Département IA & SD le 13/10/2025

Section: 4 ING Software

Module: FD1 TP N° 1

La fouille de données ou "data mining" est un processus qui consiste à explorer de grandes quantités de données pour y découvrir des **informations utiles et inconnues au préalable**.

La fouille de données est un outil puissant pour transformer des données brutes en connaissances exploitables.

Ce TP porte sur l'exploration et la manipulation de données à l'aide de Python et de la bibliothèque Scikit-learn (SkLearn), le tout dans l'environnement Jupyter Notebook d'Anaconda. Vous allez apprendre à :

1. **Installer et configurer l'environnement Anaconda (Jupyter) avec le package SkLearn.**
2. **Manipuler et explorer un fichier d'apprentissage :**
    - Ouvrir différents types de fichiers de données (benchmarks).
    - Lire et afficher les données à l'aide de Pandas.
    - Obtenir des informations de base sur les données :
        - Nombre d'instances (lignes).
        - Nombre et noms des attributs (colonnes).
        - Type de chaque attribut.
    - Calculer et afficher les 5 nombres clés pour chaque attribut :
        - Minimum.
        - Maximum.
        - Médiane.
        - Premier quartile (Q1).
        - Troisième quartile (Q3).
    - Visualiser la distribution des données :
        - Tracer les boxplots (boîtes à moustaches) de chaque attribut sur le même graphique.
        - Afficher le Scatter plot (nuage de points) du benchmark pour visualiser les relations entre les attributs.
    - Calculer et afficher les statistiques descriptives :
        - Mode.
        - Moyenne.

        - Médiane.
    - Gérer les valeurs manquantes :
        - Identifier les valeurs manquantes.
        - Remplacer les valeurs manquantes par une méthode appropriée (par exemple, la moyenne).
    - Normaliser les données :
        - Normalisation Min-Max.
        - Normalisation Z-score (standardisation).

Rapport à remettre le 27/10/2025