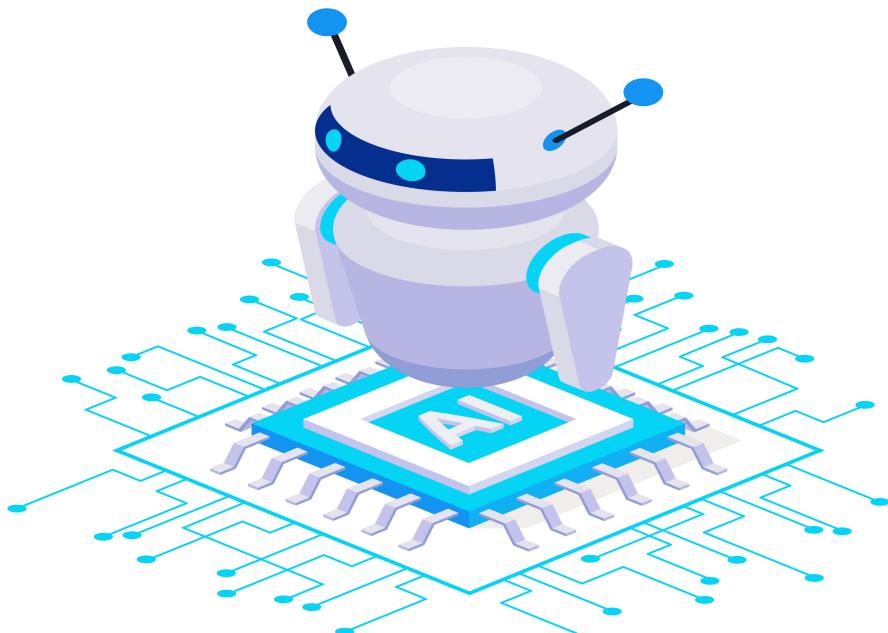


# MSPR BLOC 4



<https://github.com/jowill836/elections.git>

AGOURAM YASSAMINE

AKUETEVI WILSON

## Table des matières

<i>Table des matières</i> .....	<b>1</b>
<i>I. Introduction</i> .....	<b>2</b>
<i>II. Méthode de travail</i> .....	<b>2</b>
<i>III. Démarche suivie</i> .....	<b>3</b>
<i>IV. Recherche et sélection de bases de données</i> .....	<b>3</b>
<i>V. Constitution du dataset</i> .....	<b>4</b>
<i>VI. Pré-traitement power bi</i> .....	<b>7</b>
<i>VII. Normalisation du jeu de données</i> .....	<b>9</b>
<i>VIII. Visualisation des données</i> .....	<b>11</b>
<i>IX. MACHINE LEARNING</i> .....	<b>13</b>
1. Matrice de corrélation .....	<b>13</b>
2. Prédiction de parti politique .....	<b>13</b>
a. Modèles de machine Learning .....	13
b. Modèle retenu .....	14
<i>X. Technologies utilisées</i> .....	<b>15</b>
<i>XI. Conclusion</i> .....	<b>16</b>

## I. Introduction

Notre mission dans ce projet, est d'élaborer une preuve de concept (POC) pour la start-up de M. Jean-Edouard de la Motte Rouge, spécialisée dans le conseil en matière de campagnes électorales.

Le projet implique l'utilisation de divers indicateurs sociaux, environnementaux ou autres, notamment la sécurité, le climat, la population, l'économie ou d'autres facteurs susceptibles de refléter les choix électoraux.

Pour mettre en place cette démonstration de concept, nous passons par plusieurs étapes commençant par la recherche des données, passant par la préparation des données qui nous permettra de réaliser un modèle prédictif qui nous aidera à donner un résultat qui montre les tendances électORALES.

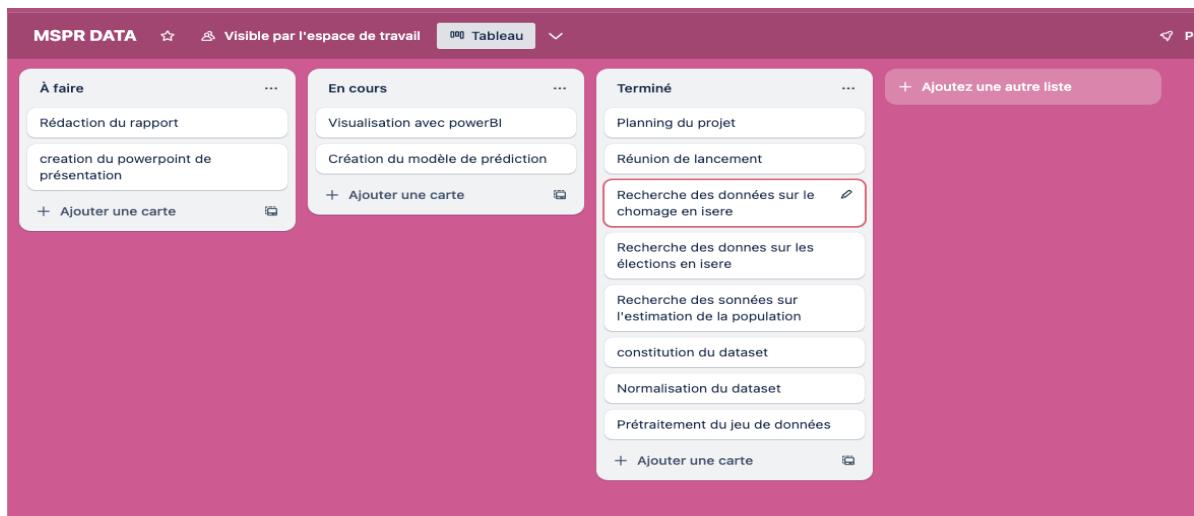
Dans ce rapport, nous mettons en œuvre le détail des différentes étapes du projet.

