UNIVERSITÉ
LAVAL

Rapport : Analyse et prétraitement des données

Driouich Meryam
Maaou Marouane
El ass Walid

Table des matières

1 Introduction

Au sein d'un processus électoral, on s'intéresse à faire une étude sur les réponses de la population canadienne, un sondage qui était fait en 2019 au but d'avoir plus de renseignements sur les électeurs, l'opinion publique du peuple canadien afin de pouvoir améliorer le processus de l'élection.

Nos données sont une étude électorale canadienne produite par le consortium de la démocratie électorale, basée sur la réponse de sondage par des électeurs à travers Canada à propos de l'élection fédérale de 2019.

Le but de notre projet est de pouvoir prédire la position politique d'un individu en utilisant ses données.

2 Analyse préliminaire des données

2.1 Découverte des données

Une analyse d'un premier pas va consister à comprendre plus nos données et se familiariser avec, comme étant dit que notre but est de pouvoir prédire pour quel parti chacun de ces individus veut voter, l'attribut *cps_votechoice* est une réponse sur cette question.

Le diagramme ci-dessous donne une idée sur la distribution des partis fédéraux.

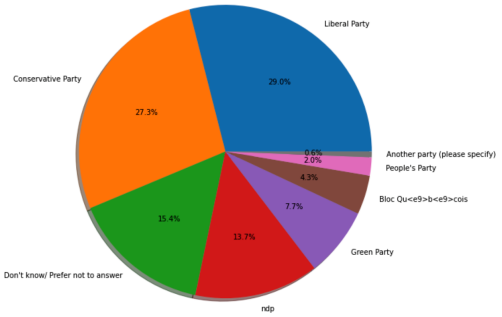


Figure 1 : Distribution des partis fédéraux

Afin de pouvoir faire une bonne prédiction, on doit essayer de comprendre l'opinion publique de nos électeurs, qui englobe soit les personnes qui sont des Canadiens ou des résidents permanents qui ne seraient pas capables de voter dans cette élection, mais on trouve un autre attribut *cps_votechoice_pr* est le choix du parti politique des résidents permanents, pour conserver plus de données et ne pas perdre des informations pertinentes, on concatène les deux attributs pour avoir un seul attribut qui répond à notre objectif (dans toute notre analyse on sera prudents de choisir les attributs qui sont en commun entre les questions ciblées aux Canadiens et aux résidents permanents).

2.2 Fréquence des partis fédéraux par rapport aux citoyennetés

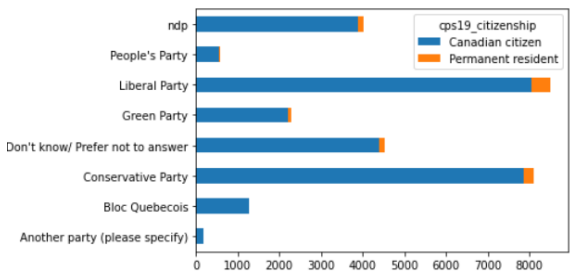


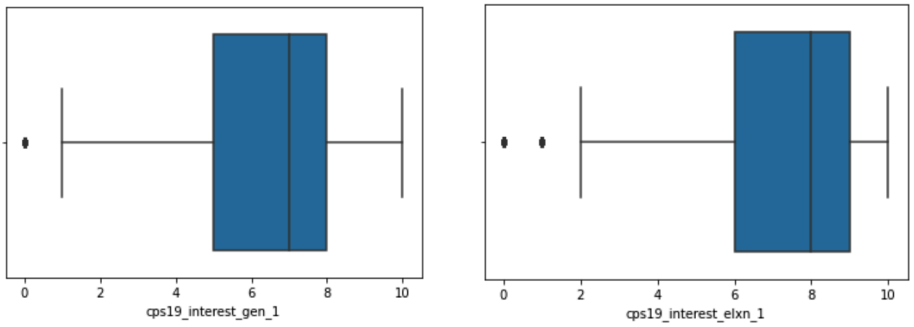
Figure 2 : Partis fédéraux par rapport aux citoyennetés

