

TP : Pipeline d'Ingestion de Données avec Scikit-learn et Pandas sur des Données Agronomiques

Objectif :

L'objectif de ce TP est de vous familiariser avec la création d'un pipeline d'ingestion de données en utilisant pandas pour la manipulation des données et scikit-learn pour la préparation des données en vue de l'apprentissage automatique. Nous travaillerons sur un jeu de données agronomiques contenant entre autres les types de cultures et les quantités d'eau utilisées, et la nature de l'eau salée ou non.

Le pipeline d'ingestion de données nettoie, encode et normalise les données agronomiques. Ces données sont prêtes à être utilisées pour entraîner un modèle de machine learning. Vous pouvez maintenant passer à l'étape de sélection et d'entraînement du modèle.

Extension :

- Essayez d'intégrer un modèle de réseau de neurones ou un random forest après le pipeline pour prédire par exemple la quantité d'eau utilisée, le rendement des cultures etc
- Explorez l'utilisation de GridSearchCV pour optimiser les hyperparamètres du modèle en utilisant le pipeline.