IA01 – Introduction à l'intelligence artificielle
Projet : Prédiction d'insuffisance cardiaque

Contexte

Le but de ce projet est de mettre en oeuvre les différentes techniques d'analyse de données et de machine learning pour prédire une insuffisance cardiaque. Pour cela, vous allez travailler sur les données disponibles à l'adresse suivante (dataset 'heart.csv' sur Moodle) :
https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data

Cet ensemble de données combine des ensembles de données de plusieurs sources. Le populaire jeu de données sur les maladies cardiaques de l'hôpital de Cleveland est combiné avec d'autres ensembles de données sur les maladies cardiaques (Hongrie, Suisse...) comme suit :

Cleveland: 303 observations
Hungarian: 294 observations
Switzerland: 123 observations
Long Beach VA: 200 observations
Stalog (Heart) Data Set: 270 observations

Total: 1190 observations
Duplicated: 272 observations
Final dataset: 918 observation

Jeu de Données

Le dataset contient les caractéristiques suivantes :

- Age: age of the patient [years]
- Sex: sex of the patient [M: Male, F: Female]
- ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- RestingBP: resting blood pressure [mm Hg]
- Cholesterol: serum cholesterol [mm/dl]
- FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
- ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
- Oldpeak: oldpeak = ST [Numeric value measured in depression]
- ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- HeartDisease: output class [1: heart disease, 0: Normal]

Objectifs du Projet

1. **Analyse exploratoire des données (EDA)** : Comprendre les caractéristiques du jeu de données, leur distribution et leur relation entre elles ainsi qu'avec la variable cible. L'objectif est de détecter les corrélations pertinentes entre les variables explicatives et

d'identifier les facteurs qui influencent le plus la sortie. Une analyse bien détaillée et pertinente est attendue de vous.

2. **Préparation des données** : Effectuer les transformations nécessaires pour rendre les données utilisables dans les modèles (par exemple, gestion des valeurs manquantes, transformation des variables catégorielles, gestion des valeurs aberrantes).
3. **Modélisation** : Utiliser des algorithmes de classification pour prédire si les sujets étudiés ont une insuffisance cardiaque. Les algorithmes à considérer sont ceux du cours, notamment : kNN, arbre de décision, forêts aléatoires, régression logistique et réseaux de neurones. **Il faudra également choisir un algorithme de classification non vu en cours, l'expliquer brièvement et le tester.**
4. **Évaluation** : Mesurer la performance des modèles en utilisant des métriques adaptées.

Livrables et dates importantes :

- **Rapport :** Un document d'une dizaine de pages est demandé. Ce rapport doit inclure les résultats de votre analyse exploratoire des données, les étapes de préparation des données suivies, les hyperparamètres testés et les valeurs retenues pour chaque algorithme, ainsi que l'évaluation et le choix final. Vous devez également décrire les principales difficultés rencontrées et commenter les résultats obtenus.
- **Travail en groupe :** Vous pouvez travailler en groupes de **deux ou trois étudiants.**
- **Présentation orale :** Quelques slides résumant le rapport doivent être présentées en 7 à 8 minutes maximum, le vendredi 19 décembre pendant la séance de TD. La présentation sera suivie de 5 minutes de questions.
- **Dépôt des documents sur Moodle** : Pour vendredi 19 décembre à 9h, veuillez déposer les documents suivants :
  - le rapport,
  - le notebook ayant servi à générer les résultats,
  - la présentation.