## II. Méthode de travail

Après avoir pris connaissance des exigences de la MSPR énoncées dans le cahier des charges, nous avons analysé les besoins et identifié les contraintes imposées. Notre objectif était de proposer une solution améliorée tout en respectant les délais imposés.

Dans un premier temps, nous avons examiné collectivement le cahier des charges, puis nous avons réparti les tâches et planifié le projet, suivant ainsi une méthodologie de travail structurée. Pour optimiser la gestion du temps et permettre une adaptation aisée aux éventuels changements et évolutions du projet, nous avons mis en œuvre une méthodologie agile pour la gestion des tâches des développeurs.

Toutes les tâches initialement définies ont été énumérées et transformées en user stories, assignées ensuite aux développeurs. Chaque développeur ou équipe de développement met à jour l'avancement de sa carte dans notre tableau de suivi de projet, comme illustré dans la figure ci-dessous. Cela offre une visibilité claire et précise sur les développements en cours, facilitant l'identification des éventuels blocages.



### **III. Démarche suivie**

Pour la réalisation de notre projet, voici en quelques lignes la méthodologie utilisée :

- Recherche dataset des élections présidentielles sur plusieurs années
- Recherche dataset d'indicateurs sociaux
- Traitement et nettoyage des jeux de données
- Visualisation des données grâce à powerBI
- Élaboration du modèle prédictif supervisé

### **IV. Recherche et sélection de bases de données**

La phase de sélection des données revêt une importance cruciale, visant à recueillir les informations les plus pertinentes pour l'ensemble du projet.

Pour donner suite à la demande de notre client, nous allons nous concentrer sur le département de l'Isère. Cette décision stratégique nous permettra de focaliser nos efforts sur une zone que nous connaissons assez, assurant ainsi la production de résultats pertinents pour notre client.

Notre objectif est de prédire les résultats des élections des partis politiques de gauche, de droit et du centre dans le département de l'Isère.

Afin de mener correctement notre projet, il est important de donner beaucoup d'importance à la recherche des données appropriées pour alimenter notre prédiction.

Dans l'ensemble des données que nous avons en main, il y a plusieurs irrégularités, notamment des différences de types de données, l'absence de certaines colonnes ...

Notre première mission serait donc d'identifier les incohérences et harmoniser l'ensemble des données en utilisant des colonnes partagées.

Dans le cadre de notre approche en matière de gestion des données, nous tenons à souligner que notre entreprise accorde une priorité particulière au respect des normes de sécurité et de confidentialité, en conformité avec le Règlement Général sur la Protection des Données (RGPD). Nous avons mis en place des protocoles de collecte, de stockage et de traitement des données qui intègrent un accès restreint aux informations sensibles, et des procédures de gestion des identités. De plus, notre méthodologie respecte les aspects juridiques, y compris les clauses contractuelles établies avec nos clients et fournisseurs.

Cette approche garantit non seulement la conformité avec les normes légales et réglementaires, mais également le renforcement de la confiance de nos partenaires et clients envers notre engagement envers la sécurité des données.

