Université de Bordeaux

Comparatif des méthodes algorithmiques d'intelligence artificielle pour la prédiction des élections

Rapport présenté par : Clément DELMAS et Gabriel MARIE--BRISSON

Date de soumission : 22 Avril 2024

Résumé

Ce rapport présente une étude comparative des méthodes algorithmiques d'intelligence artificielle utilisées pour prédire les élections présidentielles. Nous avons commencé par sélectionner des ensembles de données pertinents, puis nous les avons nettoyés et visualisés. Ensuite, nous avons appliqué différentes méthodes d'intelligence artificielle pour prédire les résultats des élections. Chaque méthode a été évaluée en termes de précision, de rappel, de score F1, ainsi que de corrélation et de coefficient de détermination (R-value) entre les valeurs réelles et prédites. Cette analyse comparative a permis d'évaluer la pertinence des algorithmes d'intelligence artificielle dans le contexte électoral.

Table des matières

I. Introduction

1. Contexte

Dans une société où les élections jouent un rôle primordial dans la vie des citoyens, la capacité à prédire les résultats des élections revêt une grande importance. Cette capacité permet une meilleure compréhension des tendances politiques, ouvrant ainsi la possibilité aux candidats ou aux médias de concevoir des stratégies de communication ciblées. Avec l'abondance de bases de données publiques disponibles, l'utilisation des algorithmes d'intelligence artificielle, qui requièrent un grand nombre de données, offre une perspective pour améliorer la fiabilité des prédictions électorales. Dans ce contexte, le rapport vise à mettre en évidence les avantages et les limites de chaque méthode, offrant ainsi un aperçu de l'efficacité de ces technologies.

2. Objectifs du rapport

Nous avons découpé les prédictions en deux catégories : la classification et la prédiction de valeurs. La classification vise à déterminer quel candidat remportera une élection. Ainsi, les sorties de notre algorithme seront binaires, impliquant une classification binaire. Parmi les nombreuses méthodes existantes, nous en avons sélectionné six : la Régression, les SVM, la Forêt Aléatoire, l'ACP, le KNN et les Réseaux de Neurones. Ce sont les méthodes que nous avons pu utiliser au sein de notre année de Master 1. Chacune de ces méthodes possède diverses variantes offrant des résultats différents, telles que les SVM et la Régression. Afin d'obtenir un comparatif fiable, nous avons sélectionné plusieurs variantes. La régression logistique est largement utilisée dans la modélisation des variables binaires, mais elle n'est pas adaptée aux relations non linéaires. La régression Ridge évite le surapprentissage lorsque les données sont fortement corrélées. La régression stochastique est davantage utilisée dans les grands ensembles de données. Pour les SVM, nous avons choisi les modèles linéaires et polynomiaux, ainsi que GridSearchCV, qui permet de trouver la meilleure combinaison maximisant les performances du modèle. En ce qui concerne la prédiction, nous nous concentrons sur la prédiction du score des candidats ainsi que sur le taux d'abstention. Pour ce faire, nous avons utilisé les mêmes algorithmes : les mêmes régressions, les SVM, la forêt aléatoire et les réseaux de neurones.

II. Méthodologie

1. Description des données

Notre premier jeu de données provenait du site gouvernemental, cependant, il contenait peu d'informations pertinentes pour notre analyse. Nous l'avons donc utilisé comme référence, en nous concentrant uniquement sur les données nécessaires à notre prédiction. Nous avons choisi de prédire les résultats du second tour des élections de 2022, en utilisant deux ensembles de données différents. Le premier, basé sur les données communales, est plus volumineux mais moins fiable en raison de la datation de certaines informations. Le second, basé sur les données départementales, est plus restreint mais plus récent.
Pour le jeu de données communal, nous avons inclus des variables telles que les tranches d'âge, les demandeurs d'emploi, le niveau d'éducation, l'immigration, les revenus, le nombre de crimes, ainsi que les résultats des élections du second tour de 2017 et du premier tour des elections de 2022. Pour le jeu de données départementales, nous avons inclus des variables telles que l'âge moyen, l'écart de revenus, le nombre de magasins bio, le niveau d'éducation, l'immigration et le taux de chômage. Nous avons obtenu ces données auprès de l'INSEE, du gouvernement et de Pôle Emploi.

