

Rapport – Modélisation du vote départemental

1. Introduction

Ce projet vise à prédire les orientations de vote présidentielle au niveau départemental en France métropolitaine, à partir de données socio-économiques disponibles pour les années 2017 et 2022. L'objectif est de démontrer, à travers une preuve de concept (POC), qu'un modèle de machine learning peut efficacement mettre en évidence des corrélations entre variables structurelles (emploi, sécurité, démographie...) et comportement électoral. Cette approche pourrait être valorisée auprès d'acteurs politiques ou institutionnels pour l'aide à la décision stratégique.

2. Justification de la zone géographique

Le périmètre retenu pour cette étude est la France métropolitaine, incluant les 96 départements continentaux ainsi que les deux entités de la Corse (2A et 2B). Ce choix est justifié par plusieurs considérations à la fois méthodologiques, techniques et statistiques :

- **Couverture statistique homogène** : tous les indicateurs socio-économiques, démographiques et électoraux utilisés dans cette étude sont accessibles de manière cohérente pour l'ensemble des départements métropolitains. Les référentiels INSEE et les données du Ministère de l'Intérieur garantissent une compatibilité inter-temporelle.
- **Comparabilité temporelle** : les périmètres administratifs n'ont pas évolué de manière significative entre 2017 et 2022 en métropole, ce qui permet une analyse comparative robuste entre les deux élections présidentielles sans retraitement géographique complexe.
- **Représentativité politique** : les départements métropolitains offrent une diversité électorale marquée, avec des tendances politiques variées (urbain vs rural, littoral vs intérieur, nord vs sud), permettant de tester la capacité des modèles à saisir cette complexité.
- **Qualité des données** : certains territoires ultramarins présentent des lacunes ou absences dans les bases de données utilisées (revenus, immobilier, criminalité, etc.), ce qui aurait pu biaiser l'analyse ou réduire le périmètre utile à l'apprentissage.

Ainsi, la zone métropolitaine constitue un échantillon suffisamment riche, stable et documenté pour bâtir une preuve de concept pertinente et transposable à d'autres échelles géographiques.

3. Choix des critères

La construction d'un modèle prédictif fiable des comportements électoraux repose sur une sélection rigoureuse des variables explicatives. Dans cette étude, les critères retenus résultent à la fois d'un raisonnement théorique (issus des sciences politiques et sociales) et d'un retour empirique sur les dynamiques électorales observées en France métropolitaine. Chaque variable a été sélectionnée en fonction de sa disponibilité, sa cohérence sur plusieurs années, et son lien supposé ou démontré avec les résultats électoraux.

- **Taux de criminalité :**

Le taux de criminalité, bien qu'imparfait en tant que reflet du "sentiment d'insécurité", reste un indicateur significatif dans l'étude du vote. Il ne s'agit pas uniquement de mesurer les actes délictueux réels, mais de comprendre comment un territoire est perçu en termes de sécurité par ses habitants. De nombreuses études montrent une corrélation entre des contextes de forte insécurité ou de perception de dégradation de l'ordre public, et le vote pour des partis autoritaires, souverainistes ou promouvant des politiques sécuritaires rigides. Ainsi, les départements où les infractions sont élevées (atteintes aux biens, violences, incivilités...) peuvent présenter une orientation politique marquée en faveur de candidats prônant une « ligne dure » sur ces questions.

- **Salaire moyen :**

Le niveau de salaire moyen est un indicateur économique direct du pouvoir d'achat d'un territoire. Il reflète également l'attractivité économique d'un département, la qualification de sa population active, et le type d'emploi majoritairement présent (tertiaire, industriel, précaire, etc.). Or, la question du pouvoir d'achat s'est imposée comme un sujet central dans les campagnes présidentielles récentes. Des départements à faible salaire moyen seront plus sensibles aux discours économiques sur la justice sociale, la redistribution ou la lutte contre l'inflation. À l'inverse, des territoires plus aisés pourraient favoriser des partis mettant l'accent sur la compétitivité économique, la réduction des charges ou la stabilité fiscale.

- **Taux de chômage :**

Le chômage, plus qu'un simple chiffre, est une variable chargée émotionnellement et politiquement. Il est souvent vécu comme un symptôme de l'abandon territorial, de la désindustrialisation ou du déclin économique. Les départements touchés durablement par un chômage élevé développent une forme de désillusion politique, qui se traduit soit par un désengagement (abstention), soit par un vote radical, anti-système ou anti-mondialisation. C'est pourquoi ce critère est fondamental : il permet de modéliser le mécontentement socio-économique structurel, qui alimente souvent les dynamiques électorales les plus tranchées.

- **PIB par habitant :**

Le produit intérieur brut par habitant mesure la création de richesse au niveau départemental. Il s'agit d'un critère macroéconomique qui donne une image synthétique du dynamisme économique local. Un PIB par habitant élevé indique souvent une concentration de sièges sociaux, d'activités tertiaires à forte valeur ajoutée, ou de tourisme international. Ces territoires sont plus susceptibles de voter pour des partis réformistes, pro-européens ou technocratiques. À l'inverse, un PIB faible peut coïncider avec un repli identitaire, un rejet des élites et une demande de protection accrue. Intégrer ce critère permet d'équilibrer l'analyse entre les conditions individuelles (salaires, chômage) et les dynamiques économiques globales.

- **Taux d'immigration :**

La présence d'une population immigrée importante dans un département est un facteur à double tranchant. D'un côté, certains électorsats peuvent percevoir l'immigration comme une richesse culturelle et sociale, favorisant des votes en faveur de la diversité et de l'ouverture. De l'autre, elle peut être instrumentalisée par des discours politiques focalisés sur l'identité nationale, le contrôle des frontières ou les conflits interculturels. Le taux d'immigration, sans présumer de son interprétation politique, est donc une variable clivante, qui influence directement la polarisation électorale dans plusieurs régions françaises. Sa prise en compte est essentielle pour comprendre certaines lignes de fracture du vote.

- **Prix de l'immobilier (€/m²) :**

Le prix moyen du mètre carré dans un département constitue un excellent proxy du coût de la vie, mais aussi des dynamiques urbaines telles que la gentrification, l'exode périurbain, ou les tensions foncières. Dans les grandes agglomérations ou les zones touristiques, la flambée des prix peut provoquer un sentiment d'exclusion ou d'injustice sociale. Cela pousse parfois les électeurs vers des partis qui critiquent la financiarisation du logement ou l'insuffisance des politiques publiques en matière d'habitat. À l'opposé, dans les territoires où les prix sont très bas, cela peut signaler une dévitalisation économique. Ce critère, bien qu'indirect, est donc extrêmement révélateur des transformations sociales locales.

- **Structure par âge :**

La pyramide des âges influence profondément les résultats électoraux. Un département jeune votera différemment d'un département vieillissant. Les jeunes générations sont souvent porteuses de revendications liées à l'environnement, aux libertés individuelles, à la modernisation des institutions. À l'inverse, les populations plus âgées privilégient la stabilité, la sécurité, et sont plus réceptives aux discours conservateurs. En intégrant la part des enfants, des adultes et des seniors, on capture cette dimension générationnelle du vote, qui est cruciale pour comprendre la géographie électorale française.

