ECF 2-juin-23

**Première partie :**

Workflow :

Workflow classique d’un projet de machine learning :

1. Prise en main du jeu de données (inspection, mise en place d’hypothèses, etc.)
2. Prétraitement (nettoyage, normalisation, etc.)
3. Choix du modèle d’apprentissage ou de clustering en fonction de la problématique
4. Évaluation des résultats (validation simple ou croisée…)
5. Optimisation d’hyperparamètres du modèle
6. Évaluation finale et interprétation

Outils utilisés :

* Jupyter notebook dans Google Collab
* Python, bibliothèque pandas, numpy, Scikit Learn et matplotlib

Dataset :

Un fichier .csv des votes des 100 sénateurs américains sur 15 lois différentes.

Outil utilisé pour la visualisation :

matplotlib

**Deuxième partie :**

Workflow :

Workflow classique d’un projet de machine learning :

1. Prise en main du jeu de données (inspection, mise en place d’hypothèses, etc.)
2. Prétraitement (nettoyage, normalisation, etc.)
3. Choix du modèle d’apprentissage ou de clustering en fonction de la problématique
4. Évaluation des résultats (validation simple ou croisée…)
5. Optimisation d’hyperparamètres du modèle
6. Évaluation finale et interprétation

Schéma du dataset :

Un fichier .csv :

* le séparateur de colonnes est la virgule
* de 101 lignes : 1 ligne d’en-tête + 100 observations, chaque observation est un sénateur)
* de 18 colonnes :
  + le nom (“name”),
  + le parti (“party”, “R” pour Républicain, “D” pour Démocrate, “I” pour Indépendant), l’état (“state”, le sigle de deux lettres majuscules de l’état),
  + et le vote pour chacun des 15 projets de lois (le nom de chaque colonne est un entier de 5 chiffres correspondant au numéro du projet de loi, les 3 premiers chiffres étant tous des 0 ; pour chaque sénateur un 0 est mis dans la colonne en cas de vote contre, un 0.5 en cas d’abstention et un 1 en cas de vote pour)

Librairies et fonctions utilisées :

* pandas pour l’utilisation des DataFrame
* numpy pour compter les valeurs uniques et randomiser les lignes
* matplotlib pour visualiser les données et les résultats des clusterisations
* sklearn pour le clustering en utilisant l’algorithme Kmeans (on crée un classifieur, puis on l’utilise)

[Lien GitHub](https://github.com/AlexandreSEBAIS/ECF_2023_06_02)