MSPR BIGDATA

Michaël ALTMAN / Yannis THOMAS / Fatima Ezzahra EL FILALI / Ibtissam DALOUCH / Mael LEGOUGE

EPSI Lyon | GROUPE i1 dev1

Table des matières

1. Justification du choix de la zone géographique

La région de Bourgogne a été fusionnée avec la région Franche-Comté en 2016 pour former la région Bourgogne-Franche-Comté. Cependant, Nous avons opté pour la région française de Bourgogne en raison de son caractère diversifié, abritant une population variée avec un éventail d'opinions et de perspectives. Cette dernière est composée de 4 départements :

- Yonne 89
- Côte d'or 21
- Nièvre 58
- Saône et Loire 71

Chacun de ces départements à ses particularités en termes de paysages, d'histoire, de culture et de traditions. Cependant, depuis la fusion des régions, la région Bourgogne-Franche-Comté a une administration et une gouvernance communes pour l'ensemble de son territoire :

- **Côte-d'Or (21)** : C'est le département le plus au nord de la région. Son chef-lieu est Dijon, la capitale historique de la Bourgogne. La Côte-d'Or est réputée pour ses vignobles, en particulier ceux de la Côte de Nuits et de la Côte de Beaune, qui produisent certains **des vins les plus célèbres du monde**, tels que le vin de Bourgogne.
- **Nièvre (58)** : Situé à l'ouest de la région, la Nièvre est un département rural avec de vastes étendues de paysages naturels préservés. Son chef-lieu est Nevers, une ville historique connue pour sa **cathédrale gothique**.
- **Saône-et-Loire (71)** : Il s'agit du département le plus au sud de la Bourgogne. Son chef-lieu est Mâcon, une ville située sur les rives de la Saône. Saône-et-Loire possède également des **vignobles renommés**, notamment ceux de la Côte Chalonnaise et du Mâconnais.
- **Yonne (89)** : À l'est de la région, l'Yonne est un département traversé par la rivière éponyme. Son chef-lieu est Auxerre, une ville médiévale célèbre pour sa cathédrale et ses maisons à colombages. L'Yonne abrite également des sites historiques tels que le village de **Vézelay**, classé au patrimoine mondial de **l'UNESCO**.

Afin d'obtenir un échantillon fournissant des informations plus détaillées, nous avons opté pour l'analyse des villes abritant des bureaux de vote. Cette approche nous permet de mettre en œuvre les critères d'analyse que vous découvrirez ci-dessous (2- Choix des critères, justification).

En ce qui concerne la politique et les élections récentes dans la région Bourgogne-Franche-Comté, nous avons enregistré les résultats des élections de 2022. Toutefois, il est important de noter que les informations les plus récentes pourraient différer de ce qui suit. Premier tour des élections présidentielles 2022

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Code du départe	Libellé du départ	Code de la circo	Libellé de la circonsc	Code de la comr	Libellé de la com	Code du b.vote	Inscrits	Abstentions	% Abs/Ins	Votants	% Vot/Ins	Blancs
2	1	Ain	4	4ème circonscription	1	L'Abergement-C	1	645	108	16,74	537	83,26	
3	1	Ain	5	5ème circonscription	2	L'Abergement-d	1	213	38	17,84	175	82,16	
4	1	Ain	5	5ème circonscription	4	Ambérieu-en-Bu	1	1129	266	23,56	863	76,44	
5	1	Ain	5	5ème circonscription	4	Ambérieu-en-Bu	2	1128	265	23,49	863	76,51	
6	1	Ain	5	5ème circonscription	4	Ambérieu-en-Bu	3	1213	246	20,28	967	79,72	
7	1	Ain	5	5ème circonscription	4	Ambérieu-en-Bu	4	933	224	24,01	709	75,99	
8	1	Ain	5	5ème circonscription	4	Ambérieu-en-Bu	5	1047	324	30,95	723	69,05	
9	1	Ain	5	5ème circonscription	4	Ambérieu-en-Bu	6	1152	278	24,13	874	75,87	
10	1	Ain	5	5ème circonscription	4	Ambérieu-en-Bu	7	1034	247	23,89	787	76,11	
11	1	Ain	5	5ème circonscription	4	Ambérieu-en-Bu	8	1129	228	20,19	901	79,81	
12	1	Ain	4	4ème circonscription	5	Ambérieux-en-D	1	1282	234	18,25	1048	81,75	
13	1	Ain	3	3ème circonscription	6	Ambléon	1	103	23	22,33	80	77,67	
14	1	Ain	5	5ème circonscription	7	Ambronay	1	1139	186	16,33	953	83,67	
15	1	Ain	5	5ème circonscription	7	Ambronay	2	892	152	17,04	740	82,96	
16	1	Ain	2	2ème circonscription	8	Ambutrix	1	554	80	14,44	474	85,56	
17	1	Ain	3	3ème circonscription	9	Andert-et-Condc	1	279	48	17,2	231	82,8	
18	1	Ain	3	3ème circonscription	10	Anglefort	1	792	168	21,21	624	78,79	
19	1	Ain	5	5ème circonscription	11	Apremont	1	287	44	15,33	243	84,67	
20	1	Ain	5	5ème circonscription	12	Aranc	1	283	62	21,91	221	78,09	
21	1	Ain	5	5ème circonscription	13	Arandas	1	125	29	23,2	96	76,8	
22	1	Ain	5	5ème circonscription	14	Arbent	1	1130	363	32,12	767	67,88	

resultat20221

2. Choix des critères, justification

Dans notre recherche visant à déterminer les critères ayant un véritable impact sur le choix d'un parti politique, nous avons entrepris une analyse des aspects susceptibles d'influencer cette décision. Nous avons donc adopté une approche méthodique en nous posant la question pour chaque parti : "Quelles sont les idées centrales de cette orientation politique (droite/gauche/centre) et qu'est-ce qui pourrait inciter un individu à lui accorder son vote ?"

Nous avons simplifié les partis politiques en deux catégories (droite et gauche, incluant leurs tendances extrêmes respectives) sans utiliser de noms spécifiques. Cette approche tient compte du fait que les candidats et les noms de partis évoluent au fil des années, tandis que les orientations politiques fondamentales restent les mêmes.

Nous avons pu relever deux thématiques où les partis de gauche et de droite affichent souvent des divergences d'opinions : l'emploi/chômage et la sécurité/criminalité.

D'une part, la question de l'emploi et du chômage, élément vital pour la prospérité économique et sociale d'un pays, occupe une place prépondérante dans les débats politiques. Les différentes formations abordent cette thématique de manière singulière, chacune proposant des solutions et des approches distinctes pour favoriser la création d'emplois, réduire le taux de chômage et améliorer les conditions de travail.

D'autre part, la sécurité et la criminalité représentent un autre enjeu central dans le choix politique. Les perspectives et les politiques proposées par les partis de gauche et de droite diffèrent souvent en ce qui concerne la manière de garantir la sécurité des citoyens et de lutter contre la criminalité. Ces divergences peuvent englober des approches variées allant de la prévention sociale à une répression plus stricte.

À terme nous réaliserons des statistiques descriptives pour chaque commune et année. Cette approche révélera des tendances et variations spécifiques à chaque zone, ainsi que des évolutions temporelles significatives

3. La démarche suivie et les méthodes employées

Démarche et méthodes :

Dans le cadre de notre démarche analytique, nous avons entrepris la collecte de données officielles à partir du site data.gouv.fr. Les éléments suivants ont été rassemblés dans le but d'effectuer une analyse approfondie :

- Résultats de l'élection présidentielle 2017 par bureau de vote.
- Résultats de l'élection présidentielle 2022 par bureau de vote.
- Statistiques relatives à la criminalité pour l'année 2022.
- Données concernant l'emploi pour les années 2017 et 2022.

Ces informations nous ont servi de base pour notre étude visant à explorer les liens entre les résultats électoraux, les indicateurs de criminalité et les données relatives à l'emploi.

Nous avons ensuite créé un flux Talend qui a permis d'établir des connexions entre chaque fichier Excel récupéré depuis le site data.gouv.fr et une base de données que nous avons mise en place sur un serveur Microsoft en local. Cette base de données centralise les informations extraites de chaque fichier Excel, formant ainsi une source complète et structurée pour notre analyse. Cette démarche nous a offert un environnement efficace pour gérer, stocker et traiter les données de manière cohérente.