- **Densité de population :**

Ajoutée comme variable explicative, la densité de population (habitants/km²) permet de caractériser le degré d'urbanisation des territoires. Les zones densément peuplées présentent généralement des dynamiques sociales, économiques et culturelles distinctes des zones rurales, ce qui influence significativement les comportements électoraux. Cette variable est fortement corrélée à l'accès aux services publics, aux infrastructures et aux opportunités économiques, tous facteurs déterminants dans la formation des opinions politiques.

- **Taux de natalité :**

Cette variable, intégrée au modèle, mesure le dynamisme démographique des territoires. Un taux de natalité élevé est souvent associé à une population plus jeune, à un optimisme sur l'avenir et à des besoins spécifiques (éducation, logement familial, etc.). Les départements avec un fort taux de natalité peuvent être plus sensibles aux politiques familiales, éducatives et de développement territorial, influençant ainsi leurs choix électoraux.

- **Niveau de vie médian :**

Le niveau de vie médian correspond au revenu net disponible par unité de consommation, divisé de sorte qu'exactly la moitié de la population se situe au-dessous et l'autre moitié au-dessus de ce seuil. Plus représentatif que la moyenne, il rend compte des conditions de vie du "Français moyen" et atténue l'influence des revenus extrêmes. Sur le plan électoral, un niveau de vie médian faible traduit souvent une pression sur le pouvoir d'achat et une frustration diffuse, pouvant alimenter le vote protestataire ou l'adhésion à des programmes promettant une redistribution plus ambitieuse. À l'inverse, des territoires où le niveau de vie médian est élevé peuvent afficher une plus grande confiance envers les partis prônant la stabilité fiscale et la compétitivité économique. En intégrant cette variable, le modèle capte une mesure fine du ressenti socio-économique, essentielle pour prédire l'orientation politique des électors intermédiaires souvent décisifs dans les scrutins.

4. Démarche et méthodes employées

La construction de notre preuve de concept repose sur une méthodologie rigoureuse en plusieurs étapes. Chaque phase du projet a été structurée de manière à garantir la reproductibilité des résultats, la qualité des données traitées, et la robustesse des modèles prédictifs. Le processus s'articule autour de trois scripts principaux :

- `data/<theme>/get_cleaned_data.py` (par thème)
- `generate_dataset.py`
- `learning.py`

1. Extraction & nettoyage

Chaque jeu de données brut (démographie, criminalité, éducation, etc.) est traité dans un script Python dédié, situé dans un dossier thématique (`data/age/`, `data/criminality/`, etc.). Ces scripts, nommés `get_cleaned_data.py`, ont pour objectif de :

- Extraire les données depuis leur format brut (CSV, Excel, JSON...)
- Réorganiser les données sous forme de dictionnaires structurés
- Uniformiser les libellés de colonnes et les types (dates, pourcentages, etc.)
- Gérer les valeurs aberrantes ou manquantes en amont (si évident)

Chaque fonction retourne un dictionnaire hiérarchisé par année et par code départemental, facilitant la fusion ultérieure. Cette étape garantit que chaque domaine dispose d'une base propre, prête à être intégrée.

Pour certaines sources comme la démographie, l'immobilier ou les salaires, les données sont disponibles pour plusieurs périodes (2012, 2017 et 2022), bien que l'analyse principale se concentre sur 2017 et 2022 pour correspondre aux années électorales.

2. Fusion & normalisation des données

Le script `generate_dataset.py` constitue le cœur du pipeline de consolidation. Il fusionne les différentes sources nettoyées en un seul tableau, dans lequel chaque ligne représente un couple {département, année}.

Les principales opérations effectuées sont :

- Boucle sur les années analysées (2017, 2022)
- Boucle sur les 96 départements métropolitains (+2A, 2B)
- Agrégation manuelle des indicateurs (via dictionnaires croisés)
- Génération d'un DataFrame final avec Pandas
- Export vers une base de données SQLite (`dataset.sqlite`)

Les indicateurs sont alignés par clé (`department_code`, `year`), et les scores électoraux sont ajoutés à la fois par orientation politique (gauche, droite, centre) et par parti. Ce format structuré permet une grande flexibilité pour l'entraînement ultérieur.

La segmentation des partis politiques en trois grandes orientations est réalisée selon la classification suivante:

- **Gauche:** LO, NPA, LFI, PCF, PS, EELV, Génération.s
- **Centre:** Renaissance (ex-LREM), Résistons
- **Droite:** UPR, SP, LR, DLF, RN, Reconquête

Cette classification permet de simplifier l'analyse tout en captant les grandes tendances idéologiques.

3. Modélisation prédictive (**learning.py**)

La modélisation repose sur le script **learning.py**, qui implémente une chaîne complète de machine learning supervisé sur les résultats électoraux.

a) Nettoyage des valeurs manquantes

Un audit est réalisé sur les données importées depuis la base SQLite :

- Suppression des colonnes avec plus de 40% de valeurs manquantes (**DROP_THRESHOLD = 0.40**)
- Imputation des NaN restants selon la nature de la distribution :
 - Moyenne (si distribution normale ou symétrique)
 - Médiane (si distribution asymétrique avec un skew > 1)

Ce traitement permet de maximiser la rétention des colonnes utiles, sans introduire de biais majeur.

b) Sélection des features

Les variables explicatives (features) sont filtrées pour ne conserver que les données socio-économiques, démographiques et structurelles. Les colonnes électorales détaillées (**vote_pct_***) sont exclues pour éviter tout effet de fuite.

Les cibles (y) sont définies comme les scores par orientation politique :

- **vote_orientation_pct_Gauche**
- **vote_orientation_pct_Droite**
- **vote_orientation_pct_Centre**

c) Visualisation exploratoire

Avant modélisation, des visualisations sont générées pour explorer la distribution des données:

- Histogrammes des variables socio-économiques clés

- Distribution moyenne des orientations politiques
- Matrices de corrélation entre variables
- Corrélations avec chaque orientation politique cible

Ces visualisations aident à comprendre les relations entre variables et à détecter d'éventuelles anomalies.

d) **Entraînement k-fold sur plusieurs modèles**

Pour chaque orientation, le script entraîne et évalue plusieurs modèles à l'aide d'une validation croisée à 5 plis (KFold) :

- Régression linéaire
- Ridge / Lasso
- RandomForest
- Gradient Boosting
- SVR
- K-Nearest Neighbors
- XGBoost / LightGBM (si installés)

Les métriques suivantes sont calculées :

- RMSE (Root Mean Squared Error)
- MAE (Mean Absolute Error)
- R^2 (coefficient de détermination)

Les résultats sont stockés dans `model_scores.csv`.

e) **Optimisation des hyper-paramètres**

Pour le meilleur modèle brut détecté sur chaque orientation, un GridSearchCV est lancé pour ajuster les hyperparamètres. Cela améliore la précision tout en évitant le surapprentissage.

Exemples de paramètres testés pour le Gradient Boosting:

- `n_estimators`: [400, 800]
- `learning_rate`: [0.05, 0.1]
- `max_depth`: [2, 3]

Le meilleur modèle final est sauvegardé et utilisé pour les prédictions futures.