2. Prétraitement des données

Pour créer notre ensemble de données, nous avons sélectionné les colonnes pertinentes en supprimant les données manquantes. Lorsque cela était possible, nous avons combiné les ensembles de données en utilisant les codes des communes, sinon nous avons utilisé les noms pour les relier. Ensuite, nous avons normalisé chaque variable pour obtenir une distribution gaussienne. Pour le taux de chômage des communes, compte tenu du faible nombre de données (2500 pour les communes de plus de 5000 habitants), nous avons choisi

d'utiliser une distribution aléatoire dans l'intervalle des données afin de mieux répartir les valeurs. Une amélioration possible consisterait à utiliser la moyenne des communes voisines ou des départements pour remplacer les valeurs manquantes. Grâce à ces méthodes, nous sommes passés de 34 955 communes à 31 356 valeurs, soit une perte de 10%, et de 107 départements à 93, avec trois départements métropolitains manquants. Pour sélectionner les colonnes pertinentes, nous avons donné l'ensemble des colonnes de notre base de données ainsi que le résultat attendu, puis nous avons conservé les P-values inférieures à 0.05. Pour optimiser les résultats, nous aurions pu ajuster la sélection de la P-value pour chaque méthode, mais cela n'ayant pas un impact significatif, nous ne l'avons pas fait.

Données des communes

```
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                31357
Model:                          Logit   Df Residuals:                    31309
Method:                           MLE   Df Model:                           47
Date:                Mon, 22 Apr 2024   Pseudo R-squ.:                 0.7136
Time:                        13:51:35   Log-Likelihood:                -6212.3
converged:                       True   LL-Null:                       -21692.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.2909      0.024    -11.883      0.000      -0.339      -0.243
x1            -0.0022      0.033     -0.067      0.947      -0.067       0.063
x2            -0.0202      0.024     -0.853      0.393      -0.067       0.026
x3             0.0062      0.027      0.232      0.817      -0.046       0.058
x4             2.1478      0.067     31.952      0.000       2.016       2.280
x5            -0.1795      0.042     -4.297      0.000      -0.261      -0.098
x6            -2.5792      0.091    -28.290      0.000      -2.758      -2.400
x7            -0.5942      0.036    -16.429      0.000      -0.665      -0.523
x8             0.2567      0.062      4.167      0.000       0.136       0.377
x9             0.1019      0.026      3.890      0.000       0.051       0.153
x10            0.4714      0.034     13.834      0.000       0.405       0.538
x11            0.2074      0.033      6.344      0.000       0.143       0.272
x12            0.0075      0.024      0.313      0.754      -0.040       0.055
x13           -0.2327      0.027     -8.720      0.000      -0.285      -0.180
x14           -0.0529      0.032     -1.642      0.101      -0.116       0.010
x15            0.5832      0.189      3.090      0.002       0.213       0.953
x16           -1.0523      0.190     -5.547      0.000      -1.424      -0.681
x17           -0.0954      0.059     -1.626      0.104      -0.210       0.020
x18           -0.0109      0.029     -0.378      0.706      -0.067       0.046
x19           -0.0827      0.062     -1.336      0.181      -0.204       0.039
x20           -0.0985      0.038     -2.566      0.010      -0.174      -0.023
x21            0.0012      0.039      0.030      0.976      -0.075       0.078
x22            0.0349      0.037      0.931      0.352      -0.039       0.108
x23            0.0557      0.059      0.947      0.343      -0.060       0.171
x24           -0.0040      0.078     -0.052      0.959      -0.156       0.148
x25           -0.0927      0.033     -2.775      0.006      -0.158      -0.027
x26           -0.0420      0.049     -0.851      0.395      -0.139       0.055
x27           -0.1040      0.040     -2.579      0.010      -0.183      -0.025
x28            0.0519      0.043      1.220      0.223      -0.032       0.135
x29            0.0352      0.046      0.766      0.444      -0.055       0.125
x30            0.1286      0.052      2.466      0.014       0.026       0.231
x31            0.0276      0.029      0.956      0.339      -0.029       0.084
x32            0.0422   2.45e+05   1.72e-07      1.000   -4.81e+05    4.81e+05
x33            0.0053   5.07e+05   1.05e-08      1.000   -9.93e+05    9.93e+05
x34           -0.0795   4.16e+05  -1.91e-07      1.000   -8.15e+05    8.15e+05
x35           -0.0298    3.4e+05  -8.75e-08      1.000   -6.67e+05    6.67e+05
x36            0.0394   4.85e+05   8.13e-08      1.000   -9.5e+05     9.5e+05
x37            0.0085    4.5e+05   1.88e-08      1.000   -8.82e+05    8.82e+05
x38            0.0654   2.81e+05   2.33e-07      1.000   -5.51e+05    5.51e+05
x39           -0.0427   4.86e+05  -8.79e-08      1.000   -9.53e+05    9.53e+05
x40            0.0140    4.5e+05   3.11e-08      1.000   -8.82e+05    8.82e+05
x41            0.0363    3.6e+05   1.01e-07      1.000   -7.05e+05    7.05e+05
x42           -0.0168   4.88e+05  -3.45e-08      1.000   -9.57e+05    9.57e+05
x43           -0.0058   2.65e+05   -2.2e-08      1.000    -5.2e+05     5.2e+05
x44            0.4733      0.072      6.529      0.000       0.331       0.615
x45            0.0063      0.020      0.324      0.746      -0.032       0.045
x46            0.0396      0.026      1.523      0.128      -0.011       0.091
x47            0.2750      0.039      6.972      0.000       0.198       0.352
==============================================================================
```

