

Décembre 2020

**RAPPORT DU PROJET**

**« Predict the party : USA »**

**8INF846**

**Intelligence artificielle**

**Automne 2020**

**Travail encadré par :**

* Mr. Mehdi ADDA

**Travail préparé par :**

* Ayman CHAFNI
* Mounia ACHOUCH
* Azize ZOUNGRANA

Table des matières :

[Introduction : 2](#_Toc58255291)

[Présentation du projet : 2](#_Toc58255292)

[Exposition des données : 3](#_Toc58255293)

[Code : 5](#_Toc58255294)

[Résultats : 8](#_Toc58255295)

[L’arbre de décision : 9](#_Toc58255296)

[Conclusion : 10](#_Toc58255297)

[Références : 10](#_Toc58255298)

# Introduction :

Selon le Larousse, l’intelligence Artificielle se définit comme étant « l’ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l’intelligence. ». Elle essaie de reproduire le processus cognitif qui caractérise l’humain afin de réaliser des actions intelligentes. Elle s’applique aujourd’hui dans tous les domaines et sous différentes formes en entrainant une révolution industrielle.

# Présentation du projet :

Afin de mesurer la vitalité de la démocratie d’un pays, les élections restent un critère indispensable. Elles sont presque le seul moyen de faire participer les humains à désigner leurs représentants. Ce choix, au début, apparait libre et transparent, mais en réalité, il est affecté par une multitude de facteurs qui ont poussé la victoire d’un parti par rapport à l’autre.

Ce projet ‘’ Predict the party : USA‘’ vise à prédire les résultats du parti gagnant aux élections des états unis en analysant les statistiques effectuées les années précédentes. En effet, USA n’a que deux grands partis politiques qui remportent une présidentielle américaine : républicain et démocrate. Ainsi, L’étude des comportements électoraux devient aujourd’hui une nécessité en cherchant les causes et les raisons qui sous-tendent les résultats grâce aux facteurs socioculturelles, économiques, politiques…

* Objectif du projet :

Prédire la partie politique qui gagne aux élections américaine 2020 en utilisant les paramètres et les statistiques collectées des années précédentes.

* Méthode utilisée :

Pour réaliser ce projet, nous avons fait appel à la méthode des arbres de décision qui est une méthode très répandue de l’apprentissage supervisé, en utilisant la classe DecisonTreeClassifier de la bibliothèque scikit-learn du langage Python. Cette classe permet, comme son nom l’indique, de classifier les données d’entrée selon des classes bien déterminés. Dans notre cas : « Democ » pour désigner le parti démocrate et « Rep » pour désigner le parti républicain.

Avant de classifier les données avec l’arbre de décision, on a effectué une technique factorielle très populaire, Elle s’agit de L’analyse en composantes principales « l’ACP ». Elle est utilisée, entres autres, pour synthétiser les informations contenues dans un ensemble de données composé de variables actives exclusivement quantitatives. Elle est largement décrite dans la très abondante littérature en langue française consacrée à l’analyse de données. L’ACP cherche à résumer de manière la plus efficace possible l’information disponible en s’intéressant à la variabilité totale portée par chaque variable de la base. Il s’agit donc d’une technique de compression, intéressante surtout lorsque l’on cherche à exploiter les facteurs dans des études subséquentes. Cette analyse est très intéressante parce qu’elle permet ainsi d’améliorer le temps de calcul aussi qu’éliminer le bruit dans les données.

# Exposition des données :

On a collecté les statistiques de chaque année depuis 1960, mais puisque le label qui est le parti vainqueur n’est disponible que chaque quatre années, on a fait la moyenne de chaque quatre années précédents les élections (sauf pour l’année 1960) pour synthétiser le travail du représentant du parti gagnant durant ces quatre années.