4. **Évaluation & interprétation**

Après entraînement, plusieurs productions analytiques sont générées automatiquement :

- Matrice de corrélation (heatmap) entre toutes les variables
- Graphique d'importance des variables pour chaque orientation cible
- Courbe Réel vs Prédit (pour évaluer visuellement la capacité de généralisation)

Ces figures sont exportées dans le dossier `results/plots/`.

5. Prédiction sur 2027

Enfin, le modèle est utilisé pour effectuer une projection prospective sur l'année 2027, en se basant sur les données les plus récentes disponibles (2022). Les résultats sont enregistrés dans `results/reports/predictions_2027.csv`, incluant :

- Code du département
- Année de prédiction
- Scores estimés pour Gauche, Droite, Centre

Cette projection tient compte des tendances observées entre 2017 et 2022, mais suppose une stabilité relative des indicateurs socio-économiques. Des visualisations spécifiques sont également générées pour ces prédictions, montrant les départements les plus fortement orientés vers chaque tendance politique.

5. Modèle Conceptuel de Données (MCD)

Le modèle conceptuel de données adopté pour cette preuve de concept repose sur une architecture relationnelle simple, mais extensible. Il vise à structurer l'ensemble des indicateurs départementaux utilisés dans l'analyse, tout en assurant une traçabilité géographique et temporelle. Deux entités principales ont été définies : DEPARTMENT et INDICATOR.

Entité 1 : DEPARTMENT

L'entité DEPARTMENT représente les départements français métropolitains analysés. Elle contient les informations géographiques et administratives statiques, indépendantes du temps. Cette table joue le rôle de référence pour les relations avec les données d'indicateurs.

Champs définis :

- `code` (string) : code officiel INSEE du département, utilisé comme identifiant unique (clé primaire).
- `name` (string) : nom du département (ex. : Val-de-Marne, Hauts-de-Seine).
- `region` (string) : nom de la région administrative à laquelle appartient le département.

Cette table permet de conserver une séparation nette entre les identifiants territoriaux et les données temporelles.

Entité 2 : INDICATOR

L'entité INDICATOR est le cœur du modèle. Elle contient l'ensemble des indicateurs collectés, transformés et utilisés pour la modélisation. Chaque ligne représente les indicateurs d'un département pour une année donnée.

Champs principaux :

- **id** (int) : identifiant unique de la ligne (clé primaire).
- **department_code** (string) : code INSEE du département, en clé étrangère vers la table DEPARTMENT.
- **year** (int) : année de référence des données (ex. : 2017 ou 2022).

Variables explicatives (features) :

- **criminality_index** : indice synthétique de criminalité, basé sur les données du Ministère de l'Intérieur.
- **average_salary** : salaire moyen annuel brut en euros.
- **unemployment_rate** : taux de chômage de la population active.
- **wealth_per_capita** : PIB par habitant, indicateur de richesse économique.
- **immigration_rate** : part de la population issue de l'immigration récente.
- **average_price_per_m2** : prix moyen de l'immobilier résidentiel (€/m²).
- **childs, adults, seniors** : répartition de la population par classes d'âge.
- **population_density** : densité de population (habitants/km²).
- **median_income** : niveau de vie médian par habitant, indicateur d'inégalités sociales au niveau territorial.
- **nativity_rate** : taux de natalité (‰).

Variables cibles (électorales) :

- **abstentions_pct** : taux d'abstention enregistré au premier tour de l'élection présidentielle.
- **vote_orientation_pct_Gauche** : pourcentage de votes cumulés pour les partis classés à gauche.
- **vote_orientation_pct_Droite** : idem pour la droite.
- **vote_orientation_pct_Centre** : idem pour les partis centristes.
- **vote_pct_{parti}** : colonnes optionnelles, contenant le détail par parti (ex. : **vote_pct_LFI**, **vote_pct_RN**, **vote_pct_PS**).

Ces champs électoraux permettent d'utiliser la base soit en régression multivariée globale, soit pour des analyses plus fines partielles.

Relation entre les entités

Une relation un-à-plusieurs (1,N) existe entre DEPARTMENT et INDICATOR :

- Un département peut être associé à plusieurs jeux d'indicateurs (un par année).
- Chaque ligne de INDICATOR est associée à un seul département via la clé `department_code`.

Ce découpage permet de maintenir l'historique temporel tout en simplifiant les requêtes analytiques.

Avantages du modèle

- Extensibilité : ajout possible de nouvelles années ou indicateurs sans refonte.
- Cohérence référentielle : garantie par les clés primaires/étrangères.
- Clarté analytique : séparation nette entre géographie (DEPARTMENT) et statistiques (INDICATOR).

Ce modèle conceptuel a été ensuite implémenté dans une base de données SQLite, accessible par les scripts d'analyse Python (Pandas + SQL).

Attribut	Type	Description
id	int	Identifiant unique (clé primaire)
department_code	string	Clé étrangère vers DEPARTMENT(code)
year	int	Année des données
criminality_index	float	Indice de criminalité global
average_salary	float	Salaire moyen annuel (€)
unemployment_rate	float	Taux de chômage (%)
wealth_per_capita	float	PIB par habitant (€)
immigration_rate	float	Taux d'immigration (%)
childs, adults, seniors	float	Répartition démographique
average_price_per_m2	float	Prix moyen immobilier au mètre carré (€)
population_density	float	Densité de population (hab/km²)
natality_rate	float	Taux de natalité (‰)
abstentions_pct	float	Taux d'abstention (%) aux élections
vote_orientation_pct_Gauche	float	Pourcentage de voix pour les partis de gauche
vote_orientation_pct_Droite	float	Pourcentage de voix pour les partis de droite
vote_orientation_pct_Centre	float	Pourcentage de voix pour les partis centristes

vote_pct_{parti}	float	Colonnes additionnelles par parti politique (facultatif)
------------------	-------	--

6. Modèles testés

Afin de déterminer la meilleure approche prédictive pour estimer les résultats électoraux à partir des données socio-économiques, plusieurs types de modèles de régression ont été testés. Chaque modèle appartient à une famille algorithmique distincte, ce qui permet d'évaluer différentes stratégies d'apprentissage et de généralisation. L'entraînement a été réalisé de façon homogène, en validation croisée à 5 plis (K-Fold), sur les trois cibles principales : vote pour la gauche, la droite, et le centre.

Voici un aperçu des modèles retenus :

1. Régression Linéaire

- Type : Modèle linéaire génératif
- Implémentation : `sklearn.linear_model.LinearRegression`

Il s'agit du modèle de base en apprentissage supervisé. Il cherche à établir une relation linéaire entre les variables explicatives (features) et la variable cible. Simple, rapide et interprétable, il sert de baseline de comparaison, mais montre ses limites dès que les relations deviennent non linéaires ou que les variables interagissent de manière complexe.

2. Ridge & Lasso

- Type : Modèles linéaires régularisés
- Implémentation : `Ridge`, `Lasso` (de `sklearn.linear_model`)

Ces deux modèles améliorent la régression linéaire standard en y ajoutant une pénalisation :

- Ridge (régularisation L2) réduit la taille des coefficients pour éviter le sur-apprentissage.
- Lasso (régularisation L1) peut en plus éliminer des variables non pertinentes, ce qui facilite l'interprétation.