#### **Sources**

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

Chômage : <https://www.insee.fr/fr/statistiques/serie/001515903>

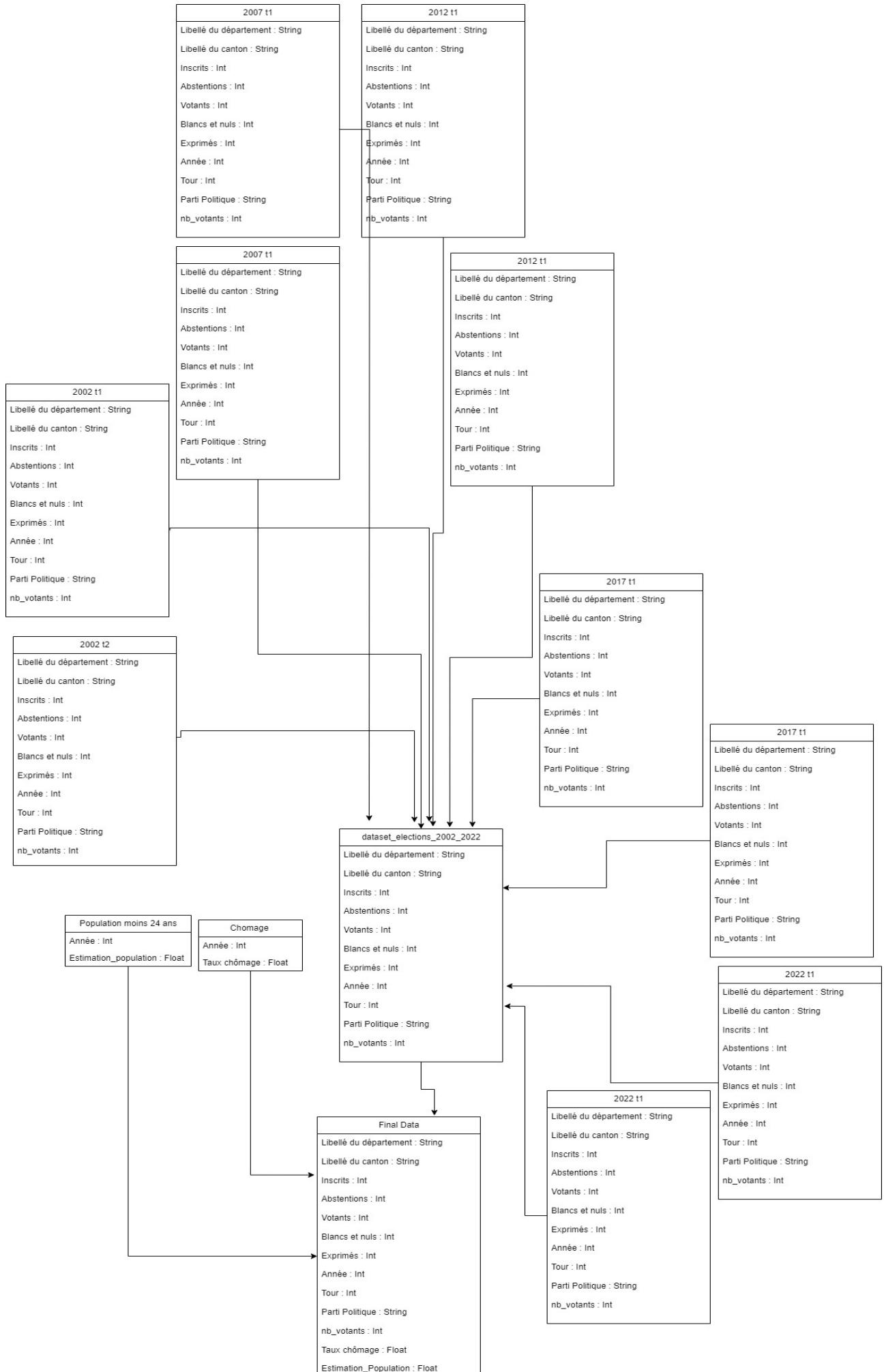
Estimation population -24 ans : <https://www.insee.fr/fr/statistiques/serie/001741269>

## V. Constitution du dataset

Nous avons agrégé des colonnes à partir de plusieurs ensembles de données en fonction des caractéristiques correspondantes, telles que la date ou le libellé du canton.

Ensuite, nous avons effectué des jointures entre ces divers ensembles de données afin de regrouper les informations.

Pour faciliter la compréhension de ce processus d'assemblage, nous avons créé un diagramme de classe représentant les différents tableaux de données. Ce diagramme permet de visualiser les éléments initialement sélectionnés et de comprendre le processus d'assemblage réalisé pour obtenir le jeu de données final, qui sera utilisé pour le modèle de prédiction.



On observe une structuration en Flocon. Tout comme dans l'organisation en étoile, on identifie des tables de faits telles que la table du chômage, population moins 24 ans, positionnement, et final data.

En ce qui concerne les données issues des élections présidentielles, une hiérarchisation distincte est perceptible, avec des sous-dimensions représentées par les diverses élections antérieures.

A noter que toutes les bases de données sont filtrées en sorte de garder uniquement les données concernant l'Isère.

