# Clustering

TP noté

L'objectif de ce travail est d'étudier une banque de données sur la consommation des voitures disponible sur le dépôt de l'UCI (`https://archive.ics.uci.edu/ml/datasets/Auto+MPG`). Pour faciliter l'importation des données, un fichier avec une virgule pour séparateur est disponible sur arche (auto_mpg.csv).

Le choix du langage, de la bibliothèque et/ou du logiciel est libre.
Le rapport et les sources sont à déposer dans le dépôt de la page arche de l'EC clustering **au plus tard le 11 décembre 2023**.

## 1 Analyse des données

1. télécharger les données (auto_mpg.data ou auto_mpg.csv) et le descriptif (cmc.names) ;
2. analyser les données ;

## 2 Régression

Sur la base de ce que vous avez vu en TP, comparez les meilleurs résultats obtenus par un arbre de régression de type CART et par une forêt aléatoires d'arbres de régression de type CART.

1. préparer les données pour prédire la variable *mpg* ;
2. apprendre un arbre de régression de type CART ;
3. visualiser l'arbre ;
4. afficher l'importance des variables ;
5. afficher les valeurs des mesures d'erreurs mean square error, absolute mean square error, $R^2$;
6. faire de même avec une forêt aléatoire en faisant varier les paramètres.

### Rendu

Les fichiers ipynb sont acceptés.

- Le rapport doit décrire votre démarche en présentant :
  - vos choix (prétraitement des données, paramètres de algorithme...) ;
  - les analyses des résultats que vous avez menées (mesures d'erreurs, visualisation de l'arbre, importance des variables...) ;

  - une conclusion précisant les points forts et les points faibles de l'étude et les améliorations possibles.

- Les sources doivent être accompagnées d'un fichier README.txt indiquant comment utiliser vos sources.

# 3 Clustering

On souhaite appliquer une méthode de clustering sur les données de consommation des voitures. Afin de choisir la méthode de clustering et ses paramètres, on observera dans un premier temps les données dans un espace à 2D en appliquant une ACP puis le résultat de ce clustering dans l'espace 2D.

1. préparer les données. La variable "car name" servira d'étiquette et ne sera donc pas utilisée comme variable d'entrée ;
2. projeter les données dans un espace 2D avec une Analyse en composantes principales ;
3. choisir et appliquer une méthode de clustering suite aux observation de la projection ;
4. visualiser dans l'espace 2D les clusters ;
5. afficher des mesures d'erreurs ;

## Rendu

- Le rapport doit décrire votre démarche en présentant :
  - vos choix (prétraitement des données, paramètres de algorithme...) ;
  - les analyses des résultats que vous avez menées (mesures d'erreurs, visualisation...) ;
  - une conclusion précisant les points forts et les points faibles de l'étude et les améliorations possibles.
- Les sources doivent être accompagnées d'un fichier README.txt indiquant comment utiliser vos sources.