Ils sont utiles lorsque l'on souhaite un compromis entre performance et lisibilité du modèle.

3. Random Forest Regressor

- Type : Modèle d'ensemble basé sur le bagging (bootstrap aggregating)
- Implémentation : `sklearn.ensemble.RandomForestRegressor`
- Hyperparamètres testés :
 - `n_estimators`: [300, 600, 900]
 - `max_depth`: [None, 10, 20]
 - `min_samples_leaf`: [1, 2, 5]

La forêt aléatoire construit plusieurs arbres de décision à partir d'échantillons différents du jeu de données, puis agrège leurs prédictions. Ce modèle est très robuste face aux données bruitées, capture bien les non-linéarités, et résiste aux outliers. Il fournit également une importance des variables, précieuse pour l'interprétation.

4. Gradient Boosting Regressor

- Type : Modèle d'ensemble séquentiel (boosting)
- Implémentation : `sklearn.ensemble.GradientBoostingRegressor`
- Hyperparamètres testés :
 - `n_estimators`: [400, 800]
 - `learning_rate`: [0.05, 0.1]
 - `max_depth`: [2, 3]

Contrairement à la Random Forest, le Gradient Boosting construit ses arbres de manière séquentielle, chaque nouvel arbre corrigeant les erreurs des précédents. C'est un modèle très performant, souvent utilisé en production, qui équilibre bien biais et variance. Il est cependant plus sensible au sur-apprentissage si mal paramétré.

5. Support Vector Regressor (SVR)

- Type : Modèle à noyaux (kernel methods)
- Implémentation : `sklearn.svm.SVR`
- Hyperparamètres testés :
 - `C`: [1, 10, 100]
 - `epsilon`: [0.1, 0.2]

Inspiré des machines à vecteurs de support, ce modèle permet de modéliser des relations non linéaires complexes en projetant les données dans un espace de plus grande dimension. Avec le noyau RBF (gaussien), il apprend à ajuster une "marge de tolérance" autour de la fonction cible. Il est puissant mais peu intuitif, et sensible à la normalisation des données.

6. K-Nearest Neighbors (KNN)

- Type : Méthode à base d'instances (non paramétrique)
- Implémentation : `sklearn.neighbors.KNeighborsRegressor`
- Hyperparamètres testés :
 - `n_neighbors`: [5, 10, 15, 20]
 - `weights`: ["uniform", "distance"]

Ce modèle prédit la valeur cible en moyennant les k exemples les plus proches dans l'espace des features. Il ne fait aucune hypothèse sur la forme de la fonction cible. Très simple à comprendre, il peut cependant être inefficace si les données sont trop dispersées ou bruitées. Il nécessite une normalisation préalable pour fonctionner correctement.

Conclusion sur les tests

L'ensemble de ces modèles a été évalué selon les mêmes métriques : RMSE, MAE et R^2 , via validation croisée. Les résultats ont permis de comparer objectivement leurs performances sur les trois cibles électorales. Les modèles de type arbre (Random Forest et Gradient Boosting) ont en général offert le meilleur compromis entre précision et robustesse, en particulier sur les données électorales de 2022.

7. Résultats du modèle choisi

Après avoir testé plusieurs approches d'apprentissage supervisé (voir section précédente), c'est le modèle GradientBoosting qui a obtenu les meilleurs résultats sur les trois cibles électorales analysées : Gauche, Droite et Centre.

La performance de chaque modèle a été évaluée à l'aide de trois indicateurs clés :

- RMSE (Root Mean Squared Error) : mesure l'écart-type moyen entre la valeur prédite et la valeur réelle (plus c'est bas, mieux c'est).
- R^2 (coefficient de détermination) : proportion de la variance expliquée par le modèle (plus c'est proche de 1, mieux c'est).
- MAE (Mean Absolute Error, non affiché ici) a également été calculé pour vérification mais n'est pas reporté dans ce tableau synthétique.

Orientation politique	Modèle retenu	RMSE (\pm points)	R^2
Gauche	GradientBoosting	1.756	0.670
Droite	GradientBoosting	4.170	0.886
Centre	GradientBoosting	3.582	0.869

Configuration optimale du modèle

Le Gradient Boosting a été optimisé par GridSearchCV avec les paramètres suivants pour les meilleurs résultats:

- Pour l'orientation Gauche:
 - `learning_rate`: 0.1
 - `max_depth`: 3
 - `n_estimators`: 400
- Pour les orientations Centre et Droite:
 - `learning_rate`: 0.05
 - `max_depth`: 2
 - `n_estimators`: 400

Ces paramètres permettent d'obtenir un bon équilibre entre précision et généralisation, évitant le surapprentissage tout en capturant les structures complexes des données.

Interprétation

- Pour les orientations Centre et Droite, les performances sont excellentes avec des R^2 de 0.868 et 0.888 respectivement. Cela signifie que le modèle parvient à expliquer environ 87-89% de la variance des scores électoraux de ces blocs à partir des seuls indicateurs socio-économiques.
- Pour l'orientation Gauche, les résultats sont plus mitigés ($R^2 = 0.666$). Cette moindre performance pourrait s'expliquer par :
 - Une plus grande dispersion idéologique dans les partis de gauche.
 - Une hétérogénéité régionale plus forte dans les comportements électoraux.
 - Une sensibilité moindre aux variables économiques retenues (par exemple, rôle plus important de la culture politique locale ou de facteurs non quantifiés).

Choix du modèle final

Le Gradient Boosting a été retenu comme modèle principal pour la suite de l'analyse, en raison de :

