Sujet de Projet - Potabilité de l'eau

Contexte

Outre les problèmes économiques ou politiques, l'accès à l'eau potable est une problématique toujours d'actualité notamment dans les pays en voie de développement.

C'est avant tout un élément indispensable et essentiel à la santé, un droit fondamental de la personne et un élément d'une politique efficace de protection de la santé. Vous faites chaque jour le simple geste d'ouvrir un robinet et de remplir un verre ou une bouteille. Mais ce n'est malheureusement pas pareil dans d'autres pays.

Dans certaines régions, il a été démontré que les investissements dans l'approvisionnement en eau et l'assainissement peuvent générer un avantage économique net puisqu'on évite d'être malade ou d'arriver dans des situations critiques où les soins peuvent coûter très cher.

Objectif

Trouver un modèle permettant de prédire si l'eau d'un cours d'eau est potable ou non afin de savoir si nous devons investir dans le développement d'un réseau d'assainissement de l'eau. Les détails du challenge sont disponibles ici.

Les données

Les différentes données sont accessibles sur Kaggle en cliquant ici. Nous travaillons avec 9 variables explicatives et des données numériques dites continues. Le dataset correspondant est intitulé **drinking_water_potability.**

Les 9 variables explicatives sont :

- Ph de l'eau
- Dureté de l'eau (calcaire dans l'eau)
- Solvabilité des solides dans l'eau (TDS)
- Chloramine
- Sulfate
- Conductivité
- Carbone organique
- Trihalométhane
- Turbidité (si l'eau est trouble ou clair)

La variable à expliquer est la potabilité de l'eau.

Prérequis

Des données sont manquantes pour certaines features et certaines instances. Il faudra décider dans un premier temps comment gérer ce problème en se référant aux différents moyens vus en cours (suppression des instances avec des données manquantes ou remplacer par la moyenne…)

A vous de voir s'il faut faire aux préalables une réduction de dimension i.e. ne prendre en compte que certaines variables explicatives parmi les 9 disponibles.

Tâches à réaliser

1. Tester plusieurs modèles de prédictions. Vous pouvez choisir autant d'approche que nécessaire notamment des modèles non vus en cours.
2. Comparer les résultats obtenus pour les différentes approches (en termes d'accuracy…)
3. Conclure en justifiant le choix du modèle le plus approprié conférant les meilleures prédictions.