Attribut	Description
Libellé du département	Nom officiel associé au département, dans notre cas, ce sera toujours l'Isère
Libellé du canton	Nom de la ville
Inscrits	Nombre total des inscrits
Abstentions	Nombre total des abstentions
Votants	Nombre total des votants
Blancs et Nuls	Nombre total des blancs et nuls
Exprimés	Nombre total des exprimés
Année	Année d'élection
Tour	Tour des élections
Parti Politique	Parti politique du candidat
Nb_votants	Nombre de votes associé à ce parti politique
Taux chômage	Taux de chômage en Isère
Estimation_population	Estimation de la population des moins de 24 ans

## VI. Pré-traitement power bi

Afin de créer notre table finale\_data nous avons procédé par plusieurs étapes commençant par l'extraction de nos données qui sont stockées sur le OneDrive de nos comptes EPSI, avec un accès sécurisé, passant par la transformation, c'est ce que nous allons détailler dans cette partie, finissant par le chargement de ces données, pour premièrement les visualiser mais aussi pour créer les modèles d'apprentissage.

Le processus de prétraitement des données vise principalement à évaluer la qualité des données. Cette évaluation de la qualité repose sur plusieurs critères, notamment :

- **Précision** : Nous avons vérifié l'exactitude des données saisies en examinant la concordance des résultats électoraux et du nombre d'électeurs.
- **Disponibilité** : Nous avons confirmé que les données étaient enregistrées légalement et accessibles en ligne, garantissant ainsi leur disponibilité.
- **Cohérence** : Nous avons vérifié que les données étaient enregistrées de manière uniforme en termes de format et de correspondances appropriées.
- **Mise à jour** : Nous avons assuré que les données étaient récemment mises à jour en obtenant des informations provenant des sources gouvernementales.
- **Crédibilité** : Avant d'utiliser les données, nous avons effectué des vérifications pour nous assurer de leur provenance officielle, garantissant ainsi leur fiabilité.
- **Interprétabilité** : Étant donné la diversité des données avec des structures et des colonnes variables, nous avons effectué plusieurs ajustements pour assurer une interprétation cohérente de l'ensemble des données.

Comme le schéma le montre, la table dataset\_elections\_2002\_2022 est constituée des données de 2002 jusqu'à 2022 des tours 1 et 2, donc le traitement des données est identique sur toutes les tables pour enfin les fusionner et avoir une table normalisée et cohérente.

Toutes les transformations se font en langage M de Power Query pour effectuer des opérations sophistiquées.

- Après le chargement des données, vient l'étape du renommage qui vise à donner des noms significatifs et unis entre les différentes tables, ceci permet aussi d'améliorer la lisibilité.
- Nous traitons uniquement les données du département "38" (Isère), permettant ainsi une analyse géographiquement ciblée, notre base de données contient tous les départements, donc nous filtrons uniquement sur le département de l'Isère.
- Des colonnes redondantes ou non nécessaires ont été supprimées, notamment les calculs de pourcentages ou encore le sexe du candidat pour simplifier la structure du tableau.
- Afin de distinguer les tours et les années nous avons créé les deux colonnes. Les résultats électoraux initialement distribués sur plusieurs colonnes, nous les avons pivoté pour regrouper les données avec une nouvelle structure "Parti Politique" et "nb\_votants" (Voir le détail sur la partie "Normalisation des données").
- Pour la finition, Nous avons été très attentionnés envers la cohérence des types de données, notamment la date qui était en type entier et aussi les noms des colonnes en respectant les majuscules et minuscules.

**Comme cité auparavant, les traitements sont identiques pour toutes les années.**

Concernant les tables "Estimations moins 24 ans" et "Chômage" nous avons gardé uniquement les données sur l'Isère et les années électorales.

Source	⚙
Navigation	⚙
En-têtes promus	⚙
Lignes filtrées	⚙
Colonnes renommées	
Colonnes supprimées1	
Personnalisée ajoutée	⚙
Personnalisée ajoutée1	⚙
Colonnes renommées1	
Colonnes supprimées	
Type modifié	
Tableau croisé dynamique de...	
Lignes triées	
Colonnes renommées2	
×	Type modifié1

## VII. Normalisation du jeu de données

Afin d'obtenir le jeu de données défini lors de notre sélection, nous avons opté pour l'utilisation de Power BI pour normaliser les formats variés des jeux de données sources des élections présidentielles, ainsi que pour effectuer une première analyse de la qualité des données obtenues. Au cours de cette démarche, nous avons constaté une incompatibilité des données générées lors de différentes élections, principalement en raison du regroupement différencié des votes par candidat. Cette disparité rendait toute analyse ultérieure impossible.

Pour remédier à cette situation, des requêtes Power Query ont été nécessaires pour préparer les données, étant donné que la mise en place d'un traitement plus automatisé n'était pas réalisable.

