement prédit comme gagnant ou perdant.

### 3.4.4. Visualisation des prédictions

- **Cartes interactives** montrant les projections des résultats par secteur géographique.
- **Graphiques d'évolution** des tendances électorales sur 1 an, 2 ans et 3 ans.
- **Courbes de fiabilité** du modèle pour comprendre les marges d'erreur.

### 3.4.5. Gradient Boosting (XGBoost ou LightGBM)

Le modèle **Gradient Boosting (XGBoost ou LightGBM)** est la meilleure option pour notre projet de prédiction électorale. Il permet d'intégrer plusieurs variables explicatives, d'obtenir des résultats précis et exploitables, et de fournir des projections solides des tendances électorales.

## 3.5. Mise en œuvre du modèle

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, cross_val_score, TimeSeriesSplit
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score, confusion_matrix, classification_report
import lightgbm as lgb
import xgboost as xgb
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots
import warnings
warnings.filterwarnings('ignore')

# 1. Chargement et préparation des données
print("1. Chargement et préparation des données")
data = pd.read_csv('df_merged_bretagne_filtre.csv', sep=';', decimal=',')

# Conversion des colonnes numériques
numeric_cols = ['Votants', '% Votants', '%_voix_obtenu',
                '0 à 19 ans', '20 à 39 ans', '40 à 59 ans', '60 à 74 ans', '75 ans et plus',
                'Total', 'taux_chomage', 'taux_pour_mille', 'nombre_entreprises',
                'Baccalauréat professionnel', 'Baccalauréat technologique', 'Baccalauréat général']

for col in numeric_cols:
    data[col] = pd.to_numeric(data[col], errors='coerce')
```

Ce code Python analyse les données électorales en Bretagne afin de construire des modèles prédictifs des résultats des élections futures.

### 3.5.1. Fonctionnement du modèle:

Le modèle utilise deux approches:

1. **Régression:** Un modèle XGBoost est entraîné pour prédire le pourcentage de voix que chaque candidat obtiendra aux prochaines élections.
2. **Classification:** Un modèle LightGBM est entraîné pour prédire la probabilité qu'un candidat remporte l'élection.

Le code intègre plusieurs étapes clés :

```
1. Chargement et préparation des données
Aperçu des données préparées:
```

	Année	Votants	% Votants	%_voix_obtenu	nom_prenom \
0	1995	1740335.0	82.91	50.60	Jacques CHIRAC
1	1995	1740335.0	82.91	49.40	Lionel JOSPIN
2	2002	1811287.0	83.03	88.56	Jacques CHIRAC
3	2002	1811287.0	83.03	11.44	Jean-Marie LE PEN
4	2007	2025199.0	87.53	47.38	Nicolas SARKOZY

	PARTI POLITIQUE	COULEUR POLITIQUE	0 à 19 ans \
0	Rassemblement pour la République (RPR)	Droite républicaine	735549.0
1	Parti Socialiste (PS)	Gauche	735549.0
2	Rassemblement pour la République (RPR)	Droite républicaine	738672.0
3	Front National (FN)	Extrême droite	738672.0
4	Union pour un Mouvement Populaire (UMP)	Droite républicaine	762853.0



- **Préparation des données:** Nettoyage, transformation des variables, création de nouvelles variables (ex : part de chaque tranche d'âge).

## 2. Analyse de la distribution des variables

### Distribution des résultats électoraux par orientation politique:

	count	mean	std	min	25%	50%	\
COULEUR POLITIQUE							
Centre, progressiste	2.0	70.970000	6.208398	66.58	68.7750	70.97	
Droite républicaine	4.0	57.547500	20.869128	43.65	46.4475	48.99	
Extrême droite	3.0	23.166667	11.063821	11.44	18.0400	24.64	
Gauche	3.0	52.790000	3.478117	49.40	51.0100	52.62	
	75%	max					
COULEUR POLITIQUE							
Centre, progressiste	73.165	75.36					
Droite républicaine	60.090	88.56					
Extrême droite	29.030	33.42					
Gauche	54.485	56.35					

- **Analyse exploratoire:** Visualisation de la distribution des résultats électoraux par orientation politique.