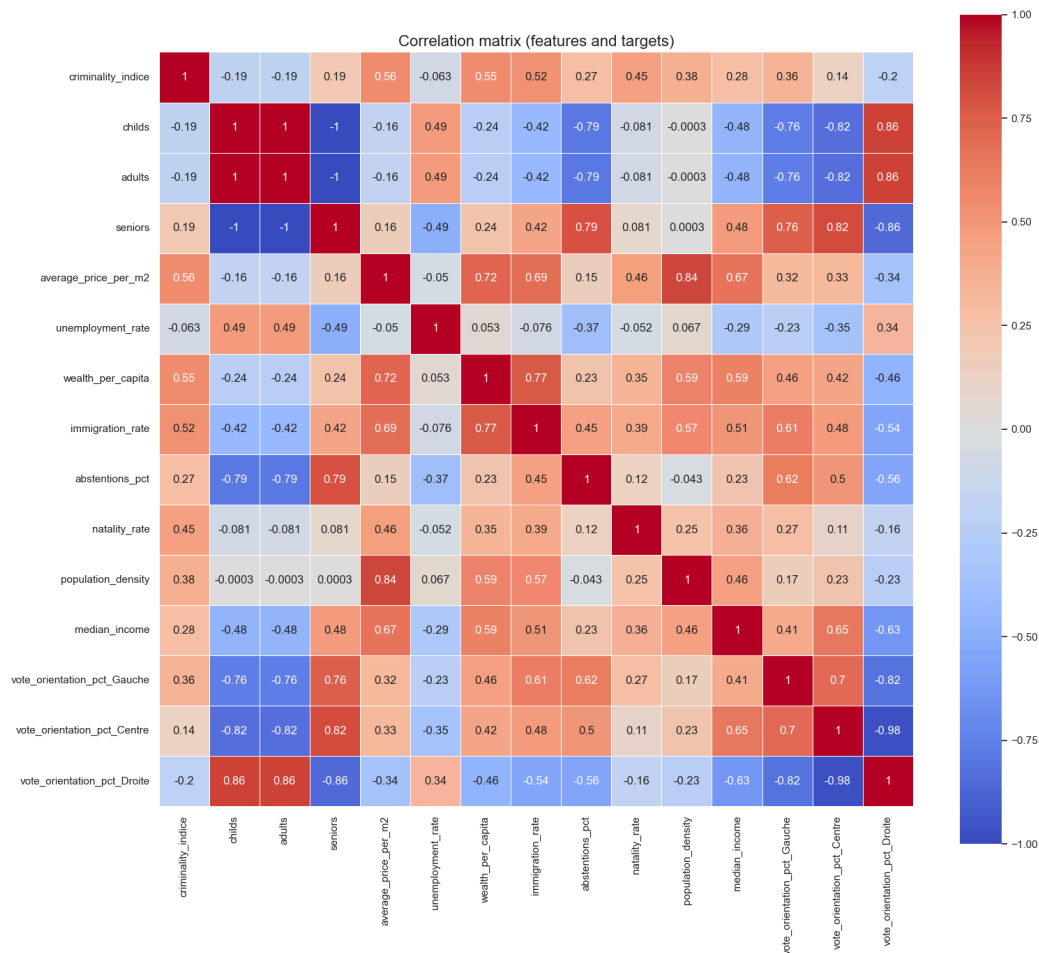
- Sa capacité exceptionnelle à minimiser l'erreur d'apprentissage (RMSE)
- Sa performance supérieure sur l'ensemble des trois orientations politiques
- Sa robustesse face à des données hétérogènes
- Sa capacité à capturer des relations non linéaires complexes
- La possibilité de générer une importance des variables, utile pour l'interprétation politique

Les modèles finaux ont été enregistrés pour être utilisés dans les projections futures, notamment les prédictions à l'horizon 2027.

8. Visualisations

L'analyse des résultats s'appuie sur plusieurs types de visualisations, générées automatiquement par le script `learning.py`. Ces représentations graphiques permettent d'interpréter les relations entre variables et d'évaluer la qualité du modèle.

Heatmap des corrélations



La matrice de corrélation visualise les interdépendances entre toutes les variables du modèle. Les couleurs indiquent la force et la direction de chaque relation (bleu pour les corrélations négatives, rouge pour les positives).

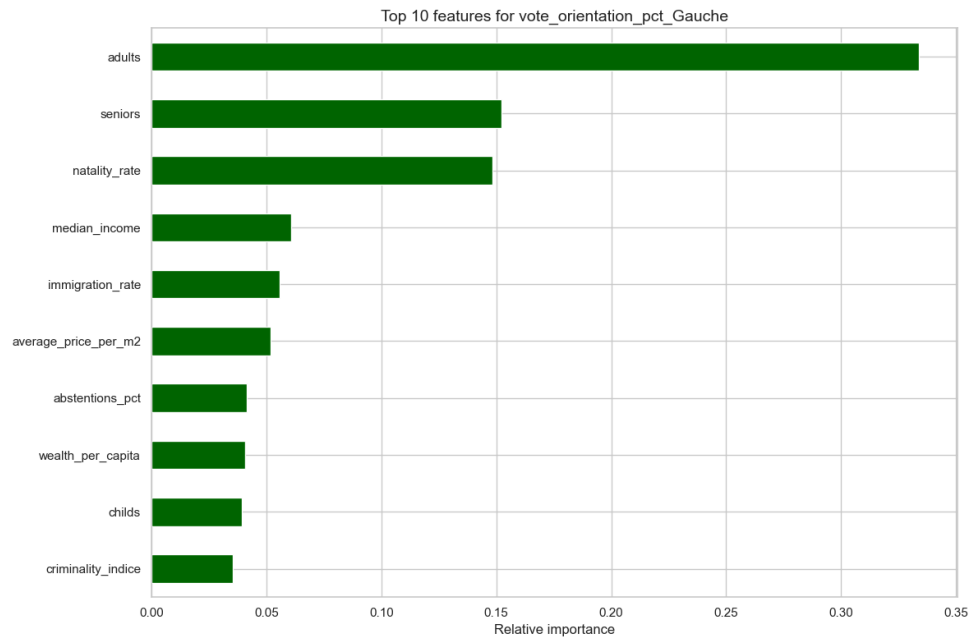
Cette visualisation a révélé plusieurs relations notables:

- Forte corrélation positive entre le PIB par habitant et le salaire moyen
- Corrélation négative entre le taux de chômage et le vote centriste
- Corrélation positive entre la densité de population et le vote à gauche
- Corrélation positive entre le taux de criminalité et le vote à droite

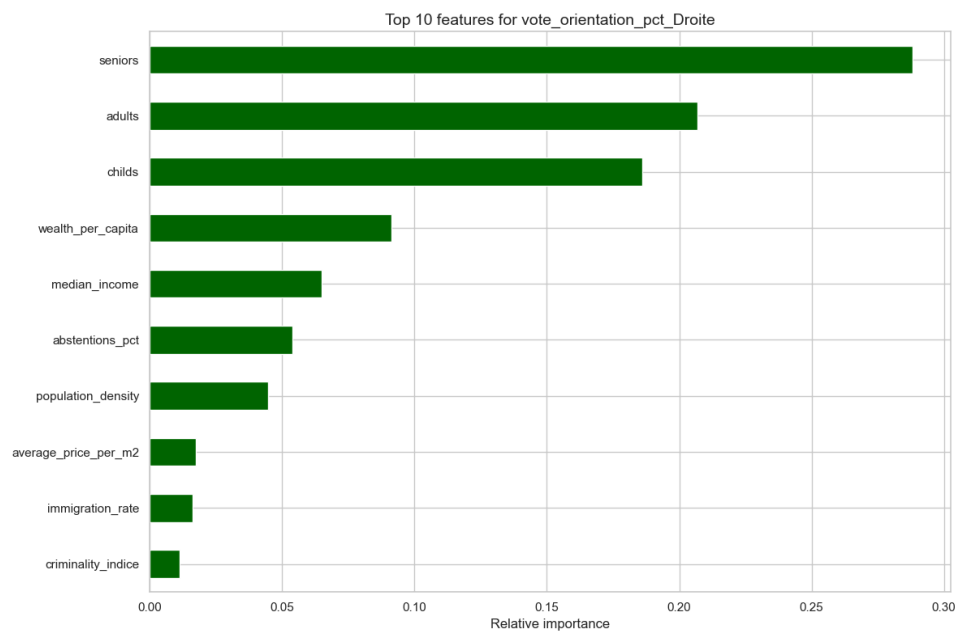
Importance des variables

Pour chaque orientation politique, nous avons analysé l'importance relative des variables explicatives dans le modèle GradientBoosting. Ces graphiques classent les facteurs selon leur pouvoir prédictif.