Le nombre des personnes de citoyenneté canadienne est grand par rapport au nombre des personnes qui sont des résidents permanents. Après la concaténation des deux attributs, on trouve la figure ci-dessus. On constate qu'une grande partie de la population canadienne (canadien et résident permanent) a l'intention de voter sur **le parti libéral**. D'autre part on ne trouve pas un grand déséquilibre dans les classes, la fréquence des partis fédéraux est bien distribuée(ne cause aucun problème dans notre objectif)

2.3 Quel est votre intérêt pour la politique et cette élection fédérale ?

Avant d'analyser en détails les réponses dans le sondage, il est important de voir l'intérêt des participants pour la politique et pour cette élection fédérale, plus ils sont intéressés plus leurs réponses sont fiables. Les participants choisissent un chiffre entre 0 (pour indiquer qu'ils n'ont aucun intérêt) à 10 (pour indiquer qu'ils ont beaucoup d'intérêt).

L'analyse sur 31 651 personnes nous donne une moyenne de 6.45, une médiane de 7, un mode de 7 aussi, le premier quartile Q1 est 5, le troisième quartile Q3 est 8 d'où la distance interquartile IQR = Q3 - Q1 est 3 pour la variable *cps19_ interest_ gen_ 1* qui exprime l'intérêt pour la politique. Du même, on a eu comme moyenne 7.08, une médiane de 8, un mode de 7, le premier quartile Q1 est 6, le troisième quartile Q3 est 9 d'où la distance interquartile IQR = Q3 - Q1 est 3 pour la variable *cps19_ interest_ elxn_ 1* qui exprime l'intérêt pour cette élection fédérale.



Selon les diagrammes de moustache, 0 est une valeur aberrante pour la variable *cps19_ interest_ gen_ 1* de l'autre côté 0 et 1 sont des valeurs aberrantes pour la variable *cps19_ interest_ elxn_ 1.* Les valeurs manquantes des deux variables sont remplacées par le **mode** de chaque variable. D'après notre analyse, on a trouvé que 75% des personnes qui ne sont pas intéressées par la politique ne sont pas aussi intéressées par cette élection fédérale.

2.4 Que pensez-vous des chefs de partis fédéraux ?

Dans cette question les participants expriment leur point de vue à propos de chaque chef de parti politique en choisissant un chiffre entre 0 (pour indiquer qu'ils ne l'aiment pas) et 100 (pour indiquer qu'ils l'aiment).

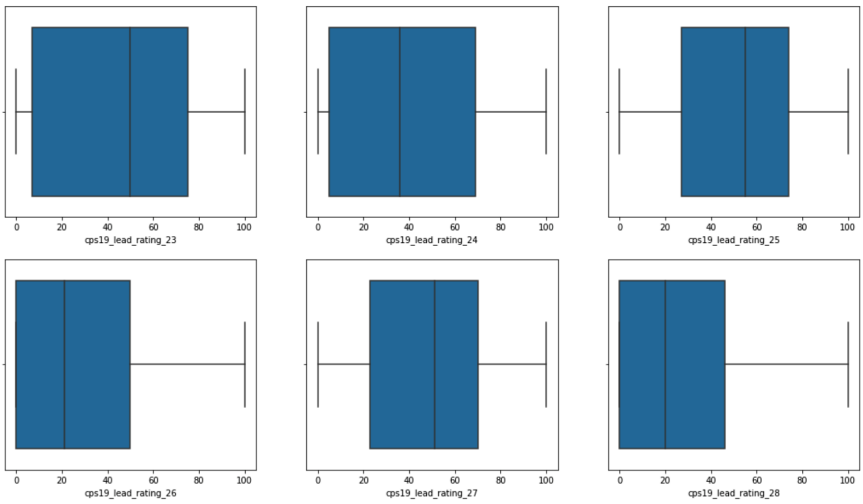


Figure 4 : Diagramme à moustache des chefs des partis fédéraux *cps19_ lead_ rating*

