**Rapport de projet de prédiction d’élections**

1. Prétraitement
2. Démographie

Après avoir chargé le jeu de données, nous avons entrepris l’étape de la sélection des colonnes pertinentes, notamment celles qui font référence aux dates des élections. Ensuite, nous avons remodelé la structure de la dataframe pour créer une version plus manipulable. Cette démarche nous a permis de créer une nouvelle dataset comportant les colonnes suivantes :

* "Libellé de la commune" : Cette colonne renferme les noms des communes de France, offrant ainsi un contexte géographique crucial pour notre étude.
* "Population" : Cette colonne contient le nombre d'habitants par commune, une information fondamentale pour comprendre les dynamiques démographiques et électorales.
* "Date" : Cette colonne indique l'année de référence des données de la dataframe, fournissant un repère temporel pour nos analyses et interprétations. Nous avions donc les données des années 2007 ;2012,2017 et de 2022.

En consolidant ces informations essentielles dans une structure de données épurée, nous avons simplifié notre processus d'analyse et mis en évidence les éléments clés nécessaires à notre étude sur les élections et leur corrélation avec la population des communes en France.

1. Education

Notre jeu de données initial concernait les taux de réussite à l'échelle nationale pour chaque année. Cependant, pour nos observations à l'échelle des communes, il était impératif de disposer de données réparties sur l'ensemble du territoire.

Ainsi, nous avons entrepris une démarche visant à agréger ces données nationales à l'échelle communale, afin d'obtenir une vision plus granulaire et représentative de la situation éducative dans chaque commune. Cette approche nous permettra de mieux appréhender les corrélations potentielles entre les résultats électoraux et le niveau d'éducation au sein des différentes communautés.

1. Génération de donnés d’education

En supposant que les taux de réussite aux différents baccalauréats (bac technologique, bac professionnel et bac général) suivent une distribution normale, nous avons élaboré une approche pour intégrer ces données dans notre analyse.

* Nous avons d'abord établi une fonction nommée "generate\_normal\_samples", prenant en paramètres la moyenne et l'écart-type d'une distribution normale, ainsi que la taille désirée de l'échantillon. Cette fonction a pour but de générer des échantillons conformes à une distribution normale.
* Pour attribuer les taux de réussite aux différents types de baccalauréat en fonction du nombre d'habitants dans chaque commune, nous avons élaboré un tableau contenant les poids associés à chaque commune. Ces poids servent de facteurs permettant d'ajuster de manière raisonnable les taux de réussite en fonction de la population de chaque commune.
* Enfin, nous avons généré les échantillons de taux de réussite pour chaque commune, en nous basant sur le type de baccalauréat, et les avons pondérés en fonction de leurs poids respectifs. Ces taux générés ont ensuite été intégrés dans la dataset "Niveau\_education", enrichissant ainsi notre analyse avec des données éducatives spécifiques à chaque commune.

1. Revenue

Dans le cadre du traitement des datasets concernant les revenus des foyers des différentes communes, nous avons initié notre démarche en effectuant une visualisation des données manquantes. Il est apparu que les cinq premières colonnes présentaient le moins d'éléments manquants. Afin d'éviter toute approximation à une échelle trop grande, risquant de biaiser notre étude, nous avons choisi de nous limiter à sélectionner les cinq premières colonnes de chaque dataset. Ensuite, pour chacune de ces cinq colonnes, nous avons identifié les valeurs manquantes que nous sommes remplacées par la médiane.

Pourquoi avons-nous opté pour la médiane comme substitut ? La médiane est préférable dans ce contexte car elle est moins sensible aux valeurs extrêmes ou aberrantes par rapport à la moyenne. En utilisant la médiane comme valeur de remplacement, nous préservons la distribution centrale des données et réduisons ainsi le risque de distorsion dans notre analyse causée par des valeurs atypiques.

Voici donc la colonne finale présente dans :

**'CODGEO' :** le code associer a chaque commune

**'NBMENFISC' :** Nombre de ménages fiscaux

**'NBPERSMENFISC12' :** Nombre de personnes dans les ménages fiscaux

**'MED' :** Médiane du niveau de vie (€)

1. Top of Form
2. Election

Nous avons consolidé les résultats des élections par candidat et par commune, en rassemblant des données telles que le nombre de voix recueillies, les bulletins nuls, etc. Cette consolidation a été réalisée en fusionnant toutes ces datasets en une seule, qui contient l'ensemble des communes ainsi que les performances de chaque candidat politique.