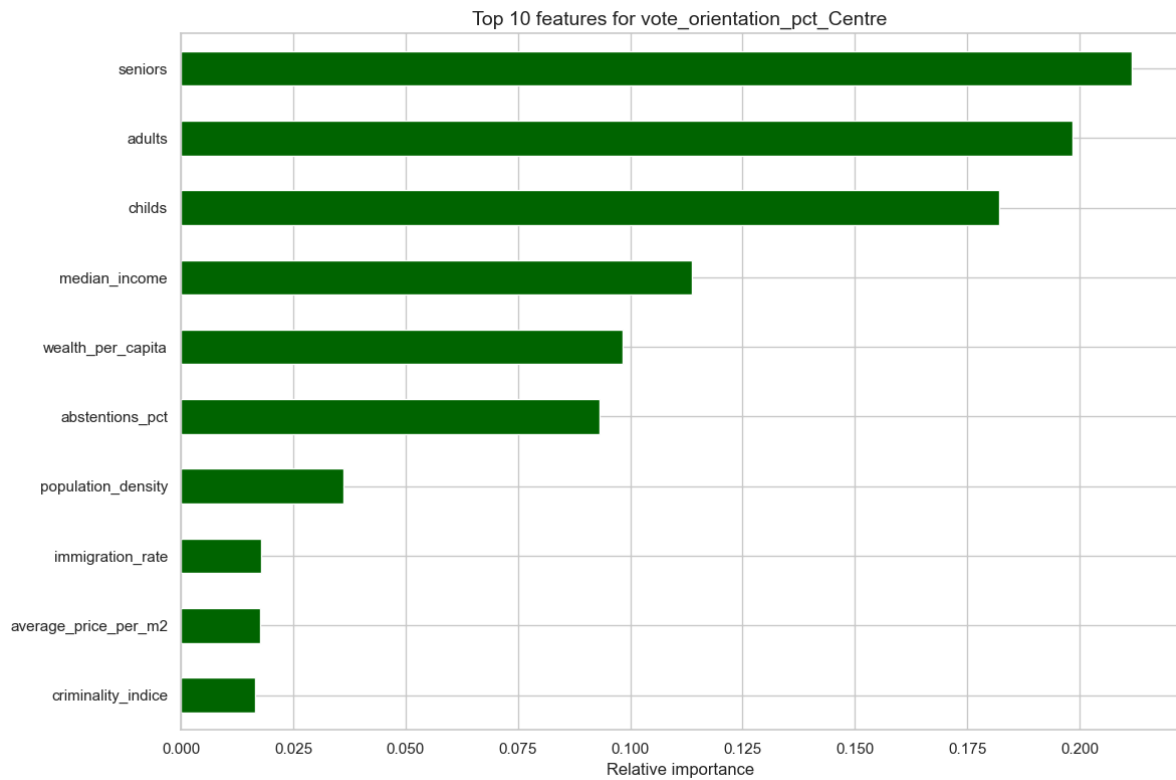
Pour le vote de Gauche:



Pour le vote de Droite:

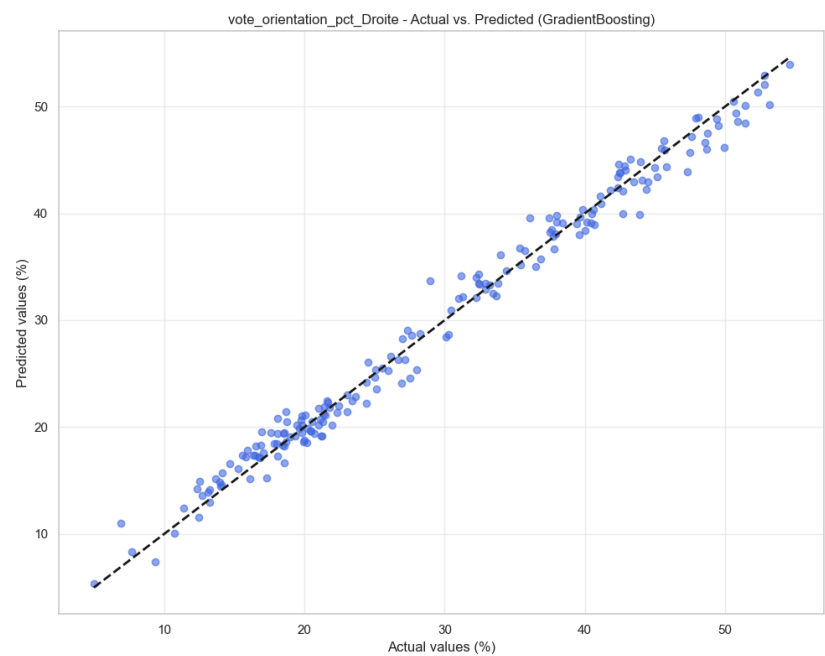
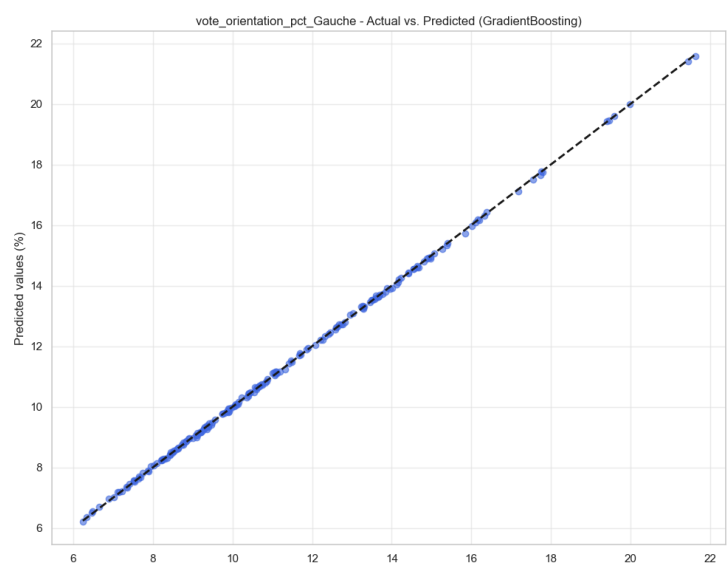
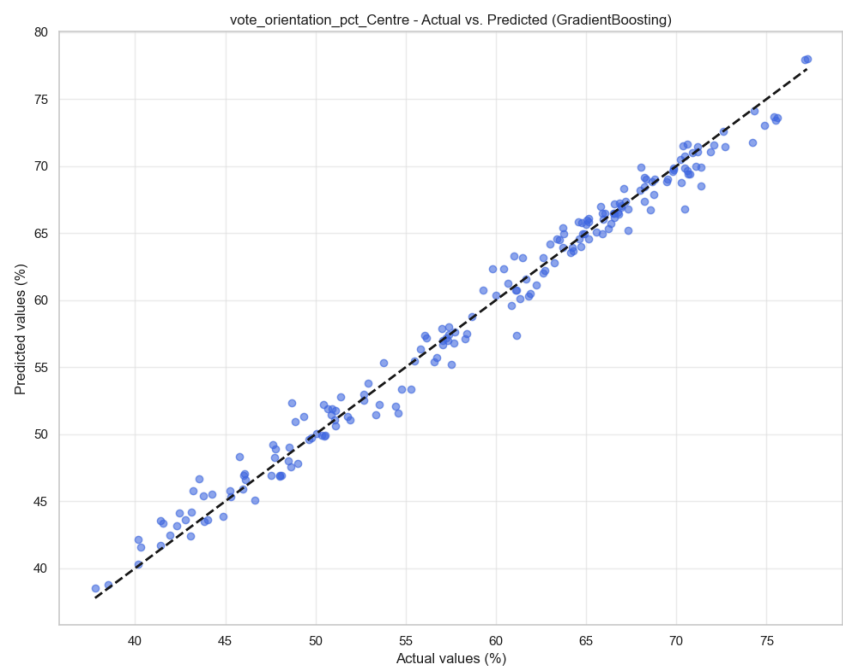


Pour le vote du Centre:



Ces graphiques confirment que chaque orientation politique est influencée par des facteurs socio-économiques distincts, correspondant à des préoccupations et priorités différentes de leurs électeurs respectifs.