### Exemple :

A <sub>B</sub> <sub>C</sub> Nom	A <sub>B</sub> <sub>C</sub> Prénom	1 <sup>2</sup> <sub>3</sub> JOLY EVA	A <sub>B</sub> <sub>C</sub> Nom_2	A <sub>B</sub> <sub>C</sub> Prénom_3	1 <sup>2</sup> <sub>3</sub> LE PEN MARINE
JOLY	Eva		167 LE PEN	Marine	823
JOLY	Eva		206 LE PEN	Marine	2332
JOLY	Eva		220 LE PEN	Marine	1233
JOLY	Eva		328 LE PEN	Marine	2836
JOLY	Eva		73 LE PEN	Marine	266
JOLY	Eva		46 LE PEN	Marine	309
JOLY	Eva		209 LE PEN	Marine	2177



1 <sup>2</sup> <sub>3</sub> JOLY EVA	1 <sup>2</sup> <sub>3</sub> LE PEN MARINE
167	823
206	2332
220	1233
328	2836
73	266
46	309
209	2177
374	4620
671	3059
440	2589
197	2092
730	1494
619	886

Nous avons également associé les candidats aux parties correspondantes afin de véritablement identifier par la suite une continuité politique :

<sup>1<sup>2</sup></sup> 3 JOLY EVA	<sup>1<sup>2</sup></sup> 3 LE PEN MARINE	<sup>1<sup>2</sup></sup> 3 SARKOZY NICOLAS	<sup>1<sup>2</sup></sup> 3 MELENCHON JEAN-LUC	<sup>1<sup>2</sup></sup> 3 POUTOU PHILIPPE	<sup>1<sup>2</sup></sup> 3 ARTHAUD NATHALIE
167	823	1018	718	64	3
206	2332	2043	1019	104	5
220	1233	2313	890	84	3
328	2836	4023	1324	127	6
73	266	440	248	17	1
46	309	436	173	13	
209	2177	2328	889	105	5



<sup>1<sup>2</sup></sup> 3 EELV	<sup>1<sup>2</sup></sup> 3 FN	<sup>1<sup>2</sup></sup> 3 UMP	<sup>1<sup>2</sup></sup> 3 LFI	<sup>1<sup>2</sup></sup> 3 NPA	<sup>1<sup>2</sup></sup> 3 LO
167	823	1018	718	64	31
206	2332	2043	1019	104	58
220	1233	2313	890	84	38
328	2836	4023	1324	127	60
73	266	440	248	17	10
46	309	436	173	13	5

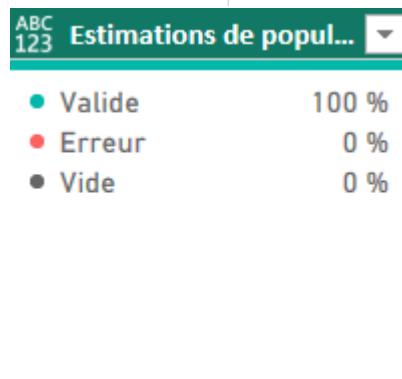
Une fois normalisé les jeux de données, nous avons pu regrouper les jeux de données sans erreurs

:

ABC 123 Année	ABC 123 Tour	ABC 123 Libellé du départeme...	ABC 123 Libellé du canton	ABC 123 Inscrits	ABC 123 Abstentions
● Valide 100 %	● Valide 100 %	● Valide 100 %	● Valide 100 %	● Valide 100 %	● Valide 100 %
● Erreur 0 %	● Erreur 0 %	● Erreur 0 %	● Erreur 0 %	● Erreur 0 %	● Erreur 0 %
● Vide 0 %	● Vide 0 %	● Vide 0 %	● Vide 0 %	● Vide 0 %	● Vide 0 %

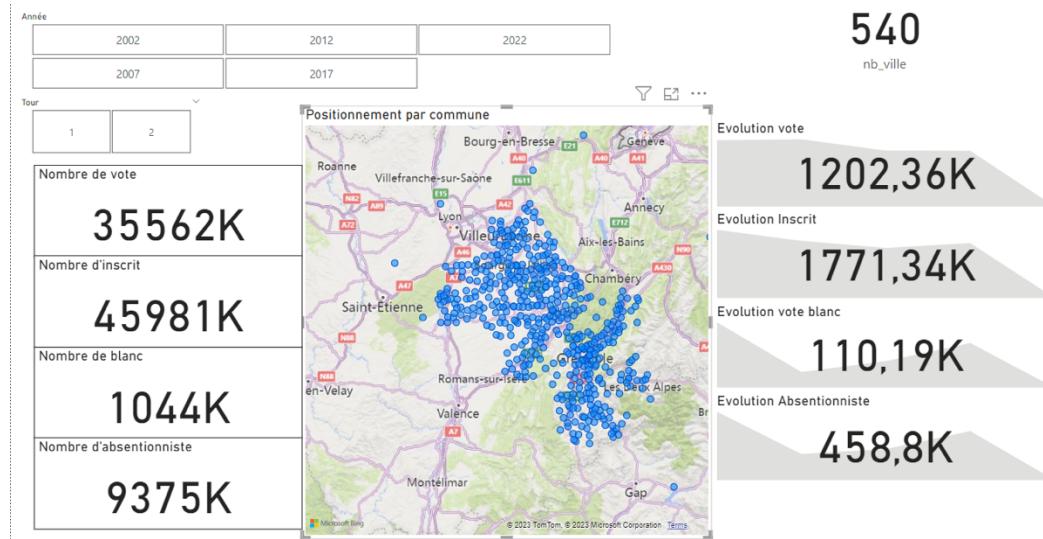
ABC 123 Votants	ABC 123 Blancs et nuls	ABC 123 Exprimés	ABC 123 Parti Politique	ABC 123 nb_votants	1.2 Chômage Isere
● Valide 100 %	● Valide 100 %	● Valide 100 %	● Valide 100 %	● Valide 100 %	● Valide 100 %
● Erreur 0 %	● Erreur 0 %	● Erreur 0 %	● Erreur 0 %	● Erreur 0 %	● Erreur 0 %
● Vide 0 %	● Vide 0 %	● Vide 0 %	● Vide 0 %	● Vide 0 %	● Vide 0 %