Dans les diagrammes ci-dessus, on remarque premièrement l'absence des valeurs aberrantes, ainsi on peut aussi remarquer que le chef du parti NDP *Jagmeet Singh* est le plus aimable avec une moyenne de 50.57 et une médiane de 55. Dans notre jeu, on a éliminé les 733 lignes qui n'ont pas noté tous les chefs des partis politiques et on a remplacé les autres valeurs manquantes par le mode de chaque variable selon la province du participant. La figure ci-dessous nous montre la distribution des chefs politique par province de chaque participant.

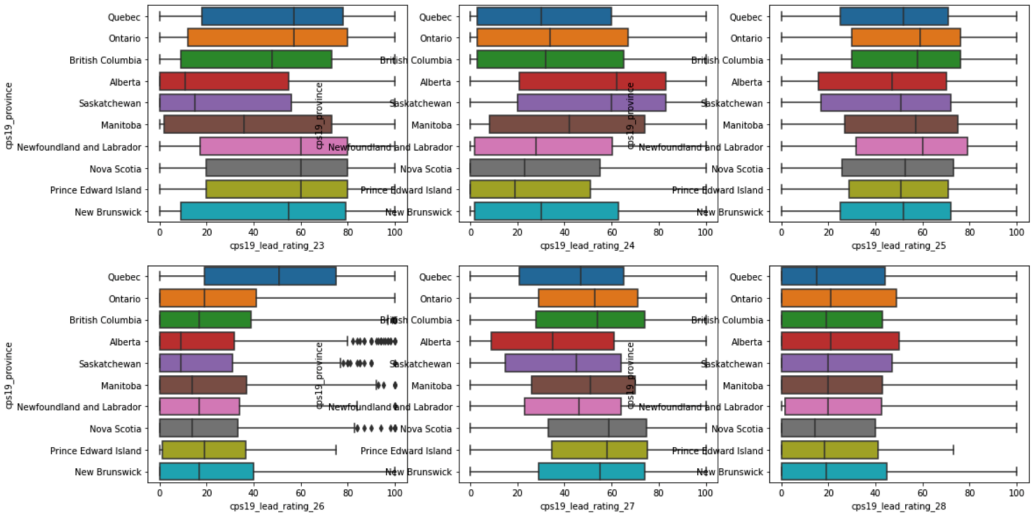


Figure 5 : Diagramme à moustache des chefs politiques selon la province *cps19_ lead_ rating*

2.5 Quel parti a le plus de chances de gagner le siège dans votre circonscription ?

Comme la section précédente, les participants un chiffre entre 0 (pour indiquer que le parti n'a pas de chance de gagner dans leur circonscription) et 100 (pour indiquer que le parti a de la chance de gagner dans leur circonscription). La figure ci-dessous nous montre la chance des partis politique par province de chaque participant.