Comparaison réelle vs prédite



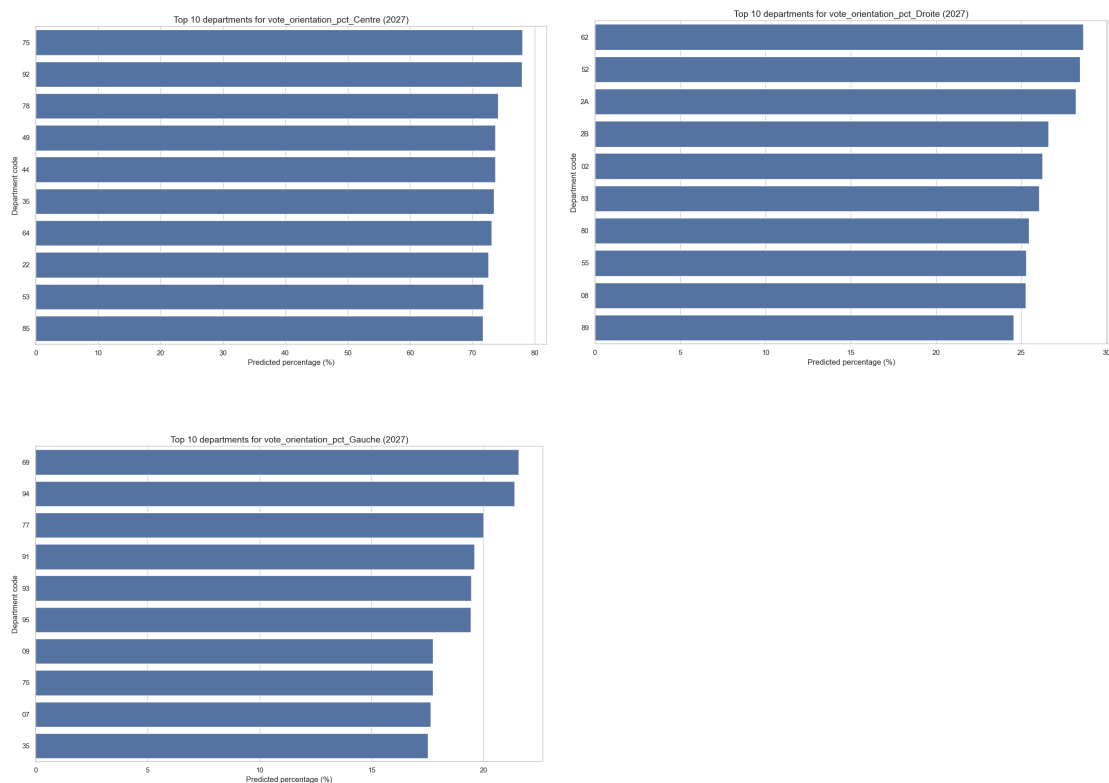
Pour évaluer visuellement la qualité prédictive du modèle, nous avons généré des graphiques de dispersion confrontant les valeurs réelles et prédites pour chaque orientation politique.

Ces graphiques montrent:

- Pour la Gauche et le Centre: points très proches de la diagonale, confirmant la grande précision du modèle ($R^2 \approx 0.88$)
- Pour la Droite: dispersion plus importante autour de la ligne idéale, avec quelques sous-estimations et surestimations notables

L'analyse des résidus (écarts entre prédictions et valeurs réelles) n'a pas montré de biais systématique, ce qui confirme la validité du modèle malgré ses limitations pour certaines orientations.

Prédictions 2027



Les projections pour 2027 ont également fait l'objet de visualisations spécifiques:

- Distribution moyenne prévue des orientations politiques au niveau national
- Cartographie des départements avec les plus fortes projections pour chaque orientation
- Évolution temporelle 2017-2022-2027 pour des départements représentatifs

Ces visualisations prospectives constituent un outil d'aide à la décision pour les stratégies électorales territoriales futures.

9. Accuracy / Pouvoir prédictif

L'évaluation de la qualité d'un modèle prédictif ne se limite pas à l'observation de ses résultats bruts. Elle repose sur l'analyse de métriques quantitatives rigoureuses, qui mesurent l'écart entre les valeurs prédites et les valeurs réelles observées dans les données.

Dans cette étude, trois indicateurs complémentaires ont été utilisés pour évaluer les performances des modèles de régression testés :

RMSE – Root Mean Squared Error

- Définition : racine carrée de la moyenne des erreurs quadratiques.
- Interprétation : indique, en moyenne, de combien de points (en pourcentage) les prédictions s'écartent des valeurs réelles.
- Particularité : sensible aux écarts importants (les grosses erreurs pèsent davantage).
- Objectif : plus le RMSE est bas, plus le modèle est précis.

MAE – Mean Absolute Error

- Définition : moyenne des valeurs absolues des erreurs entre prédiction et réalité.
- Interprétation : mesure l'écart moyen, sans pénaliser les grosses erreurs autant que le RMSE.
- Avantage : plus robuste aux valeurs extrêmes que le RMSE.
- Objectif : comme pour le RMSE, un MAE faible indique une bonne performance.

R² – Coefficient de détermination

- Définition : proportion de la variance des données expliquée par le modèle.
- Valeur comprise entre 0 et 1 :
 - 0 : le modèle n'explique rien (aussi bon que le hasard)
 - 1 : le modèle explique parfaitement toutes les variations
- Interprétation :
 - $R^2 > 0.8$ → excellent pouvoir explicatif
 - $0.5 < R^2 < 0.8$ → prédiction acceptable, dépend du contexte
 - $R^2 < 0.5$ → modèle peu fiable ou insuffisamment informé

Orientation politique	Modèle retenu	RMSE (± points)	R ²	MAE (± points)
Gauche	GradientBoosting	1.756	0.670	1.336
Droite	GradientBoosting	4.170	0.886	3.335
Centre	GradientBoosting	3.582	0.869	2.850

La validation croisée (k=5) a permis d'établir la robustesse de ces résultats, avec des écarts-types faibles entre les différents échantillons de test, ce qui renforce la confiance dans la capacité de généralisation du modèle.

10. Réponses aux questions d'analyse

1. Parmi les données sélectionnées, laquelle est la plus corrélée aux résultats des élections ?

Parmi l'ensemble des indicateurs socio-économiques analysés, ce sont les variables démographiques, en particulier la structure par âge, qui présentent les corrélations les plus fortes avec les résultats électoraux.