On n’a pas utilisé les données de 2020 puisqu’elle n’est pas encore finie.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Années** | **Populations/Million** | **Population urbaine en Million** | **Natalité brut/1000 habitants** | **Revenu personnel médian réel USD** |
| 1960 | 180,67 | 126462473 | 22,696 | 5,6 |
| 1961-1964 | 187,84 | 133274074,5 | 21,26375 | 5247,01 |
| 1965-1968 | 197,57 | 143105927,3 | 18,646 | 7126,25 |
| 1969-1972 | 206,3225 | 151737180,3 | 16,6315 | 8,42825 |
| 1973-1976 | 214,9425 | 158301292,8 | 15,3685 | 10681,5 |
| 1977-1978 | 223,775 | 164915115 | 14,886 | 14376 |
| 1981-1984 | 232,685 | 172465615 | 15,26175 | 19338,25 |
| 1985-1988 | 241,21 | 180232908,8 | 15,5575 | 24108,5 |
| 1989-1992 | 251,4825 | 190002161 | 15,5525 | 28930,5 |
| 1993-1996 | 264,68 | 203987519,5 | 15,10025 | 31894,75 |
| 1997-2000 | 277,4125 | 217922794,3 | 14,303 | 37866,75 |
| 2001-2004 | 286,3 | 229650477,5 | 14,0595 | 42416,5 |
| 2005-2008 | 298,375 | 240400553,8 | 13,87825 | 47029,5 |
| 2009-2012 | 310 | 250971178 | 13,178 | 49669,5 |
| 2013-2016 | 319 | 260623275 | 12,37125 | 53374,75 |
| 2017-2019 | 327,47 | 269514933 | 12,0073 | 62772,25 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Années** | **PIB (%)** | **Revenu personnel médian réel USD** | **Taux(%) de Criminalité/100.000Hbt** | **Taux d'emploi** | **Taux de Chômage(%)** |
| 1960 | 1,98 | 5,6 | 5,1 | 0,38 | 0,066 |
| 1961-1964 | 4,65 | 5247,01 | 4,7 | 0,5885 | 0,055 |
| 1965-1968 | 5,05 | 7126,25 | 5,6 | 0,59075 | 0,0375 |
| 1969-1972 | 2,85 | 8,42825 | 8,2 | 0,558 | 0,052 |
| 1973-1976 | 2,5725 | 10681,5 | 9,375 | 0,61 | 0,07025 |
| 1977-1978 | 3,2675 | 14376 | 9,6 | 0,63 | 0,064 |
| 1981-1984 | 3,14 | 19338,25 | 8,775 | 0,6385 | 0,08725 |
| 1985-1988 | 3,8175 | 24108,5 | 8,325 | 0,652 | 0,0615 |
| 1989-1992 | 2,2425 | 28930,5 | 9,4 | 0,6413 | 0,066 |
| 1993-1996 | 3,3075 | 31894,75 | 8,425 | 0,63175 | 0,0575 |
| 1997-2000 | 4,4525 | 37866,75 | 6 | 0,62825 | 0,0425 |
| 2001-2004 | 2,35 | 42416,5 | 5,875 | 0,587575 | 0,057 |
| 2005-2008 | 2,365 | 47029,5 | 5,65 | 0,56045 | 0,054 |
| 2009-2012 | 1 | 49669,5 | 4,8 | 0,5152 | 0,089 |
| 2013-2016 | 2 | 53374,75 | 4,85 | 0,51305 | 0,055 |
| 2017-2019 | 3,2375 | 62772,25 | 5,1 | 0,54318 | 0,046 |