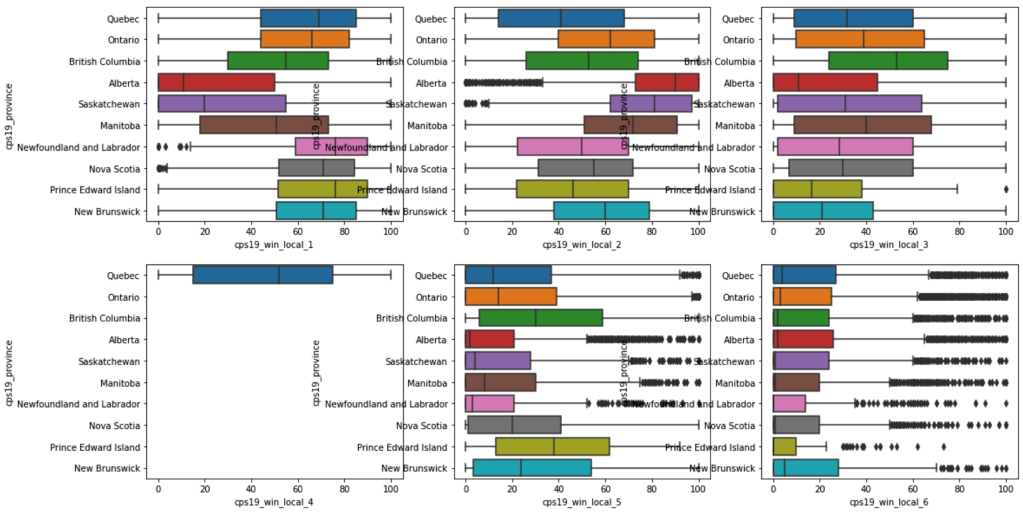


Figure 6 : Diagramme à moustache des partis politiques selon chaque province *cps19_win_local*

On remarque qu'il y a beaucoup de valeurs manquantes pour les participants hors de la province de Québec pour le parti *Bloc Québécois* (*cps19_win_local_4*), ces valeurs ne sont pas manquantes par hasard, il s'agit d'un cas où le *Bloc Québécois* a la chance de gagner seulement dans la province de Québec. Pour résoudre ce problème, on a remplacé les valeurs manquantes par des valeurs qui suivent la distribution de la variable *cps19_win_local_4* dans la province de Québec, la figure ci-dessous présente cette densité.

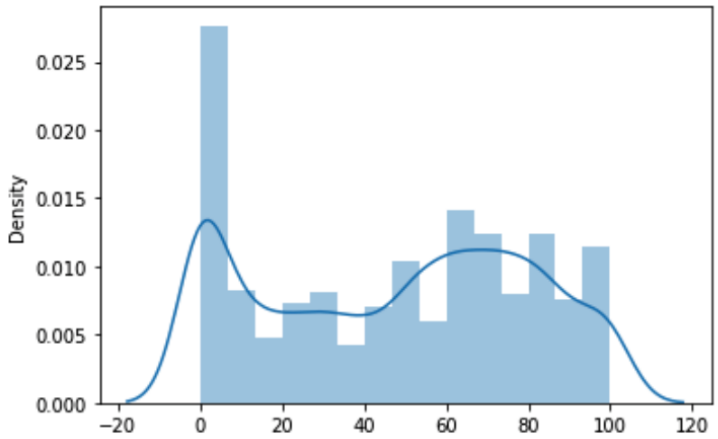


Figure 7 : Densité de la variable *cps19_win_local_4* dans la province de Québec.

2.6 Corrélation entre les variables

Avec la même logique des deux dernières sections, on analyse la variable *cps19_party_rating* dont laquelle les participants donnent leur avis à propos des partis fédéraux, la variable *cps19_cand_rating* qui note les candidat(e)s dans la circonscription locale des participants et la variable *cps19_most_seats* qui indique les chances qu'un parti gagne le plus grand nombre de sièges à la chambre des communes. La figure ci-dessous nous donne la matrice de corrélation entre ces variables pour les partis **Parti Libéral** et **Parti Conservateur**.