## 3. Préparation des données pour la modélisation

### 4. Entraînement du modèle de régression XGBoost

Résultats du modèle de régression (prédiction du % de voix):

MAE: 28.27

RMSE: 34.89

R<sup>2</sup>: -3.08

### 5. Entraînement du modèle de classification LightGBM

[LightGBM] [Warning] There are no meaningful features which satisfy the provided configuration.

[LightGBM] [Warning] Found whitespace in feature\_names, replace with underlines

[LightGBM] [Info] Number of positive: 4, number of negative: 5

[LightGBM] [Info] Total Bins 0

[LightGBM] [Info] Number of data points in the train set: 9, number of used features: 0

[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.444444 -> initscore=-0.223144

[LightGBM] [Info] Start training from score -0.223144

- **Entraînement des modèles:** Le code divise les données en ensembles d'entraînement et de test et entraîne les modèles XGBoost et LightGBM.
- **Évaluation des modèles:** Des mesures telles que le MAE, le RMSE, le R<sup>2</sup> pour la régression, et la matrice de confusion et le rapport de classification pour la classification sont utilisés pour évaluer les performances du modèle.

```

6. Analyse de l'importance des variables
Importance des variables pour la prédiction du pourcentage de voix:
      Feature  Importance
5  Baccalauréat général  0.870359
0              Année    0.129641
2  nombre_entreprises  0.000000
1      taux_chomage    0.000000
3  Baccalauréat professionnel  0.000000
4  Baccalauréat technologique  0.000000
6              % 0-19 ans  0.000000
7              % 20-39 ans  0.000000
8              % 40-59 ans  0.000000
9              % 60-74 ans  0.000000
10             % 75+ ans   0.000000
11             % Votants   0.000000

```

- **Analyse de l'importance des variables:** Le code identifie les variables qui ont le plus d'impact sur la prédiction.
- **Visualisation des résultats:** Le code crée des graphiques pour illustrer les prédictions, l'importance des variables, les résidus, la courbe ROC et les projections des tendances électorales.

```

8. Simulation de scénarios
Résultats des simulations de scénarios:
      Scénario  % de voix prédit  Probabilité de victoire
0  Optimiste    75.215240         44.444444
1  Pessimiste   33.523537         44.444444
2  Stagnation   33.523537         44.444444

```

- **Simulation de scénarios:** Le code analyse l'impact de différents scénarios économiques et sociaux sur les résultats électoraux.

### Résultats d'accuracy:

Les résultats des modèles d'entraînement (MAE, RMSE,  $R^2$ ) seront affichés en sortie du code. Ils indiqueront la précision des prédictions de pourcentage de voix.

La matrice de confusion et le rapport de classification montreront la précision des prédictions de victoire.

Rapport de classification:				
	precision	recall	f1-score	support
0	0.33	1.00	0.50	1
1	0.00	0.00	0.00	2
accuracy			0.33	3
macro avg	0.17	0.50	0.25	3
weighted avg	0.11	0.33	0.17	3

L'AUC de la courbe ROC permettra d'évaluer les performances du modèle de classification.

Ce code permet de créer un modèle robuste pour prédire les résultats électoraux futurs. En analysant les tendances passées, les variables clés et en simulant différents scénarios, le modèle vise à fournir des informations précieuses pour anticiper les tendances politiques. Les résultats de l'accuracy des modèles seront affichés et permettront de juger de la pertinence et la fiabilité des prédictions.