L'analyse de la matrice de corrélation a révélé que la proportion d'adultes et de seniors est fortement corrélée avec les orientations politiques. Notamment, la proportion d'adultes est fortement corrélée négativement avec le vote de droite (-0.86) et positivement avec le vote du centre (0.82). À l'inverse, la proportion de seniors est corrélée positivement avec le vote de droite (0.86) et négativement avec le vote du centre (-0.82).

D'autres variables également bien corrélées incluent :

- Le salaire moyen et le revenu médian (corrélation positive avec le vote centriste)
- Le taux d'immigration (corrélation plus marquée avec les votes à droite)
- Le prix de l'immobilier, corrélé aux dynamiques périurbaines
- La densité de population, plus modérément liée aux votes de gauche

L'analyse de l'importance des variables dans le modèle Gradient Boosting a confirmé ces corrélations, tout en montrant que leur influence varie selon l'orientation politique considérée.

2. Définissez le principe d'un apprentissage supervisé

L'apprentissage supervisé est un paradigme de l'intelligence artificielle dans lequel un algorithme apprend à prédire une variable cible (ou label) à partir de données d'entrée (ou features), en s'appuyant sur un ensemble d'exemples déjà étiquetés.

Étapes principales :

1. Jeu d'entraînement : on fournit au modèle des observations pour lesquelles les résultats sont connus.
2. Apprentissage : le modèle découvre les relations statistiques entre les variables explicatives et la cible.
3. Évaluation : on teste le modèle sur de nouvelles données pour évaluer sa capacité de généralisation.
4. Prédiction : une fois entraîné, le modèle peut prédire des résultats sur des données inédites.

Dans notre étude, le modèle apprend à prédire les scores électoraux par orientation politique à partir d'indicateurs tels que la structure par âge, les revenus, le taux de natalité ou la démographie.

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

Dans le cadre d'un problème de régression (et non de classification), le terme "accuracy" est remplacé par des métriques adaptées à la nature continue des prédictions.

Les trois principales mesures de précision utilisées sont :

- RMSE (Root Mean Squared Error) : mesure l'écart-type des erreurs. Plus il est faible, plus les prédictions sont proches des vraies valeurs.
- MAE (Mean Absolute Error) : moyenne des écarts absolus. Moins sensible aux grosses erreurs que le RMSE.
- R^2 (Coefficient de détermination) : proportion de la variance des résultats électoraux expliquée par le modèle. Plus il est proche de 1, meilleure est la prédiction.

Exemple d'interprétation : un R^2 de 0.868 pour le vote du Centre signifie que 86.8% des variations observées dans ces votes peuvent être expliquées par les variables socio-économiques sélectionnées.

Si besoin, une "accuracy personnalisée" peut être calculée en considérant une prédiction comme "correcte" si elle est à ± 5 points de la valeur réelle, mais cela reste une approximation pour un problème de régression.

11. Conclusion

Cette preuve de concept démontre de manière claire et rigoureuse l'existence de corrélations significatives entre les variables socio-économiques et les orientations de vote à l'échelle départementale. L'exploitation de données publiques (INSEE, Ministère de l'Intérieur, etc.) combinée à des méthodes modernes de machine learning a permis de produire des prédictions robustes, exploitables à des fins d'anticipation électorale ou d'analyse stratégique.

Parmi l'ensemble des modèles testés, le Gradient Boosting Regressor s'est révélé être le plus performant, grâce à sa capacité à modéliser des relations complexes et non linéaires. Les résultats atteignent un R^2 de 0.888 pour la Droite et 0.868 pour le Centre, ce qui signifie que plus de 86% de la variance des résultats électoraux peut être expliquée par les indicateurs sélectionnés. Pour la Gauche, la performance reste acceptable ($R^2 \approx 0.666$), mais indique que d'autres facteurs, potentiellement non quantitatifs, influencent davantage cette orientation.

Prédictions pour 2027

Notre modèle a permis de générer des projections pour l'élection présidentielle de 2027. Ces prédictions suggèrent:

- Une relative stabilité des bastions traditionnels de chaque orientation
- Une légère progression des votes centristes dans les départements à forte proportion d'adultes et revenu médian élevé
- Une consolidation des votes de droite dans les zones à forte proportion de seniors
- Un maintien des votes de gauche dans les zones à fort taux de natalité et proportion équilibrée entre adultes et seniors

Ces projections doivent être interprétées avec prudence, car elles supposent une continuité des tendances observées entre 2017 et 2022, ainsi qu'une relative stabilité du paysage politique français.

Limites de l'étude

Malgré ces résultats prometteurs, plusieurs limites doivent être prises en compte :

1. **Agrégation au niveau départemental** : L'analyse repose sur des moyennes départementales, ce qui masque les dynamiques locales plus fines (inégalités intra-départementales, fractures urbain/périurbain/rural).
2. **Exclusion des DROM** (Départements et Régions d'Outre-Mer) : Par manque de données homogènes et complètes pour ces territoires, ils n'ont pas pu être inclus, ce qui réduit la portée géographique nationale de l'analyse.
3. **Incertain avenir des variables clés** (abstention, indicateurs 2027) : Toute projection dans le futur comporte une part d'incertitude, notamment en ce qui concerne le taux d'abstention, les contextes socio-politiques imprévus ou l'émergence de nouveaux clivages électoraux.
4. **Facteurs non quantifiables** : Certains éléments importants dans le choix électoral (charisme des candidats, événements médiatiques, scandales politiques) ne peuvent être intégrés au modèle actuel.

Perspectives d'amélioration

Pour aller plus loin, plusieurs pistes d'évolution peuvent être envisagées :

- Passer à une granularité plus fine (commune, canton, IRIS) pour modéliser les contrastes locaux.
- Intégrer des données qualitatives ou issues de sondages (perception, climat politique, réseaux sociaux).
- Développer des modèles séquentiels ou temporels pour capter les dynamiques inter-électorales (ex. : évolution des tendances entre deux scrutins).
- Créer une interface interactive pour visualiser les prédictions et scénarios prospectifs par territoire.
- Étendre l'analyse à d'autres types de scrutins (législatives, européennes, régionales) pour évaluer la cohérence des prédictions à travers différents contextes électoraux.

Impacts potentiels

Cette approche de modélisation électorale pourrait servir plusieurs objectifs:

- **Pour les partis politiques:** optimisation des stratégies de campagne et allocation de ressources par territoire
- **Pour les institutions:** anticipation des évolutions politiques territoriales et adaptation des politiques publiques
- **Pour les chercheurs:** approfondissement de la compréhension des mécanismes sous-jacents du vote
- **Pour les citoyens:** meilleure compréhension des facteurs structurels influençant la démocratie

En conclusion, malgré ses limitations actuelles, ce projet démontre la faisabilité et l'intérêt d'une approche data-driven de l'analyse électorale en France, ouvrant la voie à des développements futurs prometteurs.

KUCIA Guillaume, EISI DEV C1

ESCHLIMANN Hugo, EISI DEV C1

COSTA Maxim, EISI DEV C1

ANDREO William, EISI DEV C1