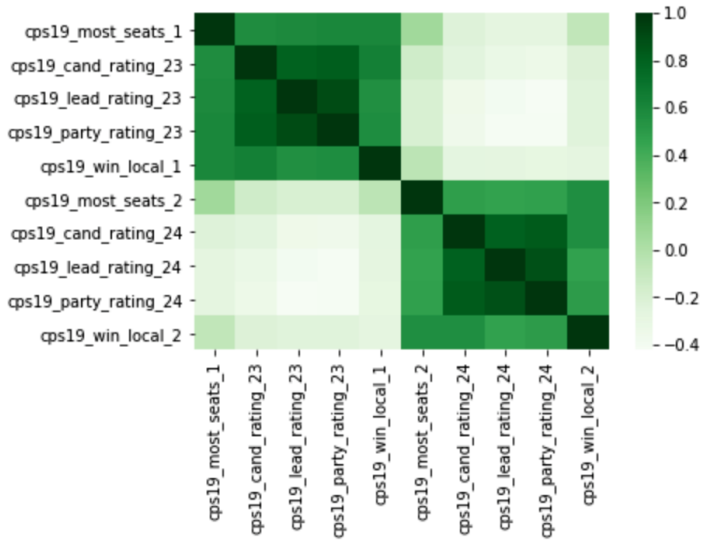


Figure 8 : Matrice de corrélation des variables *cps19_party_rating*, *cps19_most_seats*, *cps19_win_local*, *cps19_lead_rating* et *cps19_cand_rating*.

2.7 Distribution de la fréquence des partis politiques par rapport à l'âge des électeurs et leurs sexes

Une analyse de la fréquence des partis politiques par rapport à l'âge des électeurs et leurs sexes va être aussi utile. Cela revient à voir lesquels des partis politiques sont intéressants selon certaines tranches d'âge et/ou un sexe spécifique.

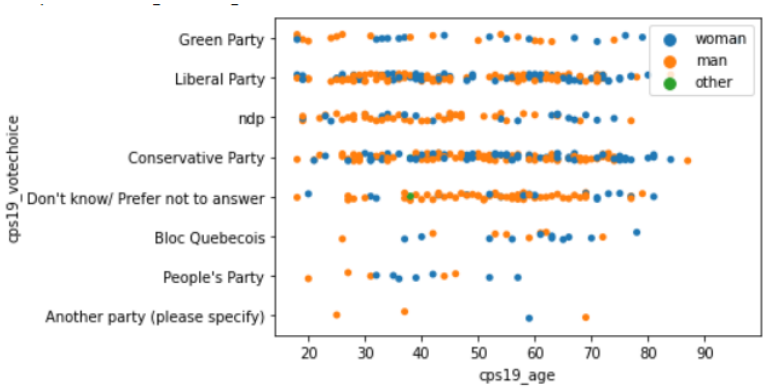


Figure 5 : Distribution de la fréquence des partis politiques par rapport à l'âge des électeurs et leurs sexes

On peut voir selon la figure ci-dessus qu'il y a une densité des points dans certains cotés, cela revient à dire qu'il y a plus de personnes dans cette tranche d'âge et la différence des couleurs signifie les différents types de sexe.

2.8 Distribution de la densité des électeurs par rapport aux tranches d'âge et la durée pour remplir le sondage

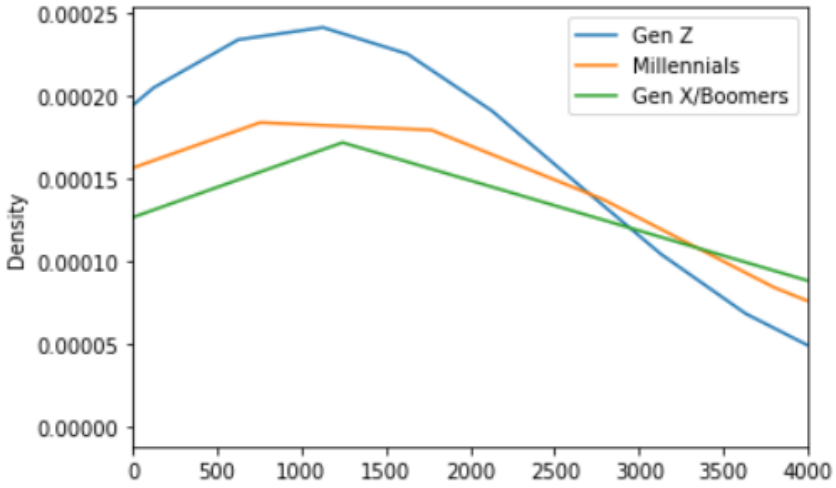


Figure 6 : Distribution de la densité des électeurs par rapport aux tranches d'âge et la durée

Pour la figure 6, on a utilisé deux attributs, **cps19_Q_TotalDuration** contient le temps écoulé pour terminer le remplissage du sondage, **cps19_yob** contient l'année de naissance des participants avec lequel on a pu extraire leurs âges dont on a divisé selon les tranches suivantes :

- **GenZ :** tranche d'age entre 18 ans et 25 ans.
- **Millennials :** tranche d'age entre 26 ans et 41 ans.
- **GenX/Boomers :** tranche d'age plus de 42 ans.