1. Fusion des datasets  
   Après avoir effectué le traitement préalable de chaque dataset, nous entamons la fusion. Cette opération se fonde sur le nom de la commune, contenu dans la colonne "Libellé de la commune", tout en tenant compte de la date à laquelle chaque donnée est affiliée.
2. Visualisation et Analyse

* A graph of a number of blue and white bars

  Description automatically generated with medium confidenceVisualisation

Figure : niveau de réussite au baccalauréat par commune et par type de BAC

A group of pie charts

Description automatically generated

Figure : résultats des élections aux sein de 5 communes les plus peuplés

**Corrélation entre les données démographiques et économiques.**

En 2012, lors des élections dans les communes de Lyon et de Nice, on observe sur le diagramme circulaire une grande part de voix (28%) pour François Hollande, comparativement à 26,9% pour Sarkozy. Cependant, lorsque l'on se réfère à la commune de Nice, nous constatons l'effet contraire avec une majorité de voix pour Sarkozy (33,7%) contre 22,4% pour Hollande. Plusieurs facteurs puissent expliquer ce revers de situation entre Lyon et Nice lors des élections de 2012.

Premièrement, prenons en compte le niveau d'éducation et les revenus médians par foyer dans ces deux villes. À Lyon, où le revenu médian par foyer est estimé à 21 669 euros, il est possible que la population ait un niveau d'éducation et des revenus salariaux plus élevés en moyenne. En revanche, à Nice, où le revenu médian par foyer est de 18 683 euros, la situation semble moins favorable sur le plan économique

Ensuite, regardons les préférences politiques associées à ces niveaux socio-économiques. Il est plausible que dans des régions avec un niveau d'éducation et des revenus salariaux plus élevés, les électeurs aient tendance à favoriser des candidats comme Nicolas Sarkozy, dont le parti est perçu comme représentant davantage les intérêts des classes aisées.

D'autre part, les populations à revenu moyen pourraient être plus enclines à soutenir des candidats comme François Hollande, dont le parti est souvent associé à des politiques sociales et économiques visant à soutenir les classes moyennes.

En somme, ces différences socio-économiques entre Lyon et Nice pourraient expliquer en partie le vote en faveur de Nicolas Sarkozy à Nice et de François Hollande à Lyon lors des élections de 2012. Cependant, d'autres facteurs tels que les dynamiques locales et les enjeux politiques spécifiques de chaque région pourraient également avoir joué un rôle dans ce schéma de vote.

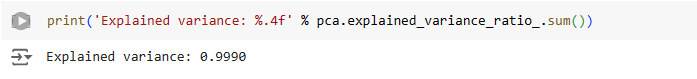
1. Nettoyage de la dataset

A grid of squares with a grid of lines

Description automatically generated with medium confidence Après avoir fusionné toutes les datasets, nous avons obtenu une dataset générale contenant 19 colonnes. La question est de savoir si toutes ces 19 colonnes sont nécessaires pour prédire le résultat. Pour répondre à cette question et simplifier le modèle, nous pouvons utiliser une méthode de nettoyage appelée Analyse en Composantes Principales (ACP).

Figure 3: heatmap des features

Avant d'utiliser l'ACP, il est nécessaire de standardiser les valeurs présentes dans la dataset. Cela permet d'éviter que l'échelle de certaines colonnes n'impacte disproportionnellement le modèle. En standardisant les données, nous ramenons les valeurs à une échelle commune, ce qui facilite la comparaison entre les différentes caractéristiques et améliore les performances de l'ACP ainsi que d'autres algorithmes d'apprentissage automatique.



Nous avons utilisé la méthode d'analyse en composantes personnelles pour sélectionner les 15 meilleures composantes principales qui expliquent 99,90% des variations présentes dans les données. En choisissant ces composantes, nous avons réussi à capturer l'essentiel de la structure et des motifs des données, tout en réduisant la dimensionnalité de manière significative. Cela nous permet de travailler avec un ensemble de caractéristiques plus compact et plus représentatif, ce qui peut simplifier les calculs et améliorer les performances des modèles prédictifs.

1. Model de predictions

Notre ensemble de données comprend les résultats des élections pour différents candidats dans chaque commune, ce qui génère une grande quantité de données (3,6 Go) si nous voulons utiliser les données des années 2012 à 2022. L'objectif principal est de prédire qui remportera les élections à l'échelle de chaque commune.