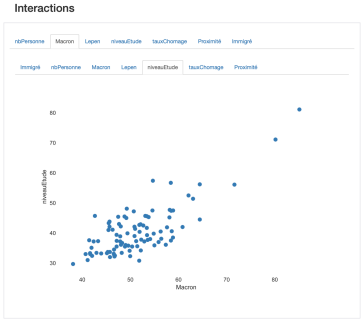
Colonnes avant la sélection via la P-value.

```
Affichage des colonnes selectionné : Index(['MACRON', 'LASSALLE', 'LEPEN', 'ZEMMOUR',
'MÉLENCHON', 'HIDALGO',
       'JADOT', 'PÉCRESSE', 'DUPONT-AIGNAN', 'Macron17', 'Lepen17',
       'P20_HNSCOL15P_BAC', 'P20_FNSCOL15P_BEPC', 'P20_FNSCOL15P_BAC',
       'P20_FNSCOL15P_SUP5', 'PopTotal', 'RevenueCommune'],
      dtype='object')
Optimization terminated successfully.
         Current function value: 0.199229
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                31357
Model:                          Logit   Df Residuals:                    31339
Method:                           MLE   Df Model:                           17
Date:                Mon, 22 Apr 2024   Pseudo R-squ.:                 0.7120
Time:                        13:51:36   Log-Likelihood:                -6247.2
converged:                       True   LL-Null:                       -21692.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.2979      0.024    -12.342      0.000      -0.345      -0.251
x1             2.1472      0.057     37.410      0.000       2.035       2.260
x2            -0.1826      0.036     -5.025      0.000      -0.254      -0.111
x3            -2.6127      0.081    -32.252      0.000      -2.772      -2.454
x4            -0.5631      0.032    -17.646      0.000      -0.626      -0.501
x5             0.2776      0.053      5.202      0.000       0.173       0.382
x6             0.1064      0.025      4.199      0.000       0.057       0.156
x7             0.4887      0.032     15.453      0.000       0.427       0.551
x8             0.2032      0.029      6.989      0.000       0.146       0.260
x9            -0.2476      0.025     -9.768      0.000      -0.297      -0.198
x10            0.5871      0.194      3.028      0.002       0.207       0.967
x11           -1.0530      0.195     -5.409      0.000      -1.435      -0.671
x12           -0.0572      0.024     -2.392      0.017      -0.104      -0.010
x13           -0.0864      0.024     -3.603      0.000      -0.133      -0.039
x14           -0.0845      0.024     -3.518      0.000      -0.132      -0.037
x15            0.2395      0.035      6.872      0.000       0.171       0.308
x16            0.5441      0.064      8.527      0.000       0.419       0.669
x17            0.3359      0.034      9.962      0.000       0.270       0.402
==============================================================================
```