Par la suite, nous avons établi plusieurs groupes de travail pour répartir les tâches selon un plan méthodique :

1 **Nettoyage et Agrégation des Données Électorales de 2017 :**
Un groupe s'est concentré sur le nettoyage des données des élections présidentielles de 2017 à l'aide d'un script python, en regroupant les résultats par commune de la région. De plus, les voix obtenues par orientation politique (gauche, centre et droite) ont été synthétisées (auparavant le nombre de voix était regroupé par candidat, désormais le nombre de voix est regroupé par orientation politique). Le parti gagnant a été identifié pour chaque commune, et les votes nuls (vote blanc, vote erroné ou abstention) ont été regroupés. Voir le script : Script2017
Vous pourrez trouver le résultat : resultat_finale_2017

2 **Nettoyage et Agrégation des Données Électorales de 2022 :**
Le même travail que pour l'année 2017 a été effectué (voir ci-dessus).
Voir le script : scriptElection2022
Vous pourrez trouver le résultat : resultat20221

3 **Création d'Indicateurs de Criminalité :**
Un troisième groupe s'est penché sur la création d'indicateurs de criminalité à partir des données recueillies. Cela a impliqué de regrouper les informations sur les types de crimes, les nombres de victimes, et les occurrences de crimes en fonction des communes.
Voir le script : Script_indice_criminalite.sql
Vous pourrez trouver le résultat : indices_crim_dep

4 **Création d'Indicateurs d'Emploi :**

Un quatrième groupe a travaillé sur la création d'indicateurs d'emploi en se basant sur les données collectées. Cela a inclus le calcul de la moyenne des demandeurs d'emploi pour chaque année ainsi que l'estimation de la population dans la tranche d'âge de 20 à 60 ans (population dite active).
Voir le script : emploie
Vous pourrez trouver le résultat : Demandeur_Emploie.csv et estim-pop-dep-sexe-gca-1975-2023.xls

5 **Fusion et Analyse des Données :**
Enfin, nous avons procédé au regroupement de toutes les années traitées dans le but de les fusionner. Cette étape essentielle nous a permis de créer une base de données complète et cohérente, prête à être analysée dans le but de dégager des tendances et des relations significatives entre les résultats électoraux, la criminalité et l'emploi.
Voir le script : scriptFinal.ipynb
Vous pourrez trouver le résultat : commune_finalrevu.xlsx

Par la suite, une fois les données soigneusement traitées et agrégées, nous les avons consolidées à la fois dans notre base de données locale et dans l'environnement Talend. À ce stade, notre objectif était d'exploiter ces données transformées pour la création de modèles de prédiction avancés.

Nous avons donc entrepris la conception de deux modèles : un modèle KNN et un réseau neuronal. Ces modèles ont été élaborés avec la perspective de prédire les résultats des élections présidentielles à venir, spécifiquement pour l'année 2022, en utilisant les données de 2017 comme données d'entraînement. Voir le code : PREDICTION.ipynb

Outils technologiques :

Pour effectuer pour cette mise en situation professionnelle nous avons utilisé un ensemble de logiciels de communication, gestion des tâches à effectuer, gestion des données (stockage, flux, traitement) et de versioning de code :

TRELLO :

DISCORD :

Logiciel gratuit de messagerie instantanée

LOOPING :

Est un logiciel de modélisation conceptuelle de données entièrement gratuit et libre d'utilisation.

EXCEL :

Logiciel de tableur de la suite Office de MicroSoft

TALEND :

C’est un éditeur de logiciel spécialisé dans l'intégration de données

talend

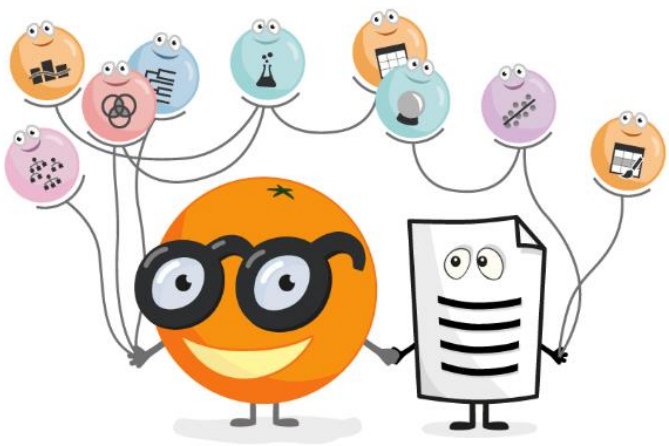
MS SQL SERVER MANAGEMENT STUDIO :

Logiciel développé par Microsoft qui est utilisée pour configurer, gérer et administrer tous les composants de Microsoft SQL Server

Microsoft® SQL Server®

ORANGE DATAMINING :

Logiciel gratuit de Machine Learning, data visualisation, diagrammes de flux de travail d'analyse de données



POWER BI :

Logiciel de création des visualisations de données personnalisées et interactives avec une interface pour créent ses propres rapports et tableaux de bord.

PYTHON :

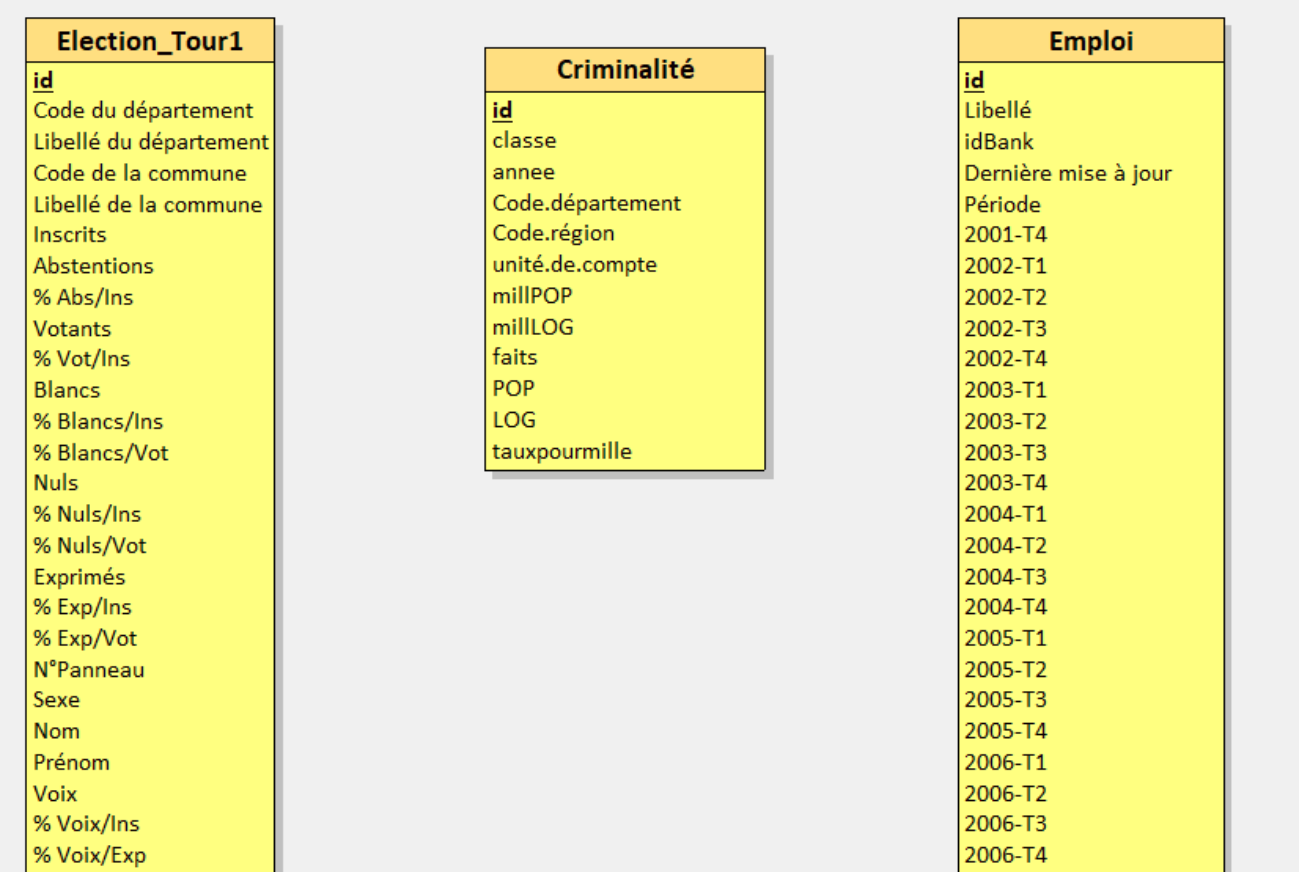
Le langage de programmation utilisé souvent en Intelligence Artificielle

GitHub :

Le logiciel de « versioning » de code le plus utilisé de nos jours https://github.com/Altmanmike/MsprBigData