On remarque que la génération Z a pris moins de temps pour terminer le sondage en comparant aux autres générations.

2.9 Analyse des données textuelles

Dans cette section, on va s'intéresser aux données textuelles, plus précisément par l'attribut **cps19_imp_iss** qui contient des réponses des participants sur l'enjeu le plus important selon eux dans cette élection fédérale.
Pour simplifier l'analyse de cet attribut, on a choisi de catégoriser les enjeux sous forme des grandes catégories par exemple : 'Économique', 'Santé', 'Environnement'...

Comme première étape, on a fait un prétraitement de l'attribut en enlevant toutes les ponctuations pour nettoyer les données. Ensuite, on a utilisé une bibliothèque **fuzzyWuzzy** pour calculer la similarité entre la liste des enjeux (liste contient des types de problèmes sociaux reconnus) et les données fournissent par les participants.

La figure ci-dessous montre la fréquence des enjeux dans nos données.

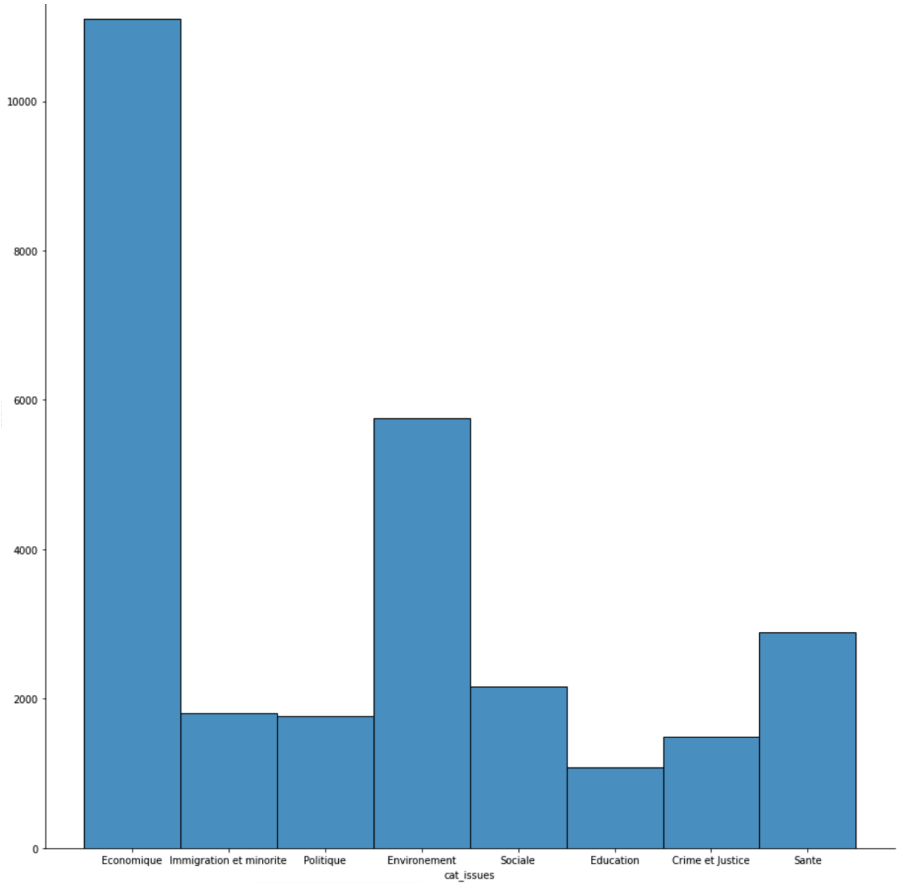


Figure 7 : Fréquence des enjeux

On remarque que les attributs les plus fréquents sont :

