*Classe : 4Data Science*
*Proposé par : M^ed Heny Selmi*

# Atelier : Analyse prédictive des sinistres automobiles corporels

## Objectif :

L'idée est de réaliser une analyse des données de sinistres automobiles corporels. Les candidats de chaque équipe doivent montrer les compétences suivantes:

- ✓ Autonomie à poser une problématique bien formulée : Formuler une question qui peut être étudiée à l'aide de données à votre disposition
- ✓ Techniques de nettoyage et transformation des données (feature engineering)
- ✓ Capacité à récupérer des données supplémentaires utiles
- ✓ Sens analytique : Analyser les données pour répondre à la question formulée en utilisant des techniques de machine learning
- ✓ Techniques de data visualisation

## Données :

- ✓ Données sinistres corporels auto : https://www.data.gouv.fr/fr/datasets/base-de-donnees-accidents-corporels-de-la-circulation/
  - Vous devez dans un premier temps agréger les données pour créer une variable **par commune**. Cette variable sera la variable cible, à prédire en fonction d'autres variables.
  - Quelques exemples d'indicateurs (tous les nombres doivent être normalisés avec la population des communes correspondantes et la prédiction de chaque indicateur nécessite un modèle)
    - nombre d'accidents sur les voies communales, par commune
    - nombre d'accidents sur les départementaux, et autres types de routes.
    - gravité des accidents (sur les différents types de routes)
    - nombre de piétons blessés
    - nombre d'accidents entre 23 et 6h du matin le weekend.
    - à vous de construire des indicateurs intéressants

- ✓ Données explicatives
  - Données insee : https://www.insee.fr/fr/statistiques?debut=0&categorie=3&geo=TOUTES_COMMUNE-1
  - Données sur Eurostat
  - Quelques variables qui peuvent sembler importantes:
    - revenu
    - population
    - répartition des âges
    - répartition des professions (CSP)

- taux de chômage
- données météo

- La clé de jointure est le numéro de commune (car les données sur les sinistres auto sont agrégées au niveau des communes), et si les données insee sont par code insee, il faudra trouver la correspondance entre les deux.
- Il est possible d'utiliser d'autres données à savoir : Open data : données publiques françaises (https://www.data.gouv.fr/fr/), association de l'open data (http://www.opendatafrance.net/), les concours (https://www.kaggle.com/ ou https ://www.datascience.net/), etc., Open API des réseaux sociaux : voir http://www.programmableweb.com/, Le web : pages web, etc.

## Descriptif :

Pour prédire un indicateur que vous avez choisi avec les autres variables géographiques, il s'agit de construire un modèle de régression.

Si vous avez le temps, vous pouvez aussi créer un modèle de segmentation des communes en fonction de typologies d'accidents et des données insee. Ainsi, il est nécessaire d'automatiser la construction de nombreux indicateurs

## Livrables :

Pour la reproductibilité de l'analyse, il est obligatoire d'utiliser Rmarkdown.

Pour la clarté du projet, il est impératif de créer un fichier zippé (avec des liens relatifs dans le code), le fichier doit être nommé avec : Nom-equipe-LAB4, et doit comprendre les éléments suivants :

- ✓ Votre DataViz sous forme de Dashboard
- ✓ Votre code R.
- ✓ Une présentation portant sur votre vision métier et les bénéfiques de la solution proposée.
- ✓ Un fichier html (sortie de rmarkdown) : N'oubliez pas vos DataViz de toutes les étapes de votre processus de travail.
- ✓ Vous pouvez travailler avec R et/ou Python.