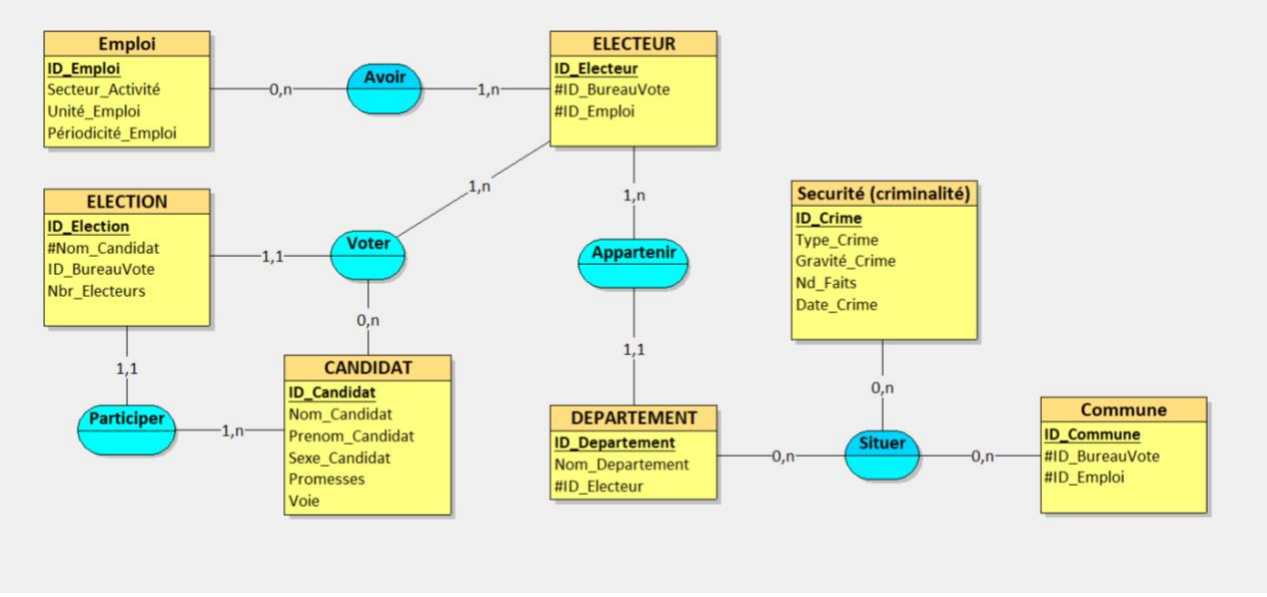
4. Un Modèle Conceptuel de Données

Lorsque nous avons débuté ce projet nous sommes partis du modèle de données suivant :



Nous avions trois tables de données sous format Excel pour chaque année : élection tour 1, criminalité et emploi. Ces dernières possédaient des données qui n'étaient pas exploitables pour notre projet.

Après avoir effectué la restructuration des tables et le nettoyage des données, nous sommes arrivés au modèle conceptuel de données suivant :



5. Les modèles testés

Nous avons entrepris des tests approfondis en utilisant deux modèles de clustering bien établis : le modèle KIH2 (K-means amélioré par la méthode de l'inertie hiérarchique) et le modèle K-MEANS.

KIH2 :

KIH2 (K-means Iterative Hypergraph Hierarchical) est un algorithme utilisé dans le domaine du Big Data pour effectuer une analyse de clustering, c'est-à-dire regrouper des données similaires ensemble.

L'algorithme KIH2 combine deux approches populaires de clustering : K-means et l'hypergraphe itératif hiérarchique.

Tout d'abord, l'algorithme K-means est utilisé pour effectuer une première étape de clustering sur les données. L'objectif du K-means est de partitionner les données en K clusters, où K est un paramètre prédéfini. Cela se fait en assignant chaque point de données à un cluster en fonction de la proximité de ses caractéristiques avec les centroïdes de ces clusters.

Ensuite, l'algorithme hypergraphe itératif hiérarchique est utilisé pour raffiner les clusters obtenus à l'étape précédente. Un hypergraphe est utilisé pour représenter les relations entre les points de données et les clusters, où les hyperarêtes représentent les similarités entre les points de données et les clusters. L'algorithme utilise une approche itérative pour fusionner et diviser les clusters afin d'améliorer la qualité globale de la partition.

L'avantage de l'algorithme KIH2 est qu'il peut gérer des ensembles de données volumineux et complexes, ce qui en fait une méthode pertinente dans le contexte du Big Data. Il peut également être utilisé pour découvrir des structures hiérarchiques dans les données, ce qui permet une analyse plus approfondie des relations entre les clusters.

En résumé, KIH2 est un algorithme de clustering qui combine les techniques de K-means et de l'hypergraphe itératif hiérarchique pour effectuer une analyse de clustering dans le domaine du Big Data.

K-MEANS :

K-means est un algorithme de clustering largement utilisé dans le domaine de l'analyse de données. L'objectif principal de K-means est de diviser un ensemble de données en K clusters, où K est un paramètre prédéfini.

Voici comment fonctionne l'algorithme K-means :

Initialisation : Sélectionnez aléatoirement K points de données comme centres de cluster initiaux, appelés "centroïdes".

Attribution : Chaque point de données est attribué au cluster représenté par le centroïde le plus proche en termes de distance euclidienne ou d'autres mesures de similarité.

Mise à jour des centroïdes : Recalculer les positions des centroïdes en prenant la moyenne des positions de tous les points de données attribués à chaque cluster. Les centroïdes se déplacent vers le centre de gravité de leur cluster.

Répéter les étapes 2 et 3 : Répétez les étapes d'attribution et de mise à jour des centroïdes jusqu'à ce qu'une condition de convergence soit satisfaite. La condition de convergence peut être définie par un nombre fixe d'itérations ou lorsque les centroïdes cessent de se déplacer de manière significative.

Résultats : Une fois que l'algorithme a convergé, les données sont divisées en K clusters, chaque cluster étant représenté par son centroïde.

Maintenant, comparons K-means et KIH2 :

Approche : K-means utilise une approche partitionnelle, où les points de données sont directement attribués à un seul cluster. En revanche, KIH2 utilise une approche hiérarchique, où les clusters sont construits en fusionnant et en divisant itérativement des hyperarêtes.

Gestion des données volumineuses : KIH2 est généralement plus adapté aux ensembles de données volumineux et complexes, car il peut utiliser des hypergraphes pour représenter les relations entre les points de données et les clusters, ce qui permet une modélisation plus riche et flexible des données. K-means peut également être utilisé pour des ensembles de données volumineux, mais il peut rencontrer des problèmes de performances et d'extensibilité lorsque la taille des données devient très grande.

Structures hiérarchiques : KIH2 permet de découvrir des structures hiérarchiques dans les données, ce qui signifie qu'il peut identifier des sous-clusters au sein de chaque cluster principal. En revanche, K-means ne produit généralement qu'une seule partition de données sans informations sur les relations hiérarchiques entre les clusters.

Complexité de l'algorithme : K-means est généralement plus simple et plus rapide que KIH2 en termes de complexité algorithmique. KIH2 implique des étapes itératives de fusion et de division des clusters, ce qui peut rendre l'algorithme plus coûteux en termes de temps de calcul.

Réseau Neuronal :

Un réseau neuronal est un modèle d'apprentissage automatique inspiré par le fonctionnement du cerveau humain. Il est composé de couches de nœuds interconnectés, appelés neurones artificiels, qui traitent et transmettent des informations. Chaque neurone effectue une combinaison linéaire de ses entrées, suivi d'une fonction d'activation non linéaire, produisant ainsi une sortie. Les réseaux neuronaux sont utilisés pour diverses tâches telles que la classification, la reconnaissance d'images et le traitement du langage naturel. L'apprentissage se fait en ajustant les poids des connexions entre les neurones à partir de données d'entraînement, permettant au réseau de capturer des modèles complexes dans les données.

Régression Linéaire :

La régression linéaire est une technique statistique utilisée pour modéliser la relation entre la variable que l'on cherche à prédire et une ou plusieurs variables indépendantes (les variables explicatives). L'objectif est de trouver la meilleure ligne droite (ou hyperplan dans le cas de plusieurs variables indépendantes) qui minimise l'écart entre les valeurs prédites par le modèle et les valeurs réelles observées dans les données. Cette ligne droite représente une approximation linéaire de la relation entre les variables.

6. Les résultats du modèle choisi

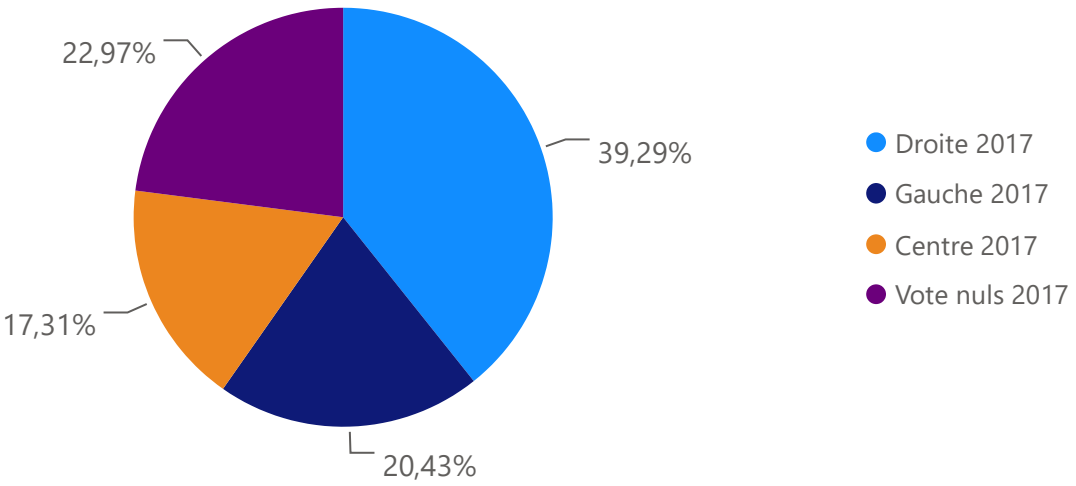
Nous avons opté pour le modèle de réseau neuronal qui affiche l'accuracy la plus haute qui est d'envions 0,86.

```
Précision du modèle de réseau neuronal : 0.8582089552238806
```