Pour ce faire, nous avons créé une colonne "label" dans notre ensemble de données. Dans cette colonne, chaque candidat dans une commune est étiqueté comme "gagnant" s'il obtient le plus grand nombre de voix dans cette commune, sinon il est étiqueté comme "perdant". Nous avons ensuite converti cette colonne de chaînes de caractères en données binaires, qui seront utilisées pour l'entraînement et la prédiction du modèle.

1. Arbre de décision

Les arbres de décision peuvent capturer des relations non linéaires entre les variables et la variable cible. Ils peuvent identifier des motifs complexes dans les données sans nécessiter de spécification explicite de la forme de la relation, ce qui nous pousse à l’implémenter dans notre cas.

1. **Resultats**

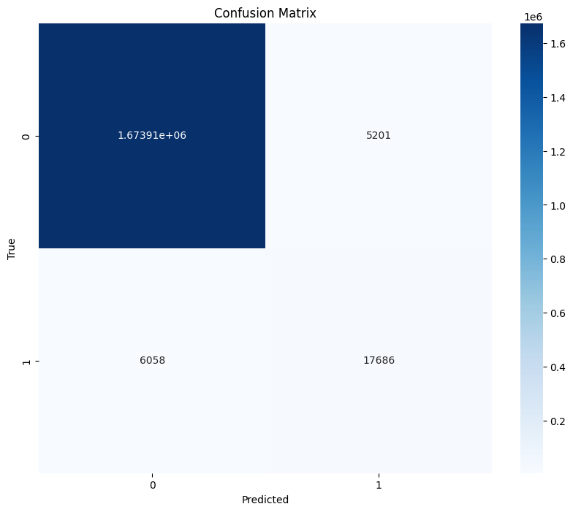


Figure 4: matrice de confusion de l'arbre de décision

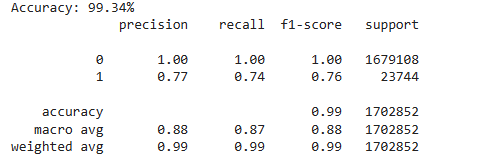


Figure 5: performance de l'arbre de décision

**Interpretation :**

La table de performance nous permet de visualiser des valeurs très acceptables de prédiction pour le F1-score, la précision et le rappel, qui sont des métriques clés pour la prise de décision d'un bon modèle. La matrice de confusion nous offre également une visibilité sur le nombre de bonne prédiction réalisé par notre modèle.

1. Régression logistique

**Resultats**

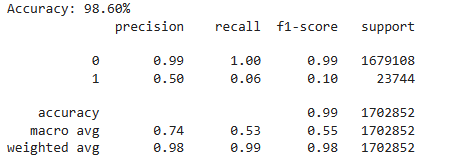


Figure 6: performance de la régression logistique

A blue squares with white text

Description automatically generated

Figure 7: matrice de confusion de la régression logistique

1. Réseau de neurone

A screenshot of a computer screen

Description automatically generated

Figure 8: réseau de neurone performance

A graph with numbers and a blue square

Description automatically generated with medium confidence

Figure 9: matrice de confusion réseau de neurones

**Interpretation**

Après avoir examiné en détail les performances de différents modèles pour la prédiction des résultats des élections à l'échelle des communes, nous constatons que l'arbre de décision pour la classification offre les meilleurs résultats de prédiction possibles pour chaque classe : gagnant (1) et perdant (0). En comparaison avec les performances des autres modèles, les performances de prédiction pour la classe (1) sont particulièrement importantes car elles représentent les résultats corrects des élections, c'est-à-dire prédire avec précision le candidat gagnant dans chaque commune.

En analysant les matrices de confusion pour chaque modèle, nous pouvons visualiser le nombre de prédictions correctes et incorrectes pour chaque classe. Ces matrices mettent en évidence que l'arbre de décision offre les performances les plus élevées en termes de prédictions correctes pour la classe (1), indiquant une capacité plus forte à identifier les gagnants des élections au niveau de la commune.

Il est important de souligner que la qualité des prédictions pour la classe (1) est cruciale pour l'objectif de notre modèle, qui vise à prédire avec une forte probabilité les résultats des suffrages à l'échelle des communes. En considérant cette importance, l'arbre de décision émerge comme le meilleur choix parmi les trois modèles évalués.

En résumé, l'arbre de décision pour la classification surpasse la régression logistique et les réseaux de neurones en termes de capacité à prédire avec précision les résultats des élections au niveau de la commune. Cette conclusion est étayée par l'analyse des matrices de confusion, qui mettent en évidence les performances supérieures de l'arbre de décision pour la prédiction des résultats des élections.