|  |  |  |
| --- | --- | --- |
| **Années** | **Export - biens/services (Billion of US $)** | **Import - biens/services (Billion of US $)** |
| 1960 | 25940 | 22432 |
| 1961-1964 | 29271,5 | 24822,25 |
| 1965-1968 | 40271,75 | 37657,5 |
| 1969-1972 | 12353,3775 | 12330,33 |
| 1973-1976 | 127,5375 | 123,1275 |
| 1977-1978 | 214,2825 | 235,2975 |
| 1981-1984 | 291,9575 | 338,6725 |
| 1985-1988 | 358,1875 | 483,2 |
| 1989-1992 | 571,035 | 628,0225 |
| 1993-1996 | 764,035 | 849,9825 |
| 1997-2000 | 998,955 | 1222,845 |
| 2001-2004 | 1059,2975 | 1538,18 |
| 2005-2008 | 1568,9375 | 2302,345 |
| 2009-2012 | 1931 | 2445 |
| 2013-2016 | 2283 | 2792 |
| 2017-2019 | 2455,81 | 2358,73 |

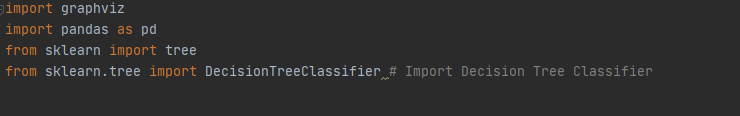
|  |  |  |  |
| --- | --- | --- | --- |
| **Années** | **Paricipation aux elections %** | **President** | **Vainqueur** |
| 1960 | 0,628 | Dwight D. Eisenhower | Rep |
| 1961-1964 | 0,614 | John F Kenn | Democ |
| 1965-1968 | 0,607 | Lyndon Johnson | Democ |
| 1969-1972 | 0,551 | Richard N. | Rep |
| 1973-1976 | 0,536 | Richard N. | Rep |
| 1977-1978 | 0,528 | Jimmy C. | Democ |
| 1981-1984 | 0,533 | Ronald R. | Rep |
| 1985-1988 | 0,503 | Ronald R. | Rep |
| 1989-1992 | 0,552 | George H. W Bush | Rep |
| 1993-1996 | 0,49 | Bill Clinton | Democ |
| 1997-2000 | 0,503 | Bill Clinton | Democ |
| 2001-2004 | 0,557 | George W Bush | Rep |
| 2005-2008 | 0,571 | George W Bush | Rep |
| 2009-2012 | 0,549 | Barack O. | Democ |
| 2013-2016 | 0,592 | Barack O. | Democ |
| 2017-2019 | 0,656 | D. Trump | Rep |

# Code :

Capture du code :

Les bibliothèques que nous avons utilisées sont :

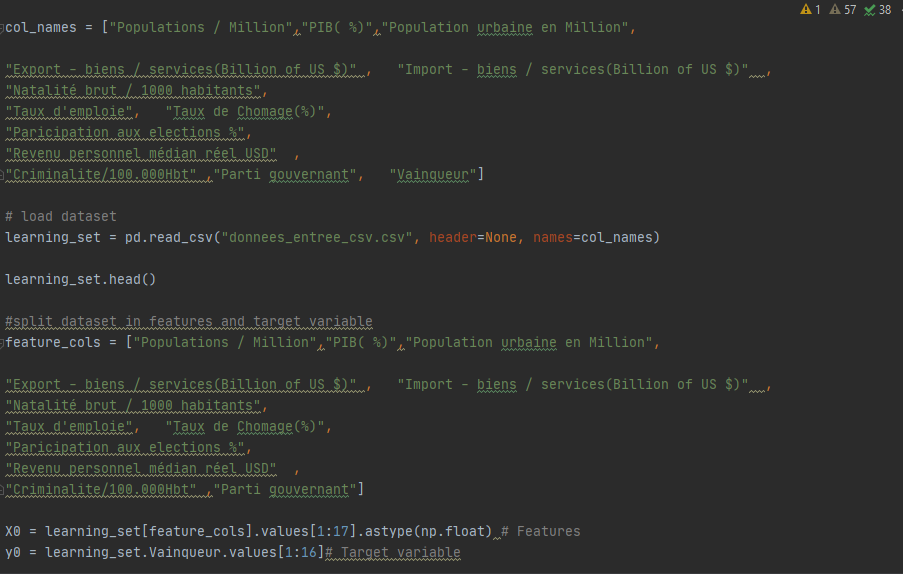
* Scikit-learn : pour les classes de classification avec les arbres de décision notamment DecisionTreeClassifier,
* Pandas : pour la lecture des données depuis un fichier csv
* GraphViz : pour la visualisation de l’arbre de décision