7. Les visualisations

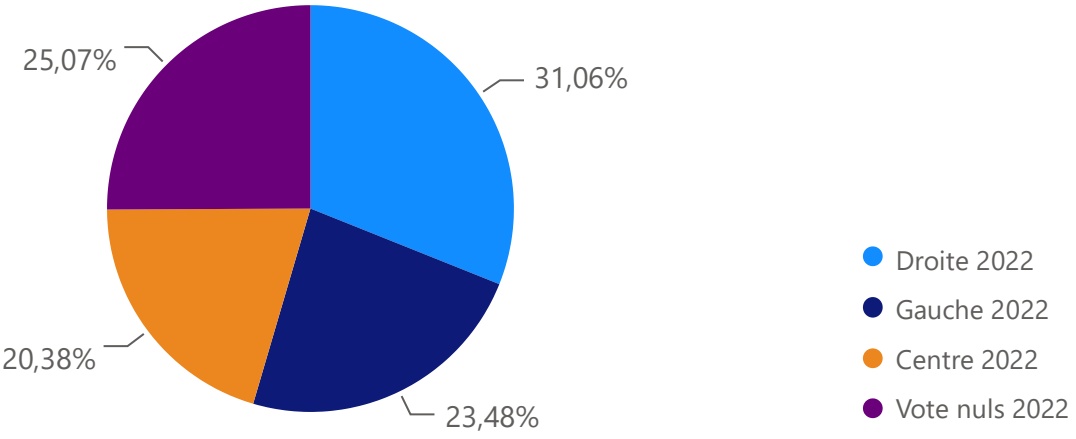
Pour la visualisation et la comparaison de nos données, nous avons choisi l'outil Power Bi.

Répartition des parties 2017



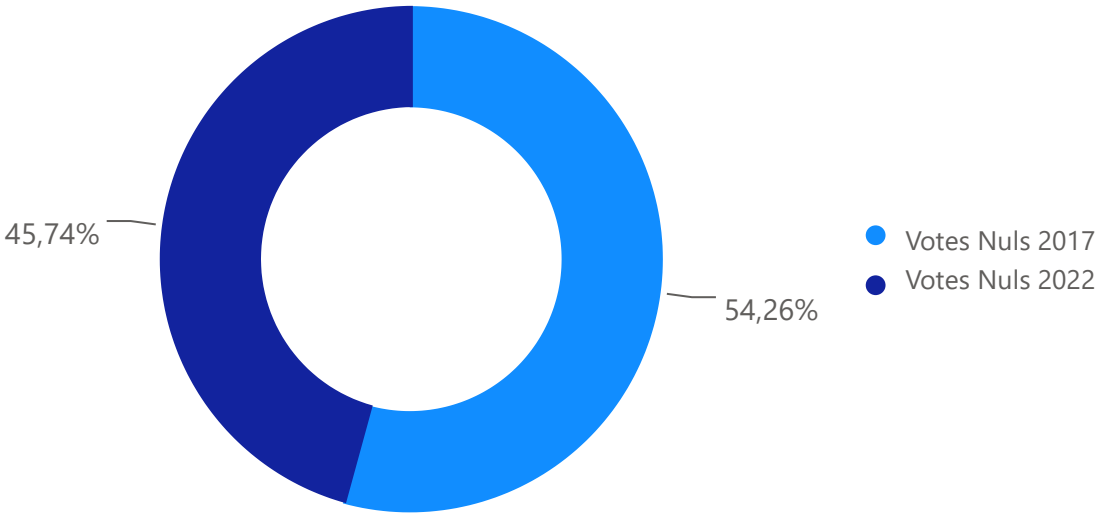
Le camembert ci-dessus illustre la répartition des votes des élections de 2017 en France. La droite domine avec la plus grande part, suivie de près par le nombre de vote nuls. Les partis de la gauche ont une part modérée, tandis que les votes du centre représentent une part plus petite mais notable.

Répartition des parties 2022



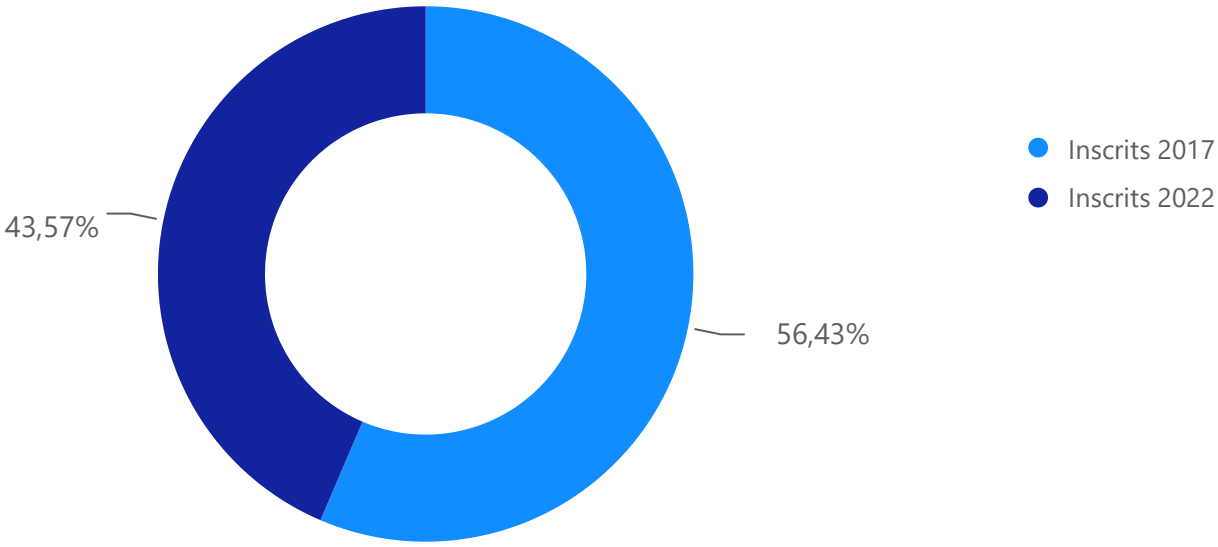
Le camembert ci-dessus illustre la répartition des votes des élections de 2022 en France. La droite domine avec la plus grande part, suivie de près par le nombre de vote nuls. Les partis de la gauche ont une part modérée, tandis que les votes du centre représentent une part plus petite mais notable.

Nombre de Votes Nuls



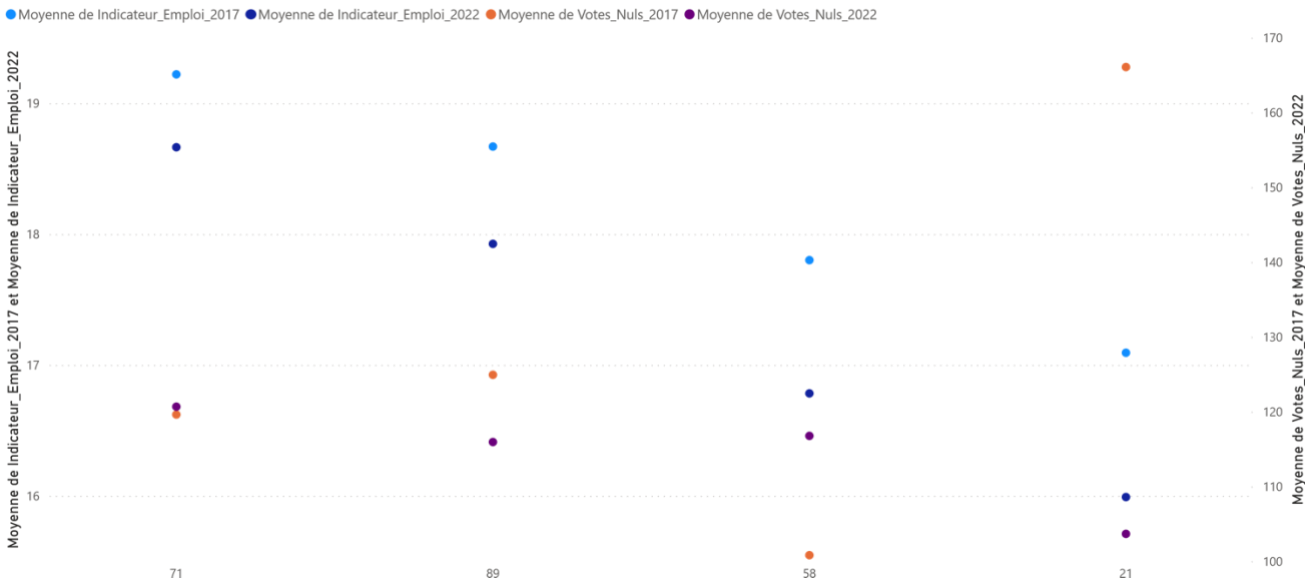
Ce graphique en anneau compare les votes nuls entre 2017 (54,26 %) et 2022 (45,74 %). Une variation notable suggère un possible changement dans la façon dont les électeurs ont abordé leur participation, illustrant un aspect changeant du processus électoral sur cette période."