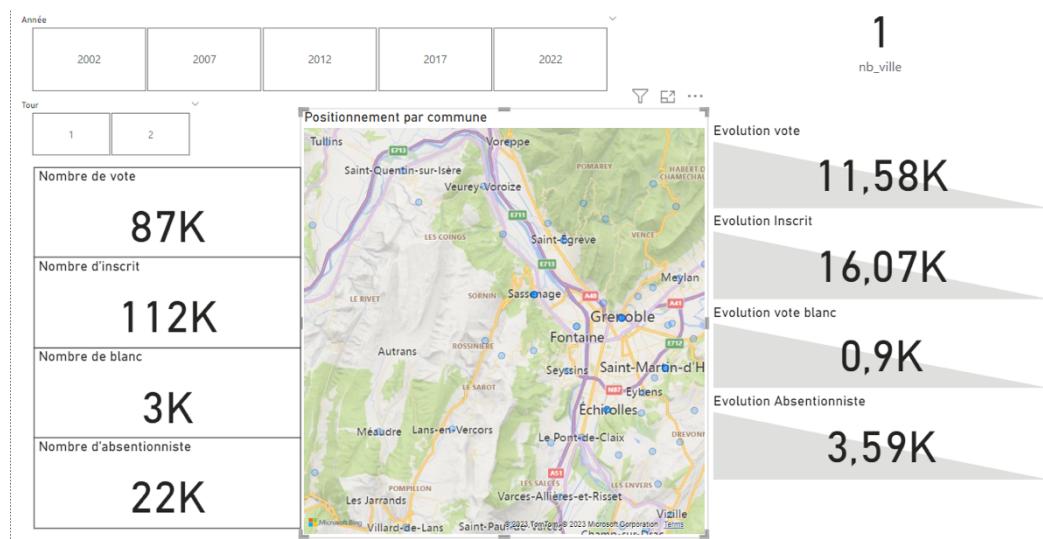
## VIII. Visualisation des données

Le tableau de bord Power BI actuel a été construit en utilisant des données normalisées.

- Sur la première page de ce rapport, la topologie de l'Isère est présentée avec divers niveaux de détail, offrant ainsi la possibilité d'analyser l'évolution des résultats à différentes granularités :



Le détail s'affiche par commune et année au clic :



- Une deuxième page est dédiée pour le détail :