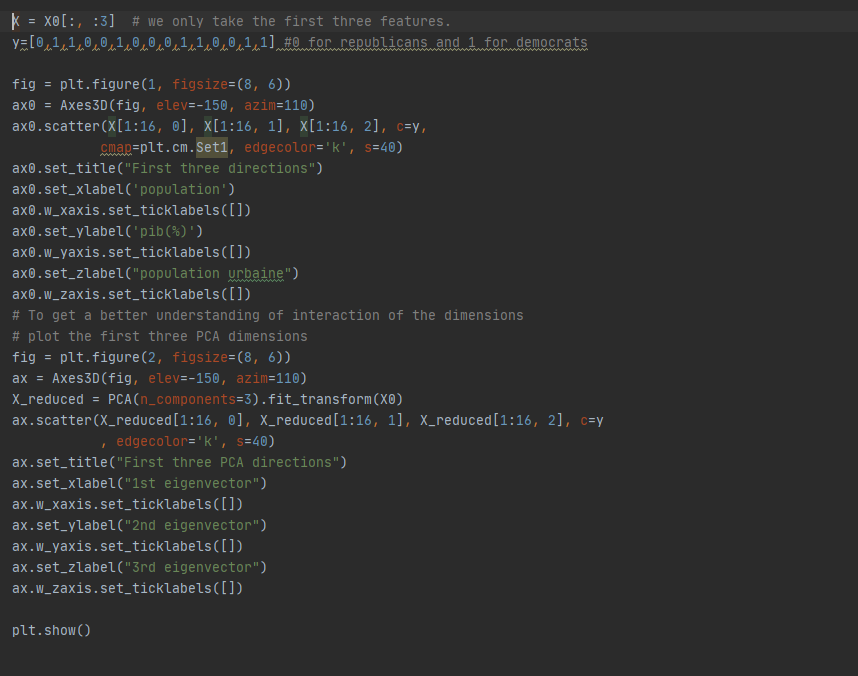
* Matplotib : pour l’affichage des graphiques.
* Numpy : pour effectuer des calculs numériques. Elle introduit une gestion facilitée des tableaux de nombres.



Après la lecture des données d’entraînement depuis le fichier « donnees\_entree\_csv.csv », il fallait transformer les listes du type fourni par Pandas à des listes Numpy par la méthode ’values’.



On a pris ensuite les trois premières colonnes des données d’entrainement pour les présenter dans un graphique 3D avec la bibliothèque Matplotlib, et on a pris aussi les trois premières composantes principales après l’application de la fonction fit\_transform() de la classe PCA du scikit-learn et on les a représenté aussi dans un autre graphique selon la même procédure. Ceci a pour but de réaliser une comparaison de la distribution des données avant et après l’ACP et comment cette dernière permet de bien mettre l’accent sur les directions qui permettent le plus de variations entre les classes.

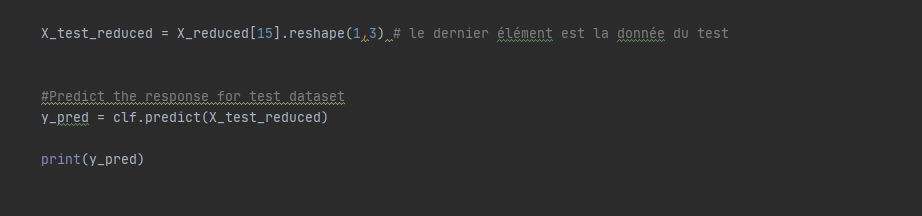


Ensuite, la création d’une instance de la classe DecisionTreeClassifier en appelant le constructeur vide. Enfin, la méthode ‘fit’ de cette classe génère un modèle d’arbre de décision à partir des données d’entrée et leurs labels.