Nombre d'inscrits

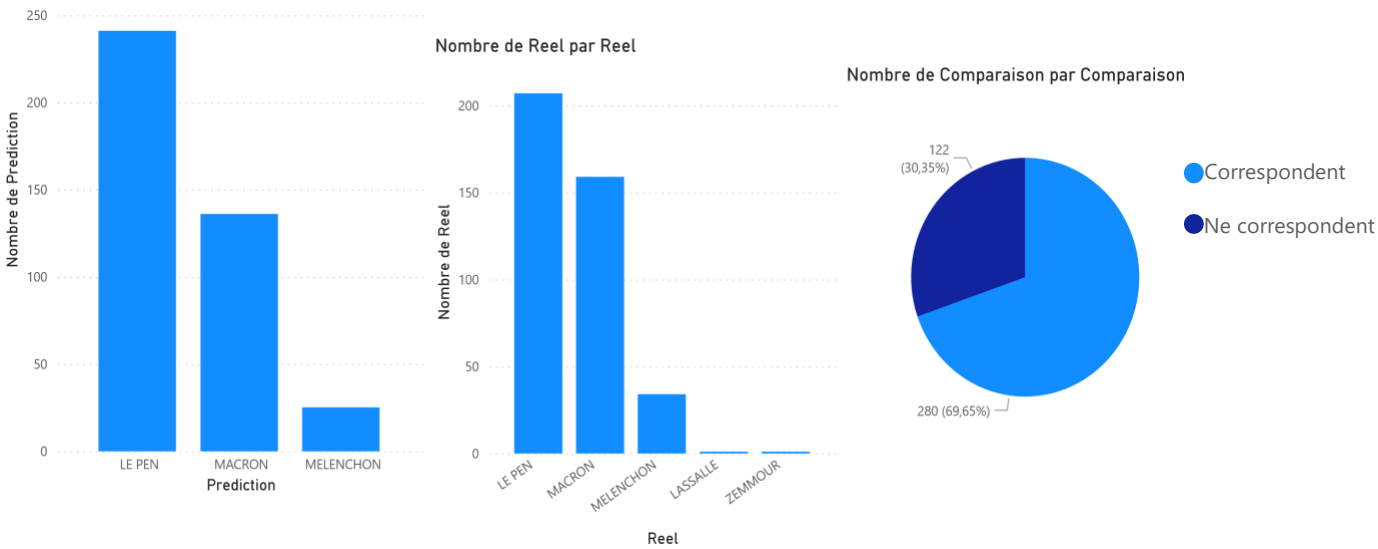


Ce graphique montre une comparaison entre le nombre d'inscrits en 2017 et en 2022. Une différence marquée entre ces deux années indique des changements dans la participation électorale et peut refléter des tendances démographiques, des intérêts politiques évoluant ou d'autres facteurs influençant le nombre d'électeurs inscrits.

Moyenne de l’Indicateur_Emploi_2017, Moyenne de l’Indicateur_Emploi_2022, Moyenne de Votes_Nuls_2017 et Moyenne de Votes_Nuls_2022 par Code du département

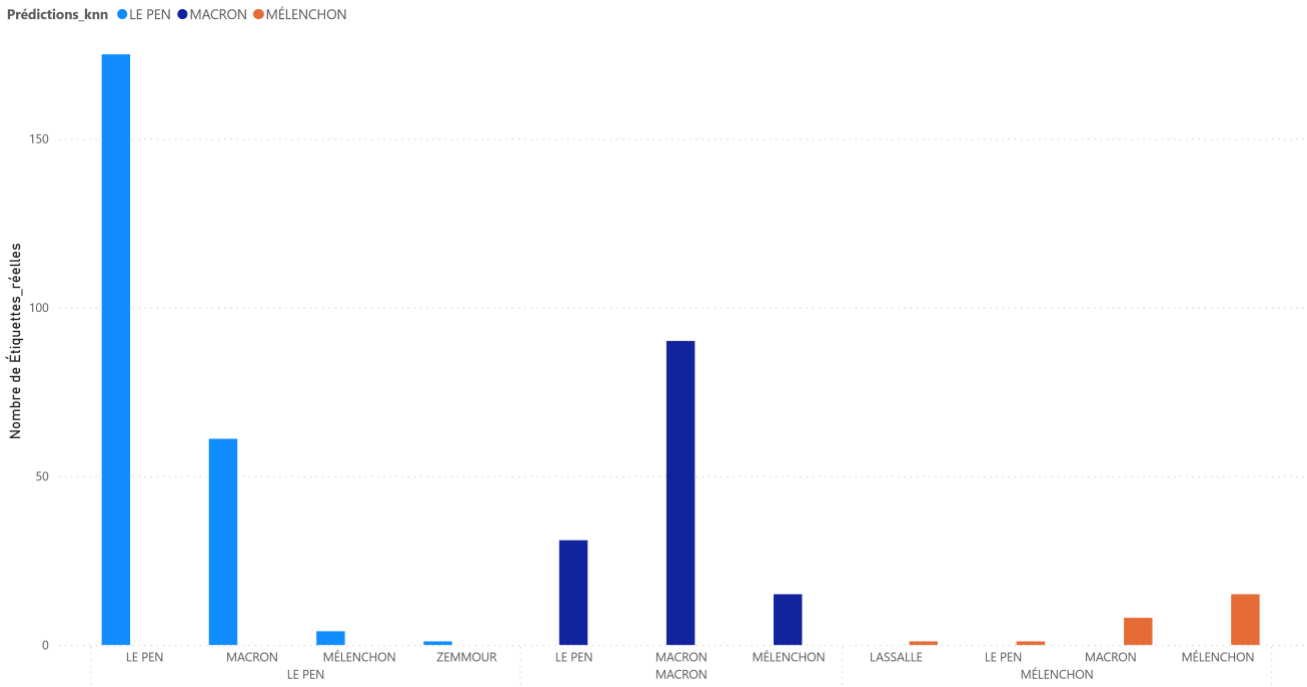


Le graphique en nuage de points compare les moyennes des indicateurs d'emploi et des votes nuls en 2017 et 2022 pour chaque code de département. Cela aide à identifier des tendances potentielles entre ces variables clés à l'échelle régionale.



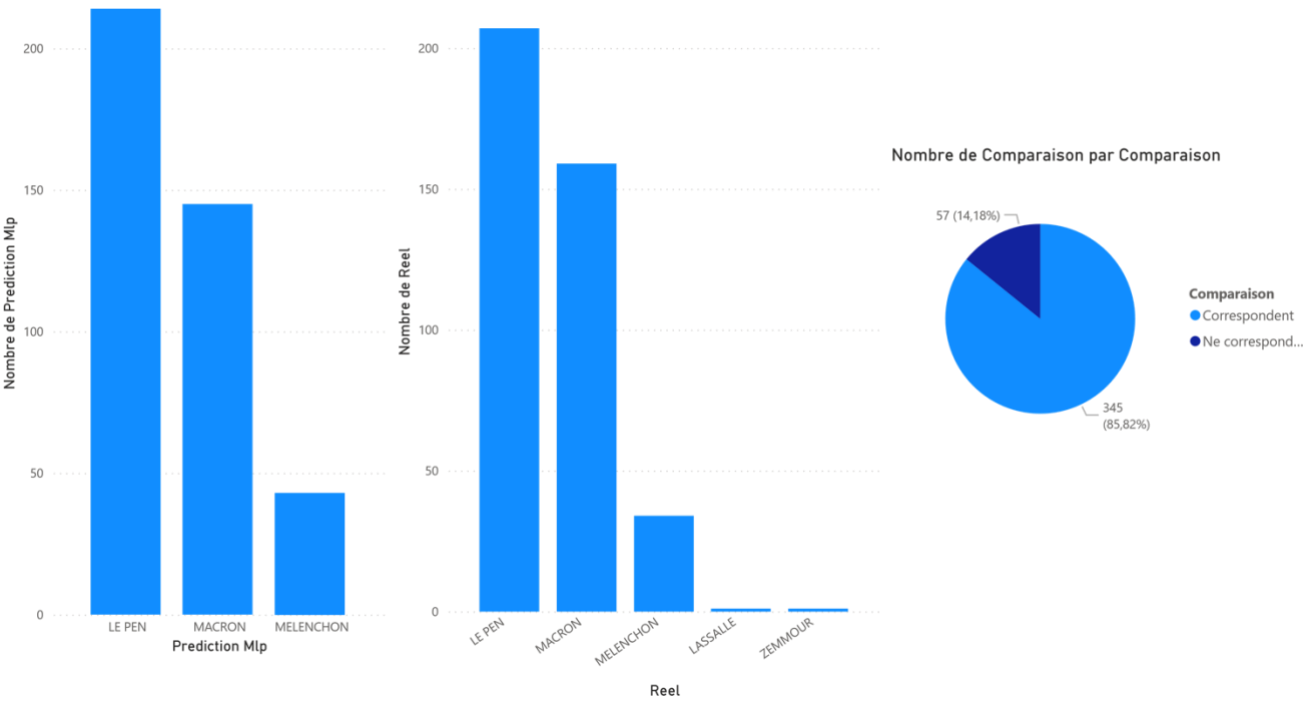
Ces graphiques comparent les prédictions d'un modèle pour les élections de 2022 avec les résultats réels de cette. Ils évaluent l'exactitude du modèle et mettent en évidence les évolutions politiques entre ces deux périodes.

