Arbre de décision et algorithme CART

Ce projet se fera en 2 temps sur 2 jeux de données différents. Le CR de ce projet est à rendre sous la forme d'un jupyter notebook.

1/ Prise en main

Nous allons utiliser une fonction de la bibliothèque scikit-learn
https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier

Sur les données de Iris de Fisher créer un arbre de décision avec l'algorithme CART et afficher cet arbre.

```
from sklearn.datasets import load_iris
from sklearn import tree

iris = load_iris()
X, y = iris.data, iris.target
```

Modifier les valeurs des paramètres *max_depth* (profondeur maximale de l'arbre) et *min_samples_leaf* (nombre minimal d'échantillons dans un nœud feuille). Ces paramètres permettent de mettre des contraintes sur la construction de l'arbre et donc de contrôler indirectement le phénomène de sur-apprentissage.

2/ A rendre

Vous devez analyser le jeu de données Heart Disease. La description complète du jeu de données se trouve **:**
https://archive.ics.uci.edu/ml/datasets/Heart+Disease

14 caractéristiques
age: age in years
sex: sex (1 = male; 0 = female)
cp: chest pain type
-- Value 1: typical angina
-- Value 2: atypical angina
-- Value 3: non-anginal pain
-- Value 4: asymptomatic
trestbps: resting blood pressure (in mm Hg on admission to the hospital)
chol: serum cholestoral in mg/dl
fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg: resting electrocardiographic results

-- Value 0: normal
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach: maximum heart rate achieved
exang: exercise induced angina (1 = yes; 0 = no)
oldpeak = ST depression induced by exercise relative to rest
slope: the slope of the peak exercise ST segment
-- Value 1: upsloping
-- Value 2: flat
-- Value 3: downsloping
ca: number of major vessels (0-3) colored by flourosopy
thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

target: diagnosis of heart disease (angiographic disease status)
-- Value 0: < 50% diameter narrowing
-- Value 1: > 50% diameter narrowing
(in any major vessel: attributes 59 through 68 are vessels)

1/ Décrire les données
2/ Proposer un arbre de décision pour prédire la target (séparer la base en phase d'apprentissage et de test et on répètera cette séparation)
3/ A chaque nouveau tirage base apprentissage / base de test visualiser l'arbre
3/ Conclusion