Pour la visualisation des données, on a utilisé la méthode ‘export\_graphviz’ de la classe tree pour transformer la forme obtenue du modèle en une forme adéquate (dot\_data) à la classe GraphViz. Cette dernière qui utilise cette variable du type adapté pour créer un graphe et le mettre dans un fichier ‘png’ qu’on a appelé « décision\_tree.png ».

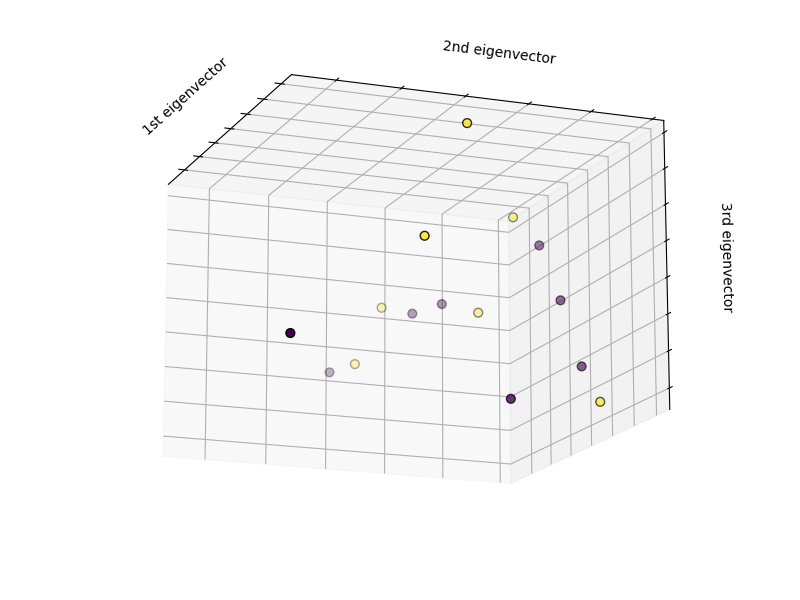
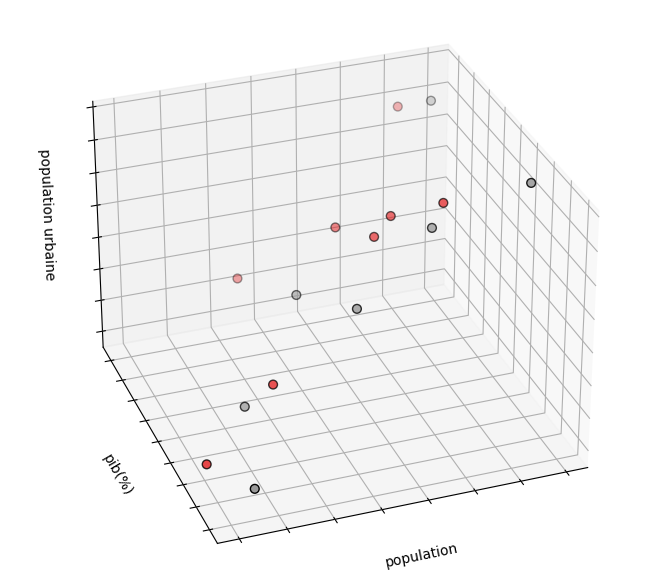


A la fin, On fait appel à la méthode ‘Predict’ de la classe DecisionTreeClassifier qui va prédire la classe des données que nous avons en entrée.