Nombre de Étiquettes réelles par Prédictions_knn, Étiquettes_réelles et Prédictions_knn



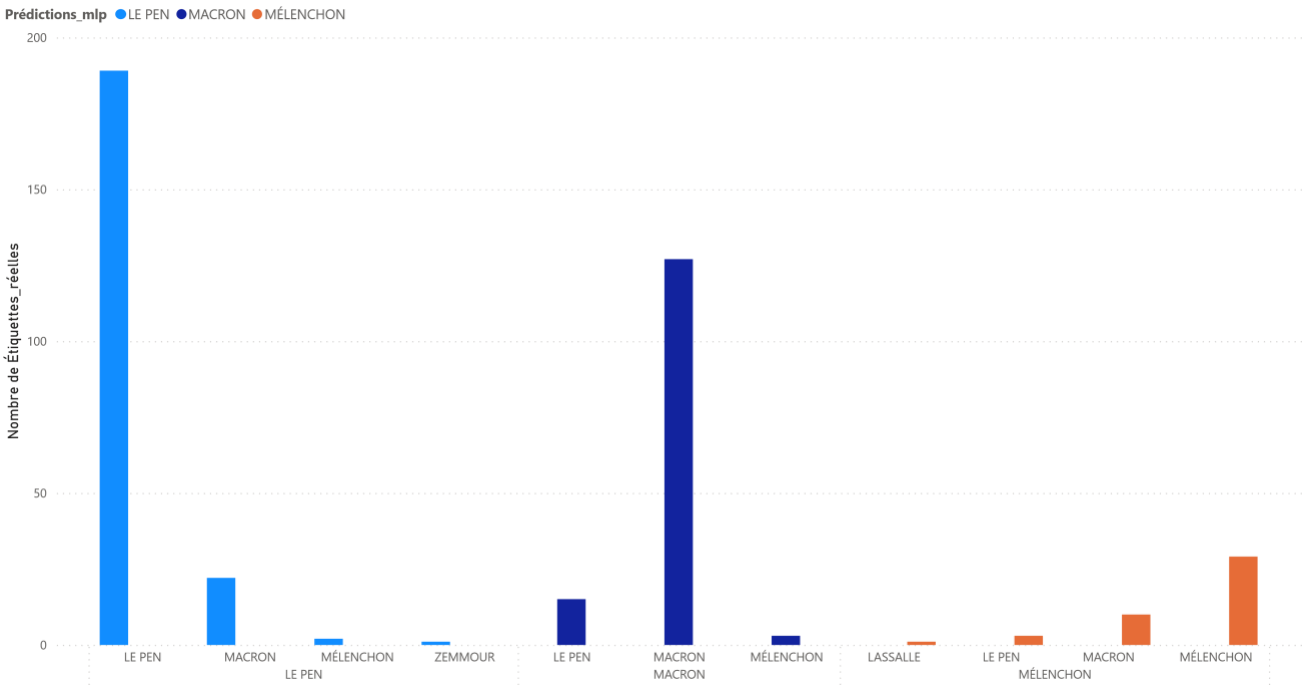
Cet histogramme compare les étiquettes réelles avec les prédictions du modèle k plus proches voisins (Prédictions knn). Il évalue visuellement à quel point les prédictions du modèle correspondent aux étiquettes réelles.

Nombre de Prediction Mlp par Prediction Mlp



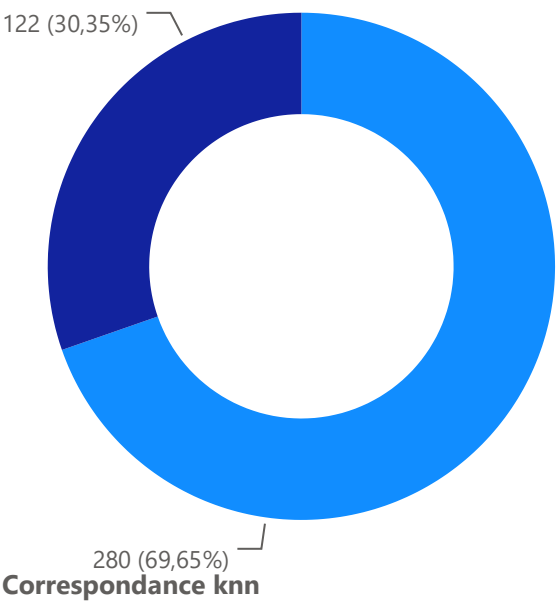
Ces graphiques comparent les prédictions d'un modèle pour les élections de 2022 avec les résultats réels de cette. Ils évaluent l'exactitude du modèle et mettent en évidence les évolutions politiques entre ces deux périodes.

Nombre de Étiquettes_réelles par Prédictions_mlp, Étiquettes_réelles et Prédictions_mlp

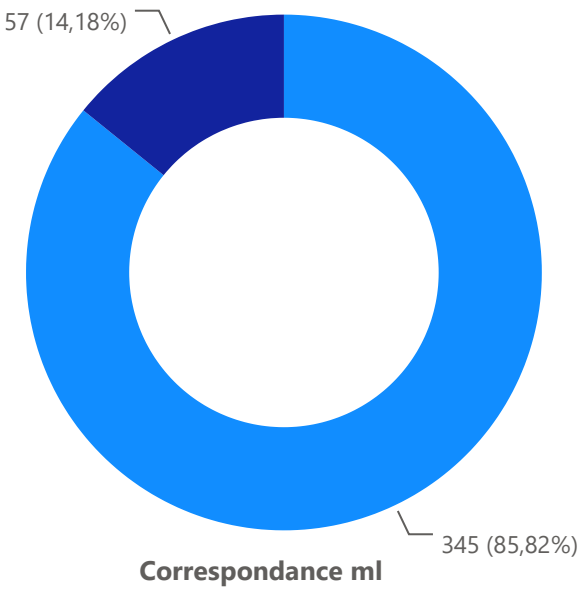


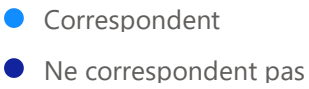
Cet histogramme compare les étiquettes réelles avec les prédictions du modèle. Il évalue visuellement à quel point les prédictions du modèle correspondent aux étiquettes réelles.

Nombre de Correspondance knn par Correspondance knn



Nombre de Correspondance mlp par Correspondance mlp

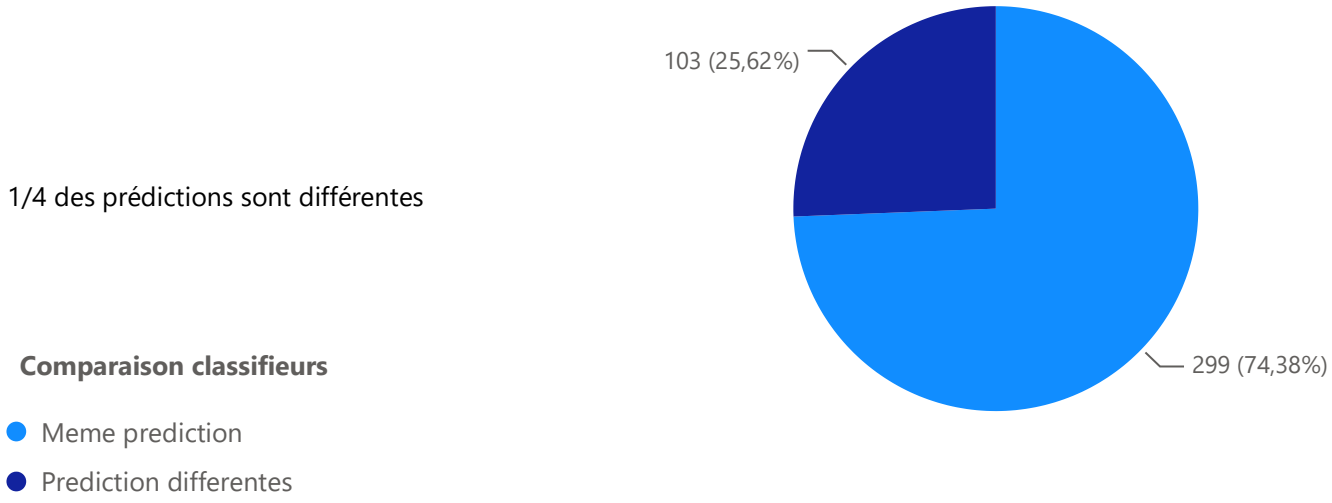




Ces deux graphiques en forme d'anneaux comparent les prédictions des deux modèles. La première partie de l'anneau montre les correspondances, tandis que la seconde met en évidence les divergences. Ils offrent une comparaison de l'efficacité des deux modèles.