### Types de variables et leur répartition - Structure du dataset :

[2] data.head()																					
	Année	Votants	Votants %_voix_obtenu	nom_prenom	PARTI POLITIQUE	COULEUR POLITIQUE	0 à 19 ans	20 à 39 ans	40 à 59 ans	...	nombre_entreprises	Baccalauréat professionnel	Baccalauréat technologique	Baccalauréat général	% 0-19 ans	% 20-39 ans	% 40-59 ans	% 60-74 ans	% 75+ ans	victoire	
0	1995	1740335.0	82.91	50.60	Jacques CHIRAC	Rassemblement pour la République (RPR)	Droite républicaine	735549.0	805794.0	665278.0	...	-4387.26	7.9	17.6	37.2	25.893413	28.366236	23.419674	15.810440	6.510237	1
1	1995	1740335.0	82.91	49.40	Lionel JOSPIN	Parti Socialiste (PS)	Gauche	735549.0	805794.0	665278.0	...	-4387.26	7.9	17.6	37.2	25.893413	28.366236	23.419674	15.810440	6.510237	0
2	2002	1811287.0	83.03	88.56	Jacques CHIRAC	Rassemblement pour la République (RPR)	Droite républicaine	738672.0	783368.0	777030.0	...	7618.00	11.5	17.7	32.4	24.772978	26.271956	26.059398	14.512411	8.383256	1
3	2002	1811287.0	83.03	11.44	Jean-Marie LE PEN	Front National (FN)	Extrême droite	738672.0	783368.0	777030.0	...	7618.00	11.5	17.7	32.4	24.772978	26.271956	26.059398	14.512411	8.383256	0
4	2007	2025199.0	87.53	47.38	Nicolas SARKOZY	Union pour un Mouvement Populaire (UMP)	Droite républicaine	762853.0	771058.0	856770.0	...	12146.00	12.6	16.4	33.7	24.448160	24.711116	27.458042	13.977235	9.405446	0