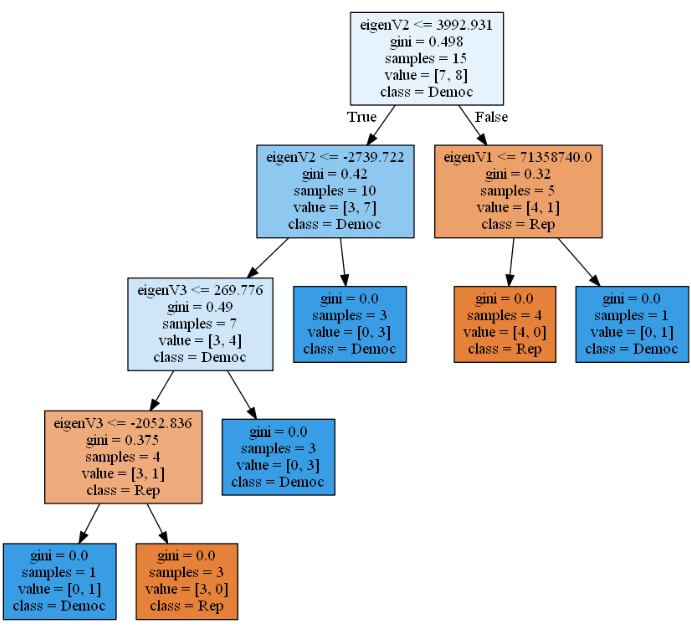


# Résultats :

Le plus grand problème pour nous est qu’on n’a pas assez de données pour montrer parfaitement l’effet de l’ACP sur l’apprentissage, ou encore pour valider la fiabilité de notre modèle étant donné que le nombre de données d’apprentissage est seulement 15 entrées et c’est le nombre de périodes présidentielles depuis 1960 jusqu’à présent.

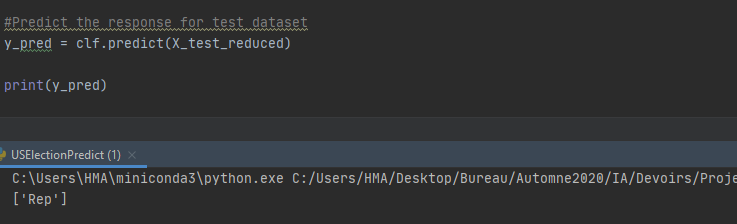


# L’arbre de décision :



Résultat de la prédiction :

* La prédiction affiche que les vainqueurs des élections d’USA pour 2020 sont les républicains.



# Conclusion :

Ce projet ne nous a pas permis de prédire les résultats des élections. La cause est principalement comme déjà mentionné dans le rapport l’insuffisance des données ainsi que le nombre de paramètres et leurs pertinences, sans oublier l’incomplétude des données de test de 2017 à 2020 car l’année 2020 n’est pas encore terminée et les données ne sont pas encore disponibles. Ceci n’empêche que la prédiction des résultats exacts des élections reste toujours possible dans le futur avec le même outil des arbres de décision en cherchant des paramètres plus pertinents et en collectant plus de données.

# Références :

* <https://www.census.gov/>
* [États-Unis - Population urbaine (% de la population totale) | Statistiques (usherbrooke.ca)](https://perspective.usherbrooke.ca/bilan/servlet/BMTendanceStatPays?langue=fr&codePays=USA&codeStat=SP.URB.TOTL.IN.ZS&codeStat2=x)
* <https://scikit-learn.org/stable/modules/tree.html>
* <https://www.datacamp.com/community/tutorials/decision-tree-classification-python>
* <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html?fbclid=IwAR2NJoB2Y0L0Bi2vgB-ZrWdw5bZO4SeE-rvoH5iJiuJqgjR0EU2hxIlpjfg#examples-using-sklearn-decomposition-pca>
* <https://dorukkilitcioglu.com/2018/08/11/pca-decision-tree.html?fbclid=IwAR0uoXZkKUGX9p_OvzG8-akzMUvJvOM45wYAU08Aoa5jl5pWF-c1aWL7HK4#:~:text=This%20makes%20PCA%20a%20natural%20fit%20to%20be,highest%20Information%20Gain%20while%20learning%20a%20Decision%20Tree>