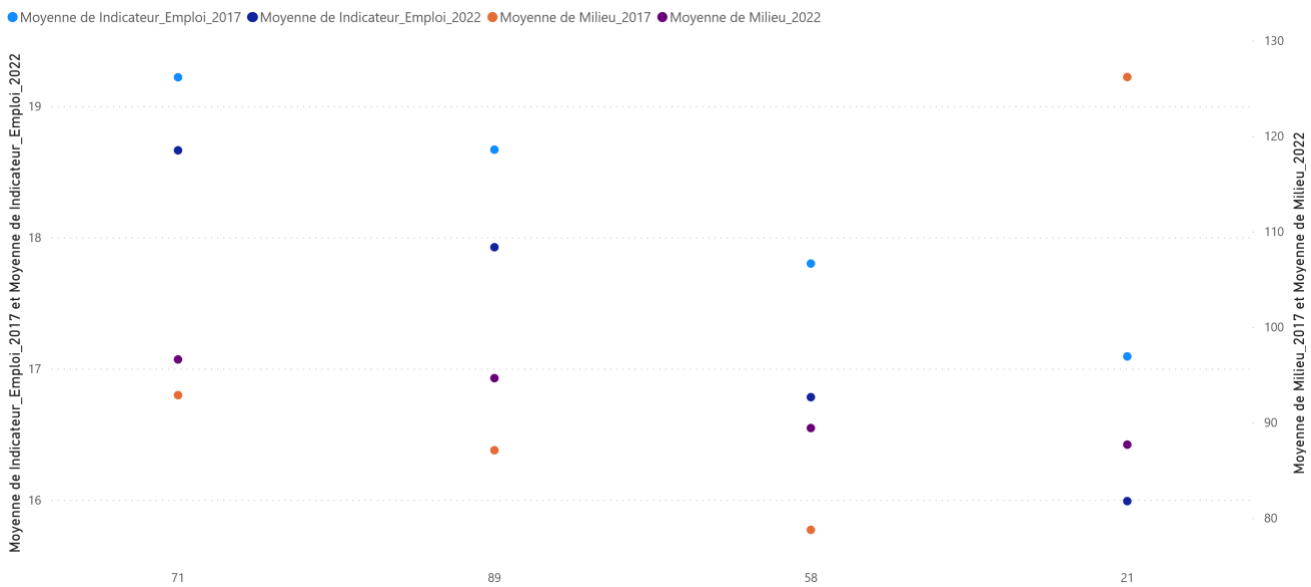
Nombre de comparaison classifieurs par comparaison classifieurs

1/4 des prédictions sont différentes



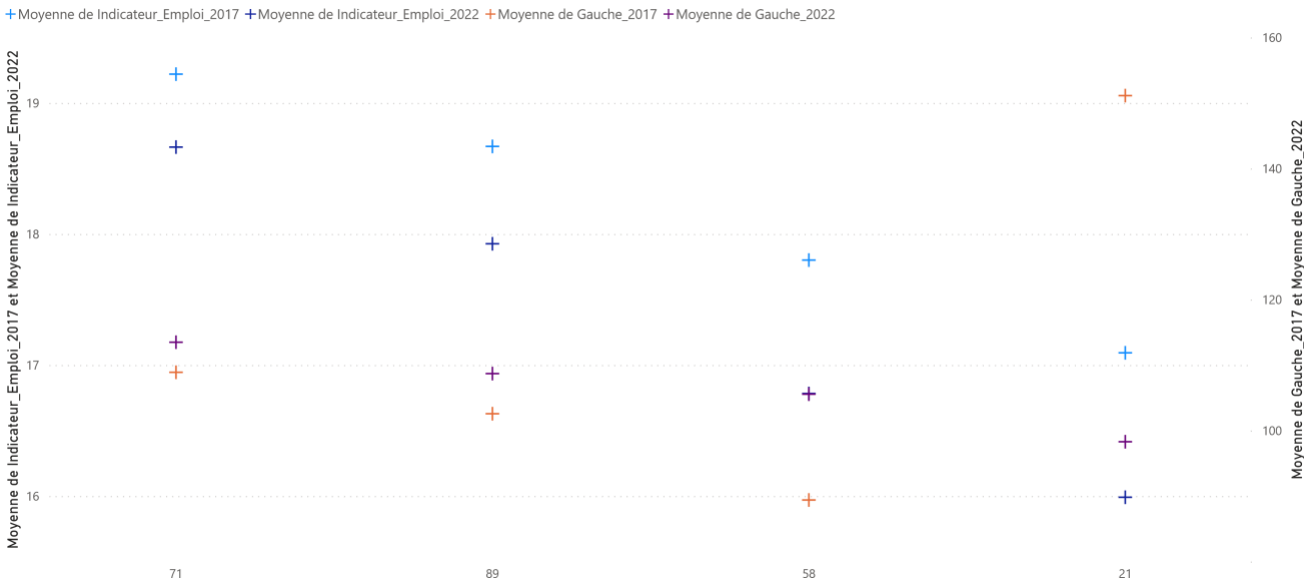
Le camembert présente une comparaison des prédictions d'un classifieur. Cette visualisation synthétise la répartition des prédictions dans différentes catégories définies par le classifieur. Cela peut aider à identifier les proportions relatives de chaque classe prédite par le modèle, offrant ainsi un aperçu global de ses performances.

Moyenne de l'Indicateur_Emploi_2017, Moyenne de l'Indicateur_Emploi_2022, Moyenne de Milieu_2017 et Moyenne de Milieu_2022 par Code du département



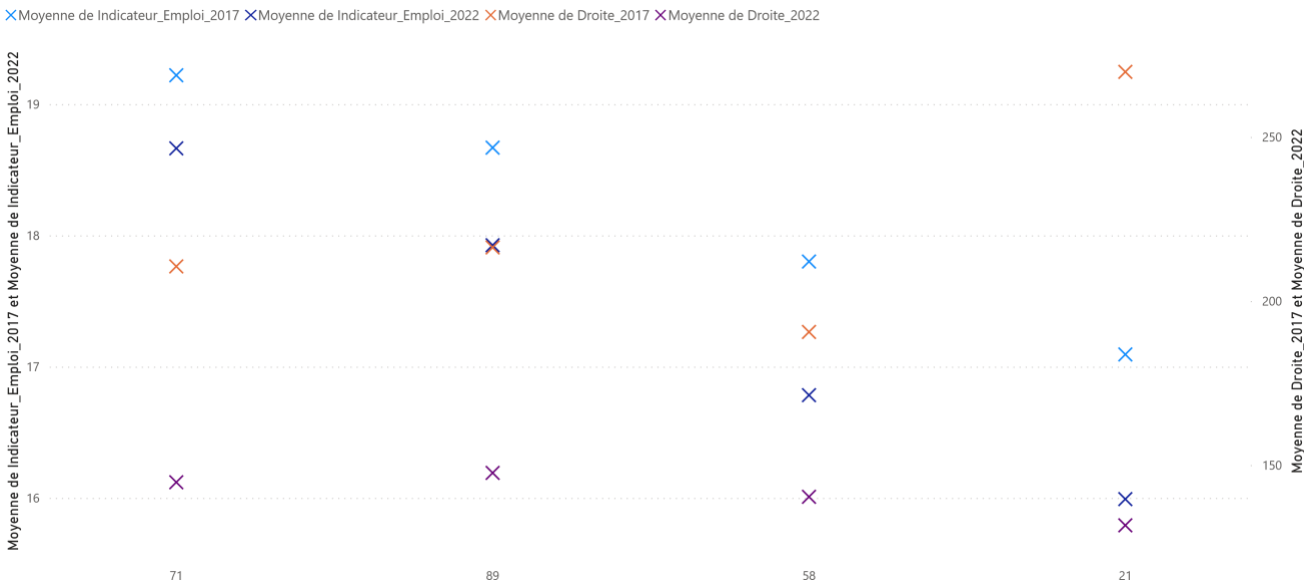
Ce graphique en nuage de points compare les moyennes de l'Indicateur Emploi et de Milieu en 2017 et 2022 pour chaque code de département. Il aide à identifier les tendances et les relations entre ces variables au fil du temps et dans différentes régions.