- Économique
- Environnement
- Santé

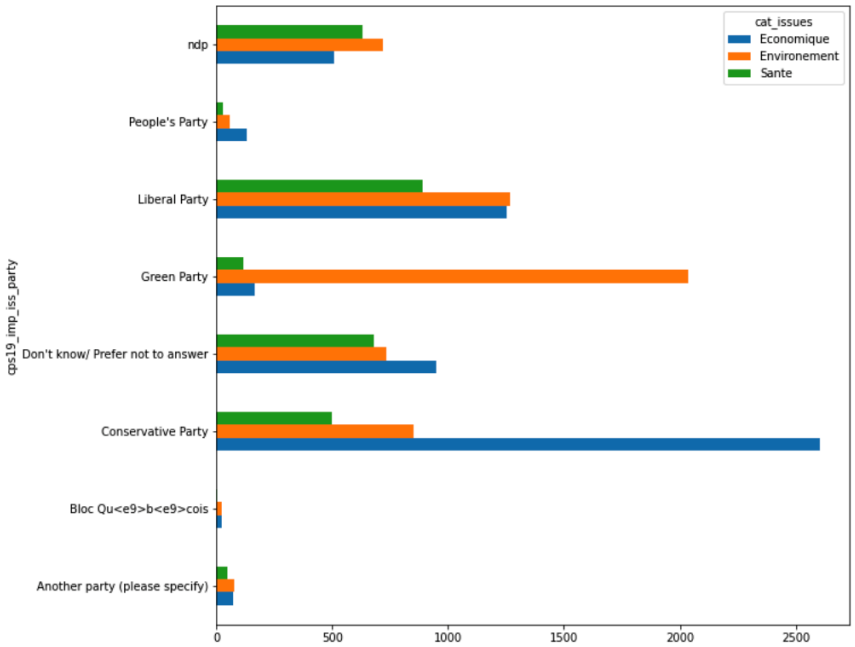


Figure 7 : Fréquence de la résolution des partis politiques des enjeux

Maintenant, on visualise l'attribut cps19_imp_iss_party qui contient les votes des participants sur les partis politiques qui résolvent leurs problèmes qu'on a classifié sous forme des catégories comme on avait déjà cité.

La figure ci-dessus montre la fréquence de la résolution des partis politiques des enjeux(ici on choisit juste 3 catégories les plus fréquents).

3 Attributs choisis

Les attributs choisis peuvent être regroupés selon les catégories suivantes :

— **attributs démographiques** :

- *cps19_win_local* : qui représente la chance qu'un parti politique gagne le siège dans le circonscription du participant.

- *cps19_cand_rating* : qui représente le point de vue des participants par rapport aux partis politiques dans leurs circonscriptions.

- *cps19_province* : qui indique la province de chaque participant.

— **méta-attributs** :

cps19_imp_iss_party : qui contient le choix du parti politique qui peut résoudre l'enjeu indiqué par répondant, Certainement ce choix va bien être corrélé l'élection du parti fédéral.

cps19_Q_TotalDuration : le temps écoulé pour terminer le remplissage des données qui nous donne une idée sur l'intérêt et la satisfaction du répondant qu'on peut tirer des informations sur le choix du vote.

cps19_lead_int et *cps19_lead_strong* : dans lesquelles les participants choisissent les chefs des partis fédéraux qui sont intelligents et manifestent un leadership fort.

— **attributs linguistiques** :

cps19_imp_iss : qui contient les enjeux des participants qu'on a classifié sous forme des grandes catégories et qui auront une relation directe avec le choix du parti politique.

4 Modèle

Notre objectif du projet est de pouvoir prédire le parti politique que chacun des individus veut voter, le choix du vote sera obligatoirement limité par le nombre des candidats. Alors on peut conclure qu'on affronte un problème de classification.