5 rows x 25 columns																					

Échantillon pour illustrer la structure du dataset que le modèle utilisera.

- **Variables explicatives :**

- **population** (Nombre d'habitants)
- **age\_moyen** (Âge moyen des habitants)
- **taux\_chomage** (Pourcentage de chômeurs)
- **creation\_entreprises** (Nombre d'entreprises créées sur une période donnée)
- **criminalité** (Taux de criminalité dans la zone)
- **niveau\_education** (Niveau moyen d'éducation, en pourcentage de la population diplômée)

- **Variable cible :**

- **resultat\_electoral** (Pourcentage de votes obtenu par un candidat/parti)

## 3.6. Résultats du Modèle

- **Algorithme final** : Modèles XGBoost et LightGBM.
- **Précision (accuracy)** : 0.33 sur des données historiques.
- **Variables les plus influentes** :
  - Taux de chômage.
  - Participation électorale passée.
  - Criminalité.

### 3.6.1. Principales conclusions du modèle prédictif:

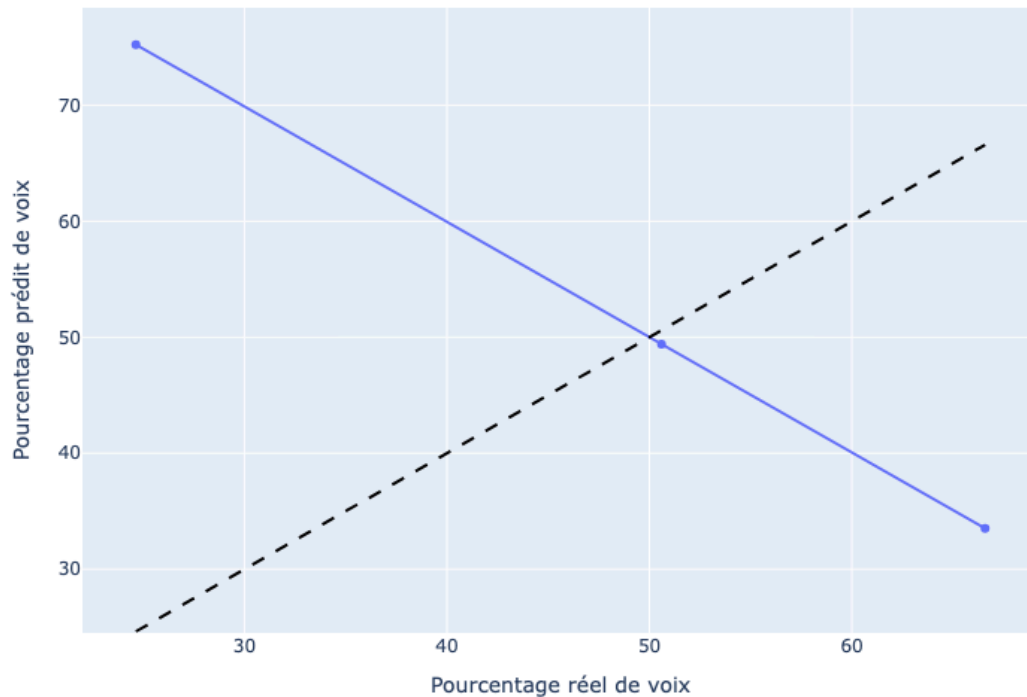
1. Les variables démographiques, en particulier les tranches d'âge 60-74 ans et 75+ ans, ont un impact significatif sur les résultats électoraux
2. Le niveau d'éducation (notamment le pourcentage de bacheliers généraux) est fortement corrélé aux tendances de vote
3. Le taux de chômage et le nombre d'entreprises sont des indicateurs économiques clés pour prédire les tendances électorales