Moyenne de l’Indicateur_Emploi_2017, Moyenne de l’Indicateur_Emploi_2022, Moyenne de Gauche_2017 et Moyenne de Gauche_2022 par Code du département

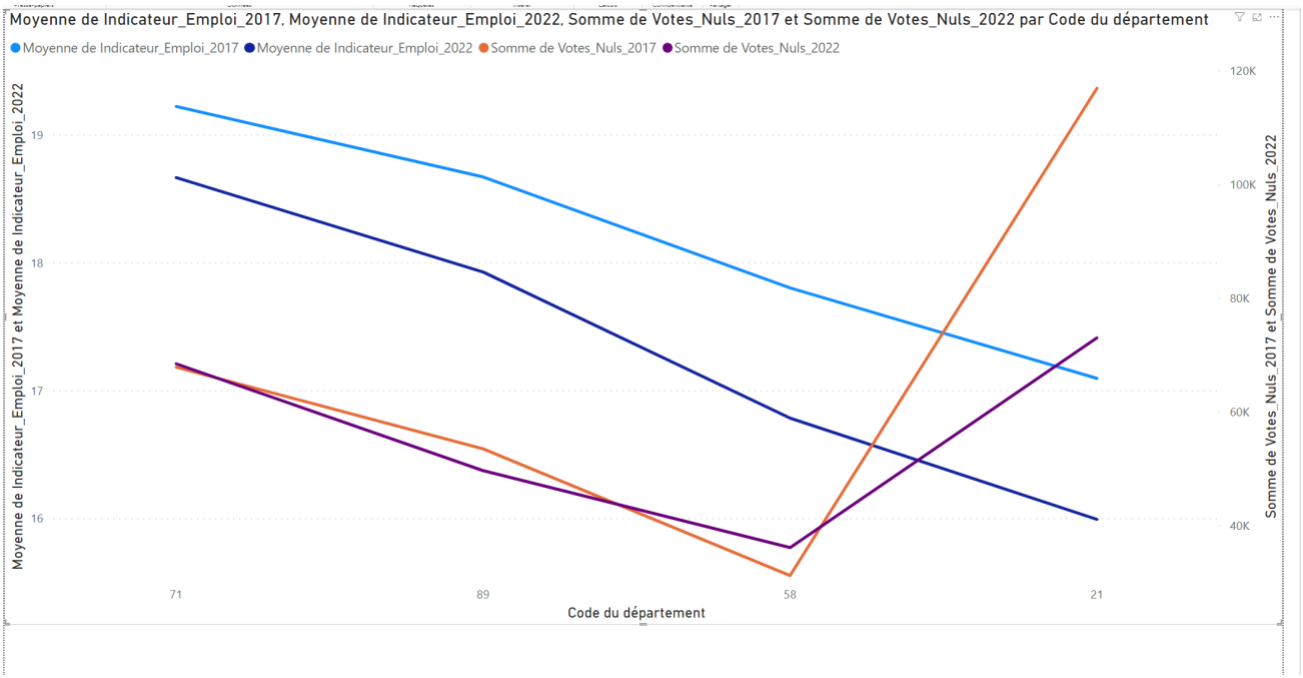


Ce graphique en nuage de points compare les moyennes de l'Indicateur Emploi en 2017 et 2022 avec les moyennes de Gauche pour les mêmes années, pour chaque code de département. Il met en lumière les variations potentielles entre les indicateurs d'emploi et les préférences politiques de gauche au fil du temps et dans différentes régions.

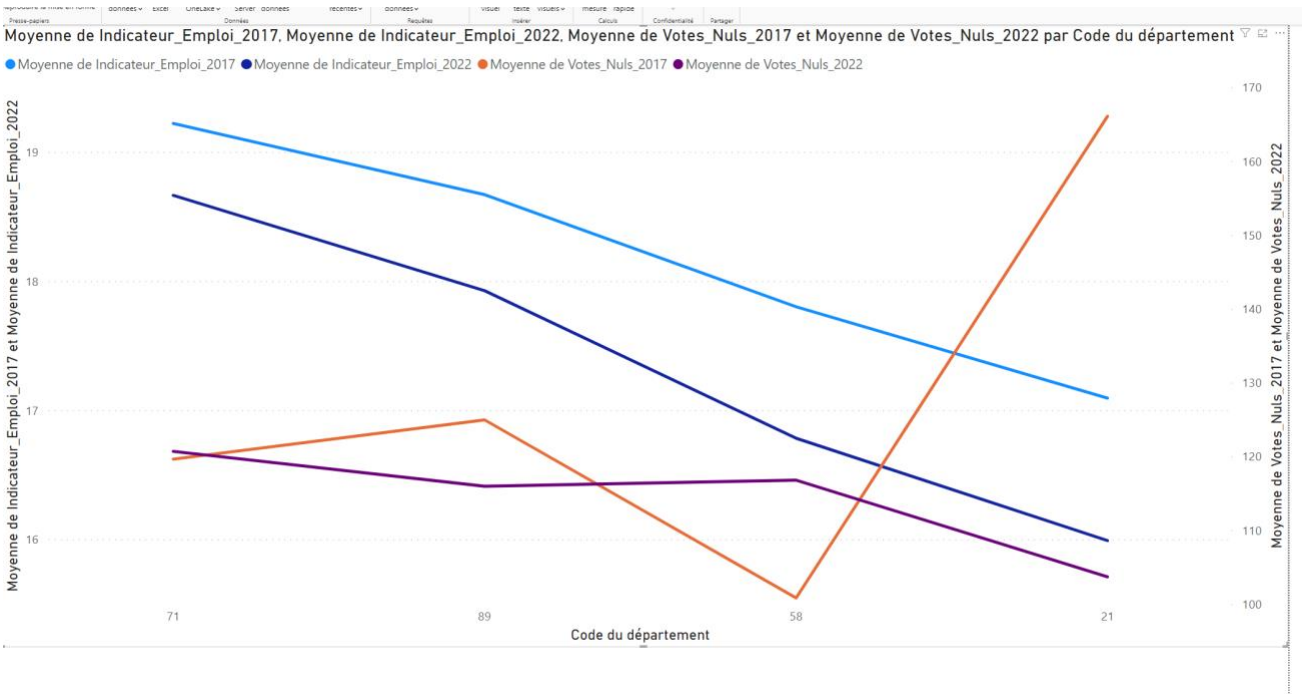
Moyenne de l’Indicateur_Emploi_2017, Moyenne de l’Indicateur_Emploi_2022, Moyenne de Droite_2017 et Moyenne de Droite_2022 par Code du département



Ce graphique en nuage de points compare les moyennes de l'Indicateur Emploi en 2017 et 2022 avec les moyennes de Droite pour les mêmes années, pour chaque code de département. Il explore les variations possibles entre les indicateurs d'emploi et les préférences politiques de droite au fil du temps et dans différentes régions.



Ce graphique présente plusieurs courbes pour chaque code de département, comparant la moyenne de l'indicateur emploi en 2017 et 2022 avec la somme des votes en 2017 et 2022. Cela permet d'examiner les tendances entre les indicateurs d'emploi et les niveaux de votes sur différentes périodes et dans diverses régions.



Ce graphique présente plusieurs courbes pour chaque code de département, comparant la moyenne de l'indicateur emploi en 2017 et 2022 avec la moyenne des votes en 2017 et 2022. Cela permet d'examiner les relations entre les indicateurs d'emploi et les niveaux moyens de votes sur différentes périodes et dans diverses régions.

8. Accuracy (pouvoir prédictif du modèle)

En utilisant les données nettoyées et les critères que nous avions recueillis, notre objectif était de développer un modèle prédictif capable d'estimer les résultats des prochaines élections pour chaque ville de cette région. PREDICTION.ipynb

Vous pourrez également retrouver la visualisation des ses prédictions ci-dessus.

9. Les réponses aux questions posées dans les exemples d'indicateurs d'analyse :

L'utilisation de données sensibles comme les résultats électoraux, les taux de criminalité et de chômage nécessite l'autorisation de la CNIL. La CNIL garantit que ces données sont traitées légalement et protège la vie privée. Cette démarche assure un équilibre entre l'analyse de données importantes et le respect des droits individuels

10. Les réponses aux questions posées dans les exemples d'indicateurs d'analyse :

Parmi les données que vous avez sélectionnées, laquelle est la plus corrélée aux résultats des élections ?

La donnée la plus corrélée est l'indicatif d'emploi avec les résultats des partis de droite (voir ci-dessous).

Définissez le principe d'un apprentissage supervisé :

L'apprentissage supervisé est une approche d'apprentissage automatique où un modèle apprend à partir d'exemples étiquetés. Ces exemples comprennent des paires d'entrées et de sorties attendues, permettant au modèle de généraliser et de faire des prédictions précises sur de nouvelles données non vues.

Comment définissez-vous le degré de précision (accuracy) de votre modèle ?

L'accuracy est le pourcentage de prédictions correctes par rapport au nombre total de prédiction. Il se calcule donc ainsi :

$$\text{Accuracy} = \frac{Nombre\ de\ préictions\ correctes}{Nombre\ total\ de\ prédictions} \times 100$$

11 Sources et liens utiles :

Lors de ce projet nous avions besoin de sources officielles. Nous sommes donc allez chercher ces dernières sur le site officiel du gouvernement data.gouv.fr :

- https://www.data.gouv.fr/fr/pages/donnees-des-elections/
- https://www.data.gouv.fr/fr/pages/donnees-securite/
- https://www.data.gouv.fr/fr/pages/donnees-emploi/
- https://www.data.gouv.fr/fr/organizations/institut-national-de-la-statistique-et-des-etudes-economiques-insee/?datasets_page=7#organization-datasets
- https://www.data.gouv.fr/fr/datasets/activite-emploi-et-chomage-enquete-emploi-en-continu-fichiers-detail/

Vous pourrez retrouver nos scripts commentés et nos résultats sur le dépôt Git suivant : Lien du Git

- Criminalité
- Emploi
- Élection 2017
- Élection 2022
- Excel final