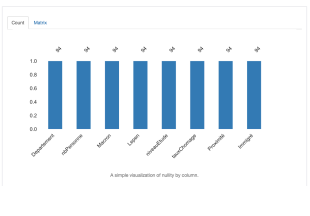
Colonnes après la sélection via la P-value, avec les noms des colonnes sélectionnées en tête de l'image.

3. Méthodes d'analyse

Pour l'analyse de nos données, nous avons utilisé la bibliothèque pandas-profiling. Cette bibliothèque nous a permis de générer une page HTML comprenant une matrice de corrélation, des interactions entre les données et une analyse détaillée des variables. Pour consulter les détails, veuillez vous rendre dans le dossier "Commune" ou "Département", puis lancez le fichier « rapport_correlations.html".

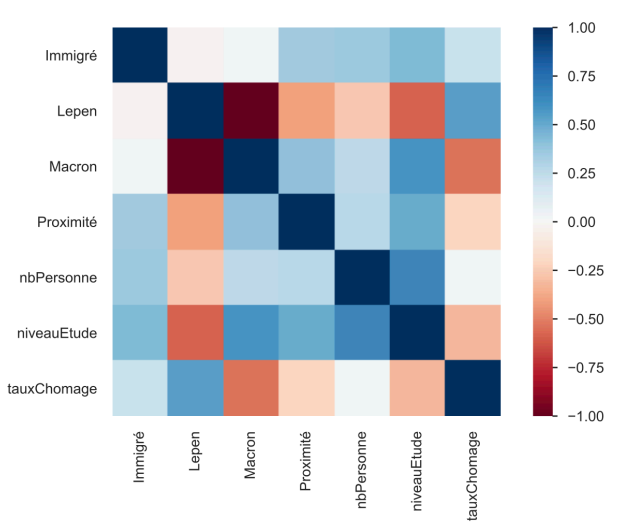


Interactions entre les données études / Votes pour Macron

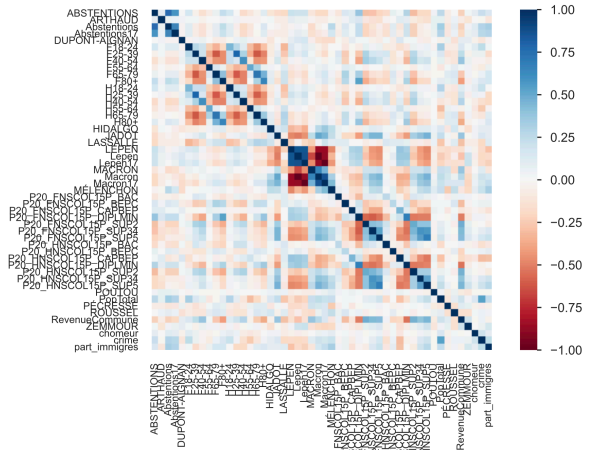


Nombre de données manquantes

Analyse détaillée des variables : hommes ayant un CAP ou un BEP

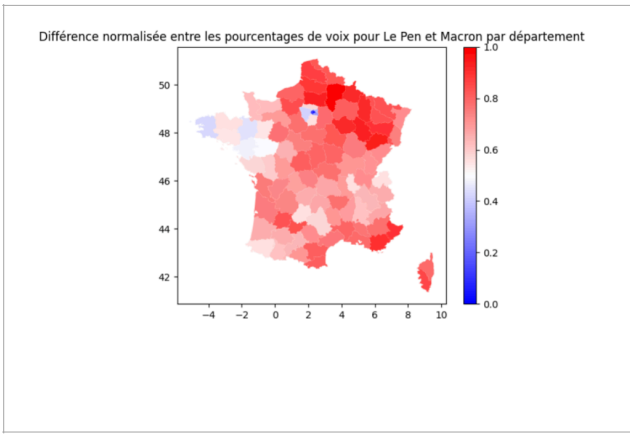


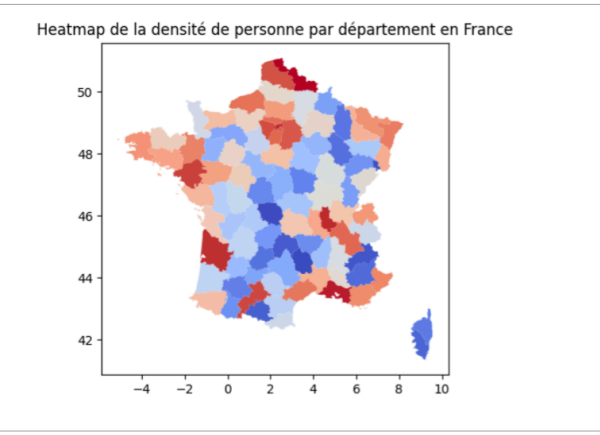