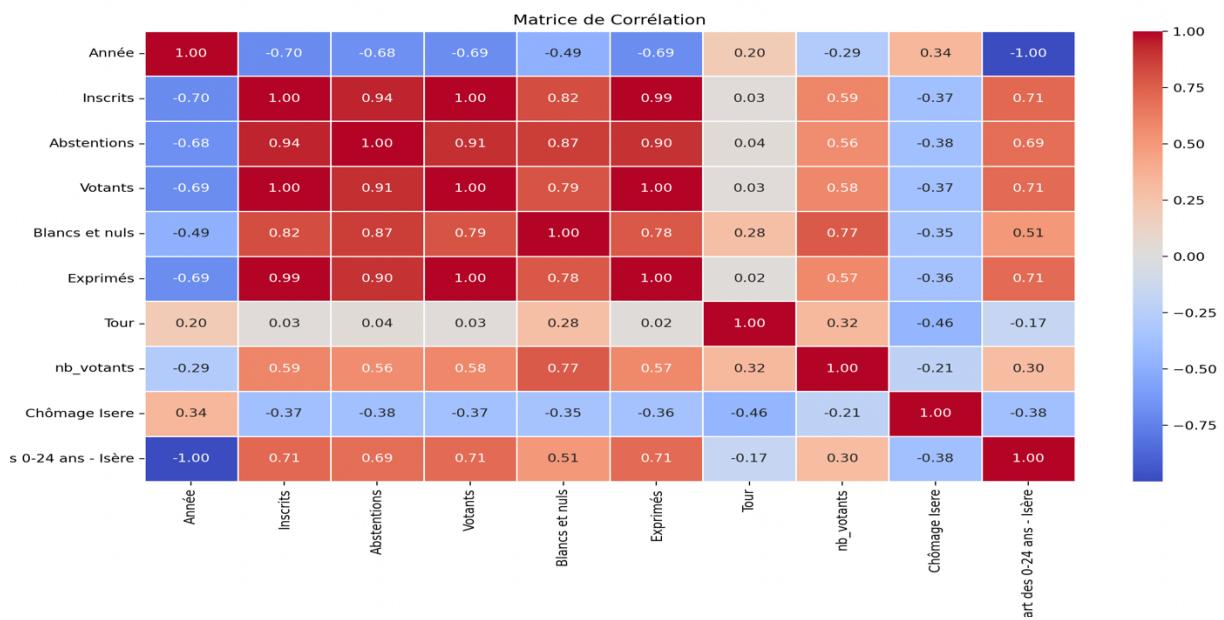
Année	Libellé du canton	Inscrits	Abstention	Votants	Blancs et nuls	Exprimés	Tour	Parti Politique	nb_votants	Chômage Isère	Estimations de population	Année
		s										2002
2022	Monestier-d'Ambel	22	5	17	2	15	2	FN	5	5,93		
2022	Monestier-d'Ambel	22	5	17	2	15	2	LREM	10	5,93		
2022	Ambel	33	6	27	1	26	2	FN	10	5,93		
2022	Ambel	33	6	27	1	26	2	LREM	16	5,93		
2022	Beaufin	34	1	33	7	26	2	FN	12	5,93		
2022	Beaufin	34	1	33	7	26	2	LREM	14	5,93		
2022	Villard-Reymond	34	4	30	7	23	2	FN	6	5,93		
2022	Villard-Reymond	34	4	30	7	23	2	LREM	17	5,93		
2022	Villard-Notre-Dame	38	5	33	3	30	2	FN	6	5,93		
2022	Villard-Notre-Dame	38	5	33	3	30	2	LREM	24	5,93		
2022	Oulles	39	9	30	4	26	2	FN	18	5,93		
2022	Oulles	39	9	30	4	26	2	LREM	8	5,93		
2022	Cognet	41	8	33	7	26	2	FN	15	5,93		
2022	Cognet	41	8	33	7	26	2	LREM	11	5,93		
2022	Sainte-Luce	55	16	39	4	35	2	FN	13	5,93		
2022	Sainte-Luce	55	16	39	4	35	2	LREM	22	5,93		
2022	La Valette	62	12	50	7	43	2	FN	18	5,93		
2022	La Valette	62	12	50	7	43	2	LREM	25	5,93		
2022	Malleval-en-Vercors	63	10	53	4	49	2	FN	15	5,93		
2022	Malleval-en-Vercors	63	10	53	4	49	2	LREM	34	5,93		
2022	Saint-Michel-en-Beaumont	65	14	51	1	50	2	FN	20	5,93		
2022	Saint-Michel-en-Beaumont	65	14	51	1	50	2	LREM	30	5,93		
2022	Marcieu	68	6	62	7	55	2	FN	31	5,93		
2022	Marcieu	68	6	62	7	55	2	LREM	24	5,93		
2022	Saint-Arey	75	8	67	13	54	2	FN	21	5,93		

## IX. MACHINE LEARNING

### 1. Matrice de corrélation

Avant de commencer toute étude sur notre jeu de données, nous avons commencé par réaliser une matrice de corrélation entre nos variables afin d'explorer les relations linéaires entre celles-ci et de ne garder que les variables pertinentes pour notre étude.

Ainsi, grâce à cette sélection de variables, nous avons pu mettre en évidence la corrélation forte entre l'estimation de la population des 0-24 ans et le nombre d'inscrit aux élections, le nombre de votant par élection, et le nombre de voix exprimés par élections.



### 2. Prédiction de parti politique

Dans cette partie, le but est de prédire les élections futures à partir des résultats des élections passées selon des facteurs qui sont le chômage et l'estimation de la population moins de 24 ans en Isère.

Pour répondre à ce besoin, nous créons une variable cible "True" pour indiquer le maximum de nb\_votants dans une certaine année, ville et tour.

Comme le modèle ne reconnaît pas les valeurs booléennes, nous les avons convertis en valeurs numériques, ainsi que Tour et Année.

Le jeu de données d'entraînement est donc composé de toutes les données des années de 2002 à 2012 et le jeu de données de test est composé des données des années 2017 et 2022.

#### a. Modèles de machine Learning

Dans notre cas, nous cherchons à prédire le parti politique avec le plus grand nombre de votants en fonction du taux de chômage et de l'estimation de la population des 0-24 ans. La variable que nous essayons de prédire (le parti politique) est une variable catégorielle.

