Projet – Apprentissage supervisé

Ce TP est à réaliser par groupe de **2 étudiants maximum**.
Date limite de rendu : **vendredi 17 décembre à 23h59** au plus tard.
Règle : **5 points de moins par jour de retard**.

Objectif : mettre en application les modèles vus en cours (linéaires, réseaux de neurones, arbres de décision, …) dans le cadre de deux problèmes de classement et un problème de régression.

Données : vous aurez à manipuler 3 jeux de données différents (sur Teams) :

- « df_breastCancer.csv » est un jeu de données uniquement quantitatives, composé de 32 caractéristiques liées à une étude sur le cancer du sein à l'université du Wisconsin. Les 32 caractéristiques correspondent à des informations calculées sur les images de radiographie de 569 patientes qui sont réparties en deux classes : B pour bénin et M pour malin (variable « Diagnosis »).
- « df_mushrooms.csv » est un jeu de données uniquement qualitatives, composé de 22 caractéristiques récoltées sur 8124 champignons répartis en 2 classes : comestibles / toxiques.
- « df_chauffage.csv » est un jeu de données uniquement quantitatives, composé de 8 caractéristiques techniques issues de 768 bâtiments à partir desquelles il faut prédire la consommation de chauffage (dernière colonne).

Vous devez :

1. Créer un Jupiter Notebook contenant l'ensemble de vos traitements sans oublier les commentaires associés.
2. Pour chaque jeu de données, réaliser une description des données (analyses descriptives, …) et proposer un prétraitement de ces données pour qu'elles puissent être mises en entrée des modèles qui seront construits (en particulier, la transformation des données qualitatives).
3. Construire des modèles permettant de classifier les deux catégories de patientes (bénin / malin), ainsi que les deux types de champignons (comestibles / toxiques).
4. Construire des modèles de régression pour les données de chauffage.
5. Pour chaque modèle, donner au moins les valeurs de l'accuracy et de la F-mesure.
6. Commenter l'ensemble de votre code.

Liens :

- Cancer :
  - https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
  - https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
- Mushrooms :
  - https://archive.ics.uci.edu/ml/datasets/mushroom
  - https://www.kaggle.com/uciml/mushroom-classification
- Chauffage :
  - https://archive.ics.uci.edu/ml/datasets/Energy+efficiency
  - https://www.kaggle.com/elikplim/eergy-efficiency-dataset