Matrice de corrélation pour les départements



Matrice de corrélation pour les communes

De plus, nous avons affiché les données sur une carte de France afin de les rendre plus facilement compréhensibles pour évaluer leur pertinence. Cette méthode aurait pu nous être utile si nous devions présenter l'efficacité de nos prédictions au grand public en comparant les cartes prédites avec les cartes réels. Vous pouvez retrouver nos 14 cartes dans le fichier « Visualisation.ipynb".

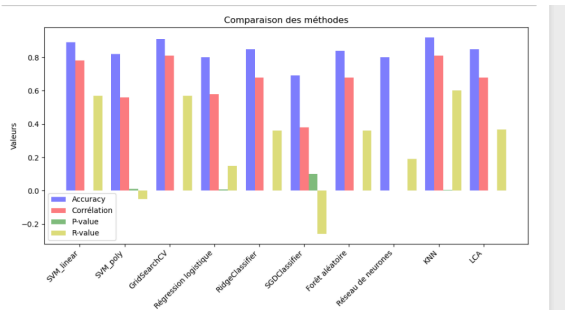




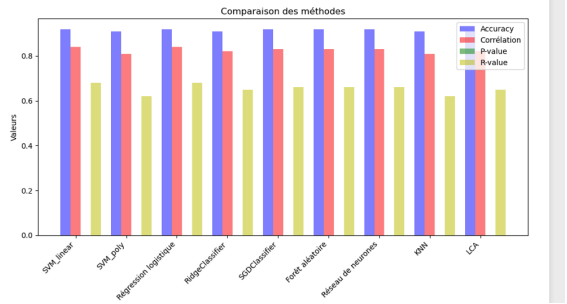
III. Résultats

Par un soucis de lisibilité, nous n'avons pas afficher les résultats du premier tour car il y a trop de candidat, 12. Nous pouvons constater qu'a cause du trop faible score des résultats de certains candidas, nous nos algorithmes n'arrive pas a prédire les classes. Cependant la regression linéaire arrive a prédire les scores.