La classification est le type de problème où nous assignons une étiquette ou une catégorie à chaque observation dans nos données, et dans notre contexte, ces catégories correspondent aux différents partis politiques.

Plus précisément, nous sommes en face d'un problème de classification multi classe car nous avons plusieurs partis politiques différents.

Les algorithmes de classification, tels que le Random Forest, SVM, KNN et Regression Logistique, sont conçus pour résoudre ce type de problème en tentant de modéliser la relation entre les caractéristiques de nos données et les différentes classes de partis politiques pour pouvoir prédire la classe correcte pour de nouvelles données.

Nous avons entraîné 4 classificateurs (Random Forest, SVM, KNN et Logistic Regression) en ajoutant l'option "balanced" afin d'enlever le déséquilibre.

Nous les avons tous évalués et répertoriés les résultats dans le tableau ci-dessous :

Modèle	Précision	Accuracy(exactitude)	recall	F1 - score
Random Forest	80,80	87,79	87,79	82,16
SVM	77,79	61,99	61,99	68,19
KNN	75,08	61,44	61,44	67,31
Regression logistique	89,31	87,82	87,82	82,13

### b. Modèle retenu

**Métrique retenue** : la précision

Dans le cadre de notre modèle de classification, un aspect crucial à évaluer est la métrique de précision. La précision mesure la proportion d'instances correctement prédictées parmi celles identifiées comme positives par le modèle, se calculant par :  $\text{précision} = \text{TP}/(\text{TP}+\text{FP})$

Le choix de la précision est motivé par :

- Impact des faux positifs : Étant donné que notre prédition concerne le "Parti Politique" et que des conséquences significatives peuvent découler d'erreurs de classification, minimiser les faux positifs est notre priorité.
- Distribution des classes : En cas de déséquilibre entre les classes comme dans notre cas, la précision offre une mesure plus détaillée des performances par rapport à l'accuracy.
- Interprétation précise : La précision, en mettant l'accent sur la qualité des prédictions positives, permet de comprendre de manière précise la capacité du modèle à identifier correctement les instances de la classe d'intérêt.

En termes de précision, la régression logistique est l'algorithme qui nous donne le meilleur résultat, c'est à dire : 89,31%

## X. Technologies utilisées

TECHNOLOGIE	JUSTIFICATION DU CHOIX
 Power BI	Microsoft propose Power BI, une suite d'outils dédiée à la business intelligence, spécifiquement conçue pour élaborer des rapports interactifs et des visualisations de données percutantes. Cette solution est particulièrement adaptée pour créer des rapports interactifs et des représentations visuelles de données impactantes, simplifiant ainsi la communication des résultats du modèle prédictif auprès des parties prenantes.
 Power Query	Power Query, intégré à la fois dans Power BI et Excel, se spécialise dans l'extraction et la transformation des données. Il occupe une fonction cruciale en préparant, nettoyant, fusionnant et ajustant les données pour répondre aux exigences spécifiques du projet, avant de les intégrer dans Power BI. Cette démarche simplifie le processus de création de visualisations et de rapports informatifs.
 python  pandas	Pandas est une bibliothèque Python puissante qui fournit des structures de données faciles à utiliser (comme les DataFrames) pour la manipulation des données tabulaires. Elle offre des fonctionnalités pour nettoyer, transformer et analyser les données, ce qui est crucial dans le processus de prétraitement des données avant l'entraînement des modèles.
	Scikit-learn offre une interface cohérente et unifiée pour une grande variété d'algorithmes d'apprentissage automatique. Cela simplifie le processus de sélection, d'entraînement et d'évaluation des modèles. Scikit-learn propose une large gamme d'algorithmes pour la classification, la régression, le clustering, la réduction de dimension, et etc
	Git est un système de contrôle de version distribué. Il permet de travailler hors ligne et facilite le suivi des modifications. GitHub facilite la collaboration entre les développeurs en fournissant des fonctionnalités telles que les pull requests, les issues, les commentaires et la gestion des branches. Github simplifie la coordination entre les membres de l'équipe et offre également un hébergement gratuit du code

## **XI. Conclusion**

Ce rapport met en lumière l'importance cruciale des représentations visuelles pour rendre les données plus accessibles et faciles à comprendre.

L'outil Power BI s'est avéré être essentiel dans le processus décisionnel, fournissant une assistance significative.

L'exploration approfondie et l'analyse des données ont dévoilé des tendances intéressantes et des corrélations significatives.

En utilisant des méthodes d'apprentissage automatique, nous avons développé un modèle prédictif supervisé capable de prédire les résultats électoraux pour un parti avec une précision relative de 83%.

En résumé, ce rapport marque une étape initiale réussie vers la création d'un système de prédiction électorale basé sur l'intelligence artificielle. Il offre des preuves solides de la viabilité du projet et ouvre la voie à des perspectives prometteuses pour son évolution future.

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