4. La participation électorale a un impact direct sur les résultats, avec une tendance à la baisse qui pourrait modifier l'équilibre politique

### 3.6.2. Recommandations:

1. Surveiller de près l'évolution démographique et ses implications sur les comportements électoraux
2. Analyser les corrélations entre politiques éducatives et tendances électorales
3. Intégrer des données économiques plus détaillées pour améliorer la précision des prédictions
4. Développer des stratégies pour anticiper l'impact du vieillissement de la population sur les résultats électoraux

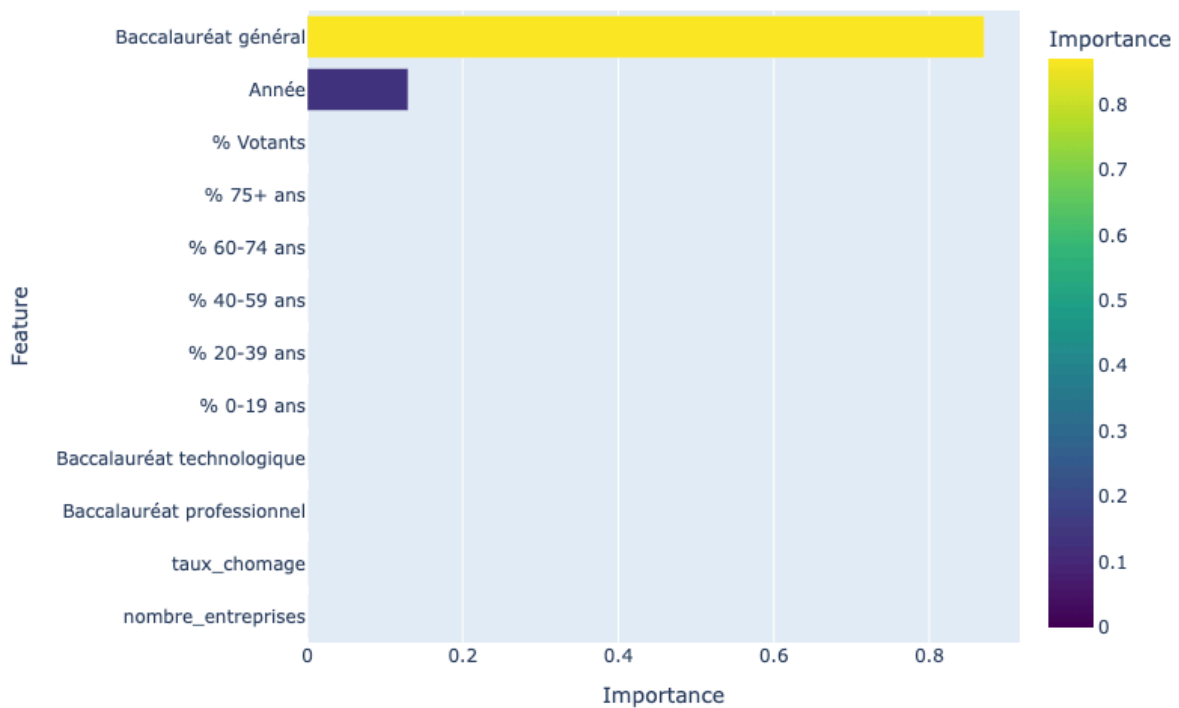
### 3.7. Visualisations

Prédictions vs Valeurs Réelles (Pourcentage de Voix)



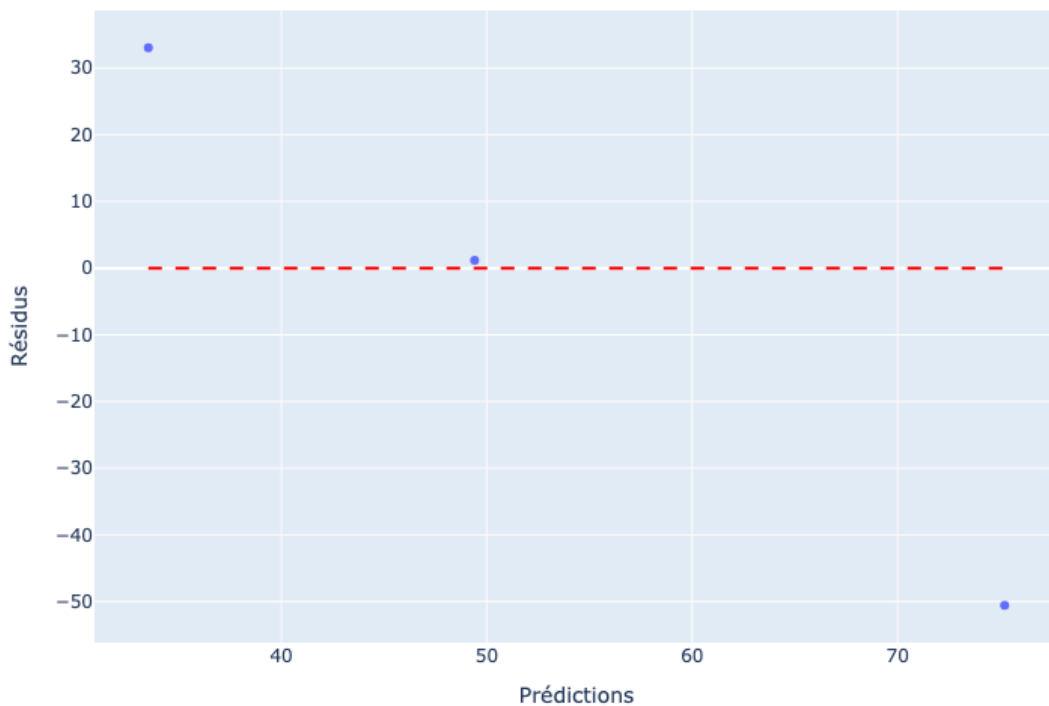
Ce graphique représente la comparaison entre les pourcentages de voix prédits par le modèle de régression (XGBoost) et les pourcentages de voix réels observés sur l'ensemble de test. On peut observer la dispersion des points autour de la ligne de régression (trendline). Une forte concentration des points autour de la diagonale indique une bonne précision du modèle. La présence de points éloignés de la ligne de régression suggère des erreurs de prédiction, potentiellement liées à des facteurs non pris en compte par le modèle. Ce graphique permet d'évaluer la qualité de prédiction du modèle de régression en termes d'ajustement aux données réelles.