Classification des résultats

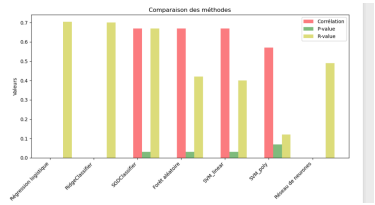


Données des départements pour le 2 eme tours

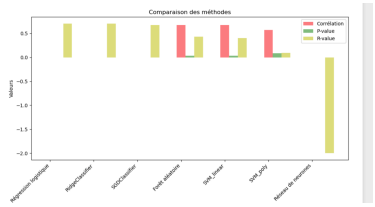


Données des communes pour le 2 eme tours

Prediction des Valeurs des candidats du 2 eme tour

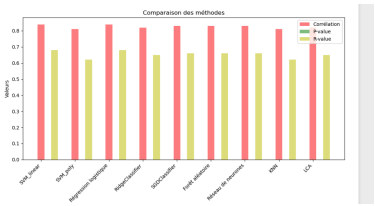


Macron par département

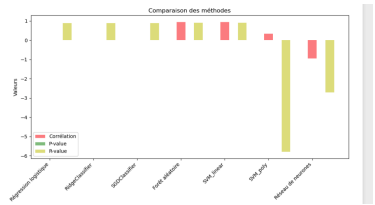


Lepen par département

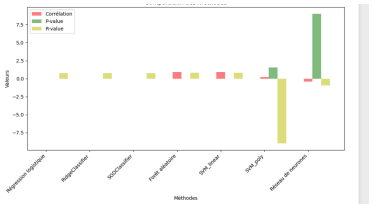
Abstention



Macron par Commune



Lepen par Commune



Abstention par Commune

IV. Discussion

1. Avantages et inconvénients

Voici les différents modèles que nous avons utilisés :

- La régression linéaire est rapide à calculer et facile à interpréter, mais elle est facilement sujette au surajustement.
- La régression Ridge réduit le surajustement mais peut rendre les coefficients moins faciles à interpréter pour le modèle.
- La régression SGD est efficace sur de grands ensembles de données, mais elle est sensible au choix des hyperparamètres et nécessite un prétraitement minutieux des données.
- Les forêts aléatoires sont parmi les meilleurs modèles connus, robustes aux données non linéaires et aux valeurs aberrantes, mais elles peuvent prendre du temps à exécuter si le nombre de paramètres est élevé.
- SVR convient généralement aux ensembles de données de taille moyenne et gère les données non linéaires, mais peut aussi prendre du temps si le volume de données est important.
- Les réseaux de neurones offrent de bonnes performances, mais ils nécessitent souvent beaucoup de données pour obtenir des résultats précis.

2. Pertinence dans le contexte électoral

Pour les classifications avec un faible volume de données, le KNN et le SVM se démarquent largement. Ils sont plus adaptés pour gérer des ensembles de données de petite taille. Le KNN se base sur la proximité entre les points, ce qui le rend efficace. Le SVM cherche à séparer les classes, ainsi avec un petit nombre de données, il trouve un hyperplan optimal.

En revanche, lorsque le volume de données augmente, toutes les méthodes se valent généralement. Cela est dû à la capacité des modèles à généraliser, réduisant ainsi les écarts de performance entre les différentes approches.

Pour les prédictions de valeur avec un faible volume de données, la régression SGD se démarque en raison de sa capacité à s'adapter rapidement aux mises à jour des données, via les itérations. Elle converge vers une solution même avec un faible nombre de données. Cependant, avec un grand volume de données, les réseaux de neurones et le SVM polynomial peuvent afficher des valeurs incohérentes. Cette incohérence peut être due à une complexité excessive des modèles par rapport à la taille des données, ce qui peut entraîner un surajustement et une perte de capacité de généralisation.

V. Conclusion

Nous constatons que ces méthodes sont vraiment pertinentes en cas d'élections, elles nous offrent des résultats fiables. Plus nous avons de données, moins nous sommes dépendants du modèle choisi, cependant, la complexité de certains modèles nécessite plus d’attention afin d'éviter le surapprentissage. Avec cette analyse, nous pouvons imaginer la création d’un algorithme qui, lorsqu'on entre les données demandées, nous indique pour qui nous allons voter.

Références

nombre d’immigration
https://www.insee.fr/fr/statistiques/2012727

ecart type entre les revenue
https://www.insee.fr/fr/statistiques/5371235?sommaire=5371304

magasin bio
https://www.insee.fr/fr/statistiques/4240612#graphique-figure2

taux de chomage
https://www.insee.fr/fr/statistiques/2012804
moyenne d'age Departement
https://www.insee.fr/fr/statistiques/2012692#tableau-TCRD_021_tab1_departements

étudiant par département
https://www.insee.fr/fr/statistiques/5020064?sommaire=5040030#graphique-figure4

ecart type des revenue
https://www.insee.fr/fr/statistiques/5371235?sommaire=5371304

Revenus des Français à la commune
https://www.data.gouv.fr/fr/datasets/revenus-des-francais-a-la-commune/