esiea
ECOLE D'INGENIEURS
DU MONDE NUMERIQUE

5A : MACHINE LEARNING 1

PROJET

Livrables

DTM : 1 rapport sur l'étude 1 →1 fichier pdf : nom : projetML1_nom1_nom2(_nom3).pdf
DTE : 1 rapport sur les études 1 et 2 → 1 fichier pdf : nom : projetML1_nom1_nom2(_nom3).pdf

Deadline du dépôt sur moodle : lundi 19/11 23h59

Nombre d'étudiants par équipe projet : minimum 2, maximum 3. Pas de « mix » DTM/DTE.

Base de données

La base de données adultdata.txt (https://archive.ics.uci.edu/ml/datasets/Adult) a été rassemblée pour prédire à partir de différentes caractéristiques (âge, genre, niveau d'étude, pays d'origine …), si le niveau de revenu d'un citoyen américain dépassait ou non un certain seuil (50K $).

Elle comporte plus de 48K observations, décrites par 14 caractéristiques. Certaines d'entre elles sont qualitatives, d'autres, quantitatives.

Pour charger la base :

```
import pandas as pd
data = pd.read_csv("adult.data.txt",
    names=["Age", "Workclass", "fnlwgt", "Education", "Education-Num", "Martial Status", "Occupation", "Relationship", "Race", "Sex", "Capital Gain", "Capital Loss", "Hours per week", "Country", "Target"],
        sep=r'\s*,\s*',
        engine='python',
        na_values="?")
data.tail()
```

Analyse préliminaire et encodage

Analyser les différentes variables et les caractériser : intervalle de variation, nombre de modalités …

Les variables qualitatives peuvent être, suivant les algorithmes d'apprentissage mis en œuvre, soit utilisées directement, soit encodées par « one hot encoding ».

Etude 1 : DTM+DTE

Tester les différents algorithmes (plus-proches-voisins, réseaux multicouches, svm, arbre de décision …) vus pendant le cours sur ces données. Choisir les hyper-paramètres par une procédure de type k-fold validation.

Comparer les performances des différentes méthodes considérées séparément. Justifier les choix effectués de façon claire et rigoureuse.

Combiner (en utilisant un vote majoritaire par exemple) les décisions des différentes méthodes et

évaluer les performances de cette combinaison.

Conclure.

Etude 2 : DTE

Evaluer les algorithmes d'apprentissage d'ensemble de classifieurs vus en cours (forêts aléatoires, bagging, boosting) sur ces données. Comparer les performances des différentes méthodes.

Mesurer, quand c'est possible, l'impact des variables sur la prédiction.

Conclure.