Après notre analyse des données qui nous a aidé à faire un choix pertinent des attributs, dont on sera capable de faire une bonne prédiction avec un bon choix de modèle qui va s'adapter à nos attributs et notre problème.

Le modèle de classification envisagé est un modèle d'apprentissage par ensemble. Nous prévoyons utiliser des différents modèles de classification. Le but étant d'avoir des résultats qui sont plus interprétables, ainsi la combinaison des prédictions de plusieurs algorithmes induise à des prédictions plus précises que n'importe quel modèle individuel. Cela est réalisé grâce à la librairie *sklearn.ensemble*.

Les modèles choisis sont notamment KNN(K plus proche voisin), Decision Tree, GradientBoosting-Classifier, qui est un modèle additif de manière progressive.

Nous utiliserons aussi Random Forest (Bagging technique) en construisant une multitude d'arbres de décision au moment de l'apprentissage.

4.1 Mesure de confiance sur les scores estimées

Une bien meilleure façon d'évaluer les performances d'un classifieur est d'examiner la matrice de confusion. Elle compare les données réelles pour une variable cible à celles prédites par le modèle. Elle permet de connaître d'une part les différentes erreurs commises par un algorithme de prédiction, mais plus important encore, de connaître les différents types d'erreurs commis. En les analysant, il est possible de déterminer les résultats qui indiquent comment ces erreurs ont eu lieu. Les résultats d'une matrice de confusion sont classés en quatre grandes catégories : les vrais positifs (TP), les vrais négatifs (TN), les faux positifs (FP) et les faux négatifs (FN).

Différentes métriques peuvent être calculées à partir du tableau de contingence afin d'en faciliter l'interprétation. C'est par exemple le cas de la précision, du rappel et du F1 score. Ces indicateurs permettent de mieux apprécier la qualité de précision du modèle.

— **Précision** répond à la question suivante : sur tous les enregistrements positifs prédits, combien sont réellement positifs ?

$$\text{Précision} = \frac{TP}{TP+FP}$$

— **Rappel** permet de répondre à la question suivante : sur tous les enregistrements positifs, combien ont été correctement prédits ?

$$\text{Rappel} = \frac{TP}{TP+FN}$$

— **Score F1** est une moyenne harmonique de la précision et du rappel. Il équivaut au double du produit de ces deux paramètres sur leur somme. Sa valeur est maximale lorsque le rappel et la précision sont équivalents.

$$\text{F1} = 2 \text{ x } \frac{precisionxrecall}{precision+recall}$$

5 Test

5.1 Validation croisée

Dans un premier lieu nous allons répartir notre jeu de données étiquetées en 70% (données d'entrainement) et 30% (données de validation), nous appliquerons une validation croisée a 3 ou 5 folds sur notre modele d'apprentissage par ensemble afin d'éviter le sur-apprentissage et avoir une bonne estimation du modele sur de nouvelles données.

5.2 Recherche des Hyperparamètres

Les hyperparamètres sont les coefficients internes ou les poids spécifiés lors de la configuration du modèle. Il est difficile de savoir quelles valeurs utiliser pour les hyperparamètres d'un algorithme donné sur un ensemble de donnée, il est donc courant d'utiliser des stratégies de recherche aléatoire ou de grille pour différentes valeurs d'hyperparamètres. Plus on a d'hyperparamètres d'un algorithme, plus le processus de réglage est lent. Par conséquent, on va sélectionner un sous-ensemble minimum d'hyperparamètres de modèle à rechercher. Par exemple, Le paramètre le plus important pour Bagged Decision Trees est le nombre d'arbres ($n\_estimators$). Idéalement, cela devrait être augmenté jusqu'à ce qu'aucune amélioration supplémentaire ne soit observée dans le modèle. Les bonnes valeurs peuvent être une échelle logarithmique de 10 à 1 000.