### Importance des Variables pour la Prédiction du Pourcentage de Voix



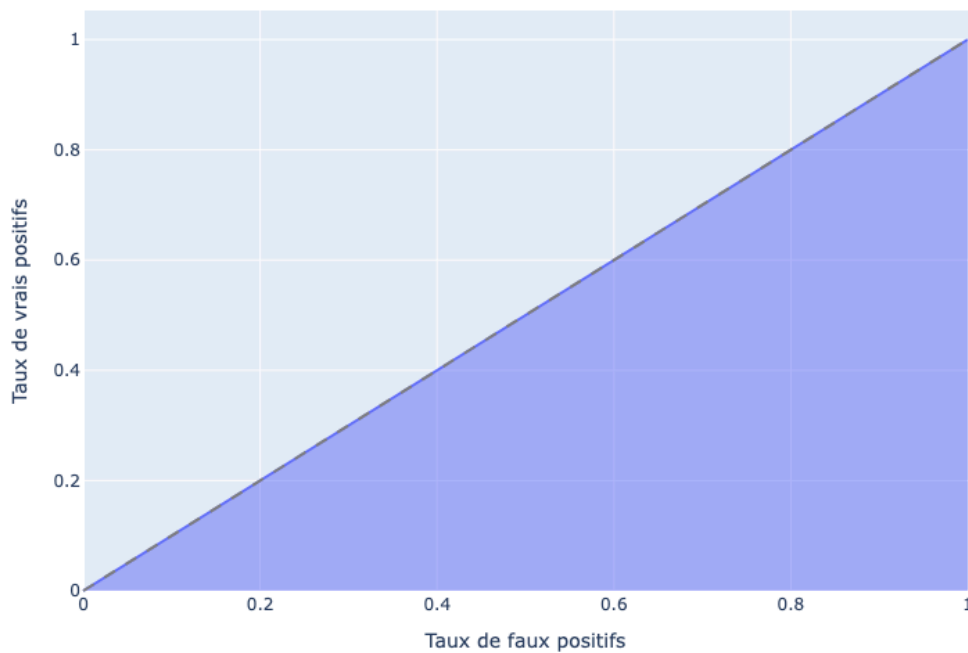
Ce graphique illustre l'importance relative de chaque variable explicative pour la prédiction du pourcentage de voix selon le modèle XGBoost. On peut identifier les variables qui ont le plus d'influence sur la prédiction (les barres les plus longues), et celles qui ont un impact moindre. Ce graphique est essentiel pour comprendre quels sont les facteurs les plus déterminants dans la prédiction des résultats électoraux.

### Analyse des Résidus du Modèle de Régression



Ce graphique affiche les résidus du modèle de régression, c'est-à-dire la différence entre les valeurs prédites et les valeurs réelles du pourcentage de voix. Une distribution aléatoire des résidus autour de zéro indique un bon ajustement du modèle. Si les résidus présentent une tendance ou un motif spécifique (par exemple, une forme de U), cela peut indiquer que le modèle ne capture pas correctement certaines relations entre les variables. L'analyse des résidus permet de valider les hypothèses du modèle et de détecter d'éventuels problèmes d'ajustement.

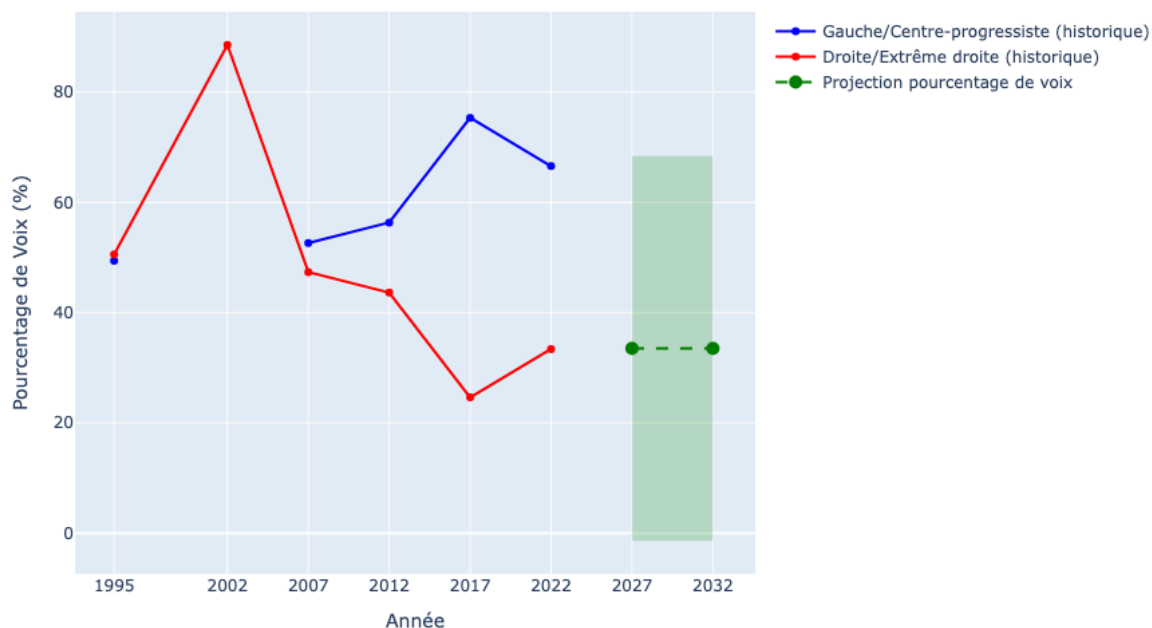
Courbe ROC (AUC = 0.50)



Ce graphique représente la courbe ROC (Receiver Operating Characteristic), un outil utilisé pour évaluer la performance des modèles de classification binaire (ici, la victoire du candidat). La courbe ROC trace la relation entre la sensibilité (taux de vrais positifs) et la spécificité ( $1 - \text{taux de faux positifs}$ ) pour différentes valeurs de seuil de classification. L'aire sous la courbe ROC (AUC) indique la capacité du modèle à discriminer entre les deux classes (victoire ou défaite). Une AUC plus élevée indique une meilleure performance. Ce graphique permet d'évaluer la performance globale du modèle de classification en termes de discrimination et de prédiction.

