# Les essentiels du Data Scientist – Projet

## Prédire le parti politique victorieux de chaque État U.S. aux élections de 2020

## Description du projet

L'objectif de ce projet est de prédire le parti politique gagnant des élections présidentielles de 2020 aux États-Unis à partir de données socio-démographiques.

Il s'agit donc d'un projet de **classification binaire**.

## Les étapes du projet

Pour ce projet nous vous demandons de suivre les étapes suivantes :

1. *Constituer les données*

Pour ce projet, il vous faudra reconstituer les données. Vous disposerez de plusieurs jeux de données qu'il faudra joindre pour obtenir un DataFrame exploitable.

Le fichier des résultats de 2020 vous servira à créer votre target :

1 pour le Parti Républicain
0 Pour le Parti Démocrate

Vous disposerez également des résultats de 2008 à 2016. **Ce fichier ne vous servira que dans vos analyses exploratoires pour faire des comparaisons avec les élections de 2020 mais ne devra pas être utilisé dans les modèles que vous allez faire**.

En effet l'objectif du projet est de construire un modèle explicatif sur la base des données socio démographique et non sur le résultat des élections passées.

## 2. *L'analyse exploratoire*

Pour ce projet vous devrez mettre en place une analyse exploratoire des données. Vous disposerez d'un grand nombre de variables et vous devrez donc choisir un sous-ensemble de features sur la base de vos connaissances du sujet ainsi que de votre intuition compte tenu de la problématique.

- Vous devrez commencer par **valider la qualité de vos données** : Contrôler la présence ou non de doublons, de valeurs manquantes et de valeurs aberrantes.
- **Déterminez les agrégats et statistiques classiques** pour le sous-ensemble de variables d'intérêts : moyenne, médiane, écart-type, etc.
- Tout au long de votre analyse exploratoire **vous produirez des graphiques** permettant de mieux comprendre les données sur lesquelles vous travaillez.
- Vous ferez une **analyse univariée** pour chaque variable qui vous semblera avoir un intérêt compte tenu de votre problématique : distributions, répartitions, …
- Vous produirez ensuite une **analyse bivariée** de vos données : Analyse des corrélations, box-plots, scatter-plots, etc.

Remarque : Faites attention à la lisibilité de vos graphiques : netteté, titre, noms des axes, taille, légende

## 3. *La modélisation*

Ce projet est une tache de classification binaire. Pour cette partie vous devrez :

- Vous questionner sur le **features engineering** : Création de variables, transformations ou non à appliquer à vos données numériques (ex : log-transformation, normalisation, standardisation, etc.). Cette étape peut être cruciale en fonction du choix de l'algorithme que vous allez retenir (comme vous l'avez vu dans le cours de Machine Learning Avancé).
- **Encoder vos variables catégorielles** en variables numériques. Il vous faudra choisir entre les techniques que vous aurez vu telles que le **label encoding** ou le **one hot encoding**.
- **Procéder à une sélection de variables**.
- **Séparer** vos données en train et en test afin de mesurer la capacité de vos modèles à se généraliser. Attention à la répartition des différents États dans les deux jeux de données.

- **Tester différents modèles de classification** : Vous devrez obligatoirement réaliser une régression logistique comme modèle baseline que vous devrez challenger par des modèles non-linéaires.
- **Vous hyper-paramétriserez vos modèles avec un GridSearchCV**.
- Pour construire vos modèles vous utiliserez des **pipelines**.
- Vous proposerez une **analyse de l'importance des variables globales** (exemples : les coefficients de la régression logistique ou les features importance d'une random forest) **et locale** (shap)
- Dans tout le projet, vous attacherez une grande importance à la **qualité de votre code**.

Conseil : Il est recommandé de commencer par des modèles simples et de complexifier votre approche (features engineering, choix des modèles) au fur et à mesure.

### *4. Évaluation*

Pour évaluer la qualité de vos modèles **vous utiliserez le F1-Score**. Vous pourrez également regarder les métriques telles que le Recall, la Précision ou encore l'Accuracy.

Vous devrez systématiquement afficher les résultats sur votre jeu de train et de test pour justifier la capacité de votre modèle à se généraliser.

## Les rendus

- Vous avez **jusqu'au dimanche 24 Février 2024** pour rendre le projet
- Vous devrez nous remettre le **Notebook commenté contenant vos analyses exploratoires et vos modélisations** ainsi que la version **html (avec le code exécuté et sans erreurs)** de ce Notebook.
- Votre projet doit être au maximum reproductible.
- La semaine suivant le rendu sera consacrée à la préparation de votre soutenance (création de vos slides)

## La soutenance

- Les soutenances (**30min**) auront lieu la **semaine du 04 Mars 2024**
- Pour la soutenance, **15min seront consacrées à la présentation de vos slides**
- Vos slides comprendront :
    - La présentation des données
    - Quelques analyses exploratoires pertinentes
    - Vos choix de features engineering & d'encoding
    - Vos choix de modélisation
    - L'analyse de vos résultats
- Les **15 dernières minutes seront consacrées à la réponse aux questions** que l'évaluateur vous posera.