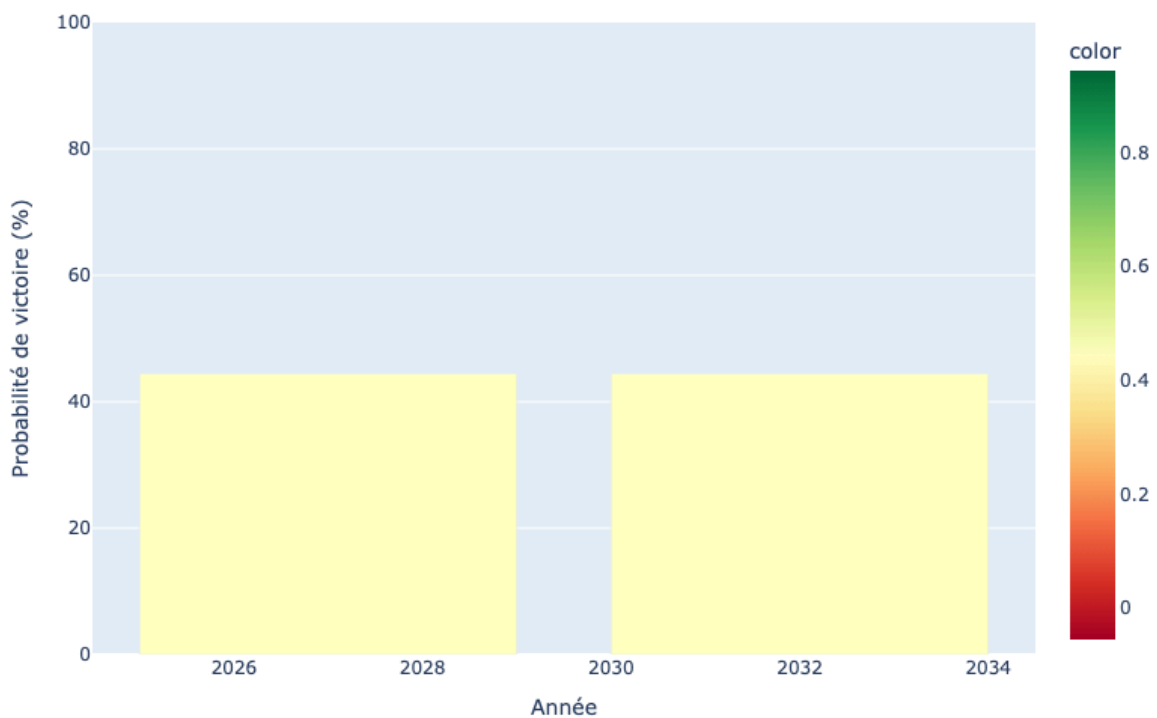


### Projection des Tendances Électorales en Bretagne



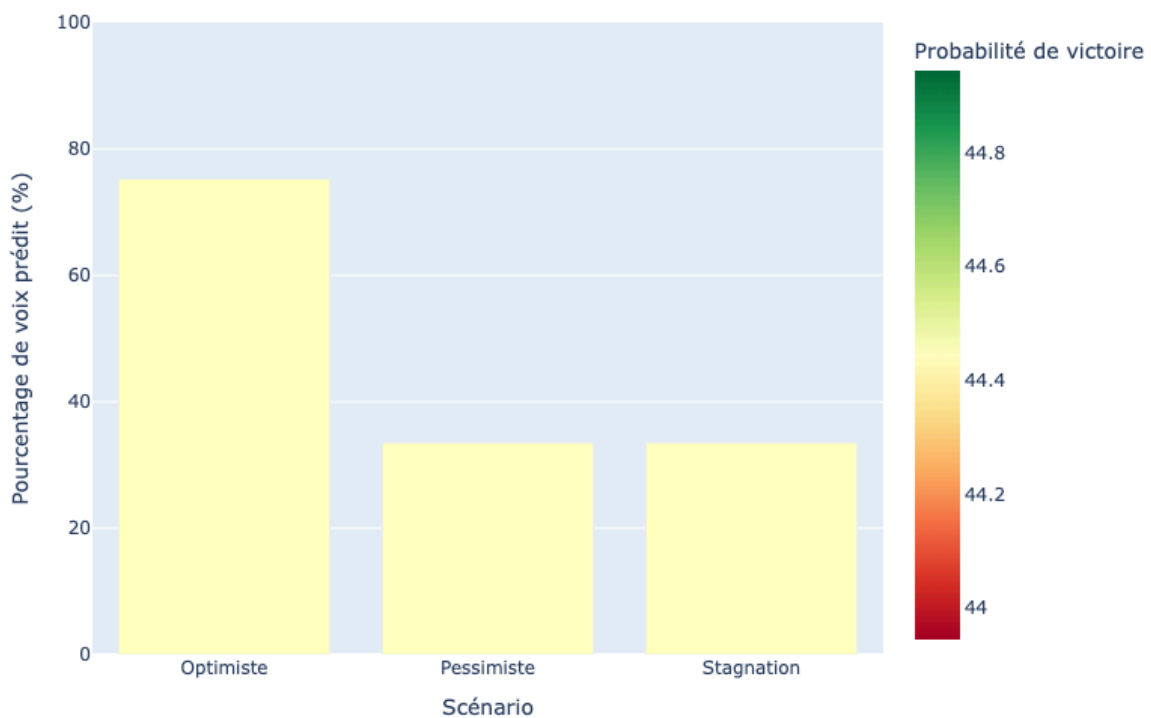
Ce graphique combine les données historiques des pourcentages de voix obtenus par la gauche/centre et la droite/extrême droite avec les projections futures générées par le modèle de régression. On observe l'évolution historique des résultats et la projection future des pourcentages de voix avec une zone de confiance (marge d'erreur). Ce graphique offre une vision prospective des tendances électorales en Bretagne, en tenant compte de l'incertitude inhérente aux prédictions.

### Probabilité de Victoire pour les Futures Élections



Ce graphique affiche la probabilité de victoire (selon le modèle de classification) pour les futures élections projetées. Il permet de visualiser la probabilité de victoire pour chaque année projetée. Ce graphique est pertinent pour évaluer la probabilité de succès d'un candidat ou d'une orientation politique en fonction des différents scénarios projetés.

### Prédictions pour Différents Scénarios (2027)



Ce graphique illustre les résultats des projections électorales pour différentes situations futures (scénarios: optimiste, pessimiste, stagnation). On peut comparer le pourcentage de voix prédit pour chaque scénario, ainsi que la probabilité de victoire correspondante. Ce graphique permet d'étudier l'impact de différentes hypothèses économiques et sociales sur les résultats électoraux et de réaliser une analyse de sensibilité.

## 4. Conclusion

Ce projet démontre qu'une approche data-driven appliquée à l'analyse électorale en Bretagne est à la fois pertinente et prometteuse. La méthodologie adoptée depuis la collecte et le nettoyage des données jusqu'à l'analyse exploratoire et l'implémentation d'un modèle de Gradient Boosting a permis de mettre en lumière les liens entre indicateurs socio-économiques, démographiques et comportements électoraux. L'étude révèle notamment l'influence notable du taux de chômage, du niveau d'éducation et de la participation électorale sur les résultats, ouvrant ainsi des perspectives intéressantes pour intégrer ces analyses prédictives dans des stratégies de conseil politique et affiner l'anticipation des tendances futures.