Stage de Master 2 en Data Science

Exploration d'approches de machine learning/deep learning pour l'amélioration des prédictions du rendement chez le palmier à huile

Introduction :

La **sélection génomique** est un processus d'amélioration des espèces végétales ou animales basé sur l'analyse des informations génétiques pour prédire et sélectionner les individus présentant des caractéristiques souhaitables. Cette approche présente un atout majeur pour les plantes pérennes (comme le palmier à huile), dont les cycles de croissance sont longs et ralentissent fortement les programmes de sélection classiques. En permettant d'identifier précocement les croisements les plus prometteurs – par exemple en termes de rendement, de tolérance aux maladies ou d'adaptation à l'environnement – la sélection génomique constitue un levier essentiel pour accélérer et rendre plus efficace l'amélioration variétale. Le **CIRAD** (Centre de coopération internationale en recherche agronomique pour le développement) conduit des recherches concernant la sélection génomique pour l'amélioration de différentes plantes tropicales. L'équipe GSP (Génome et Sélection des Pérennes) de l'[UMR AGAP Institut](#) (Unité Mixte de Recherche Amélioration Génétique et Adaptation des Plantes) recherche **un.e stagiaire de Master 2 pour tester et comparer des méthodes d'apprentissage pour la prédiction génomique du Palmier à Huile.**

Objectifs :

Ce stage a pour objectif de tester différentes approches de **machine learning** (ML), en particulier de **deep learning**, et de comparer les résultats avec des approches statistiques classiques basées sur le modèle linéaire mixte pour la prédiction des performances de croisements de palmier à huile.

Ce stage de Master 2 offre une opportunité d'explorer de nouvelles approches pour améliorer la sélection génomique en utilisant des méthodes de pointe en apprentissage automatique et en utilisant un jeu de données de grande taille dans le domaine de l'amélioration des plantes.

Méthodologie :

Le jeu de données à disposition comprend **deux sites** sur lesquels ont été évalués **plusieurs centaines de croisements** de palmier à huile phénotypés pendant **plusieurs années** pour **plusieurs composantes du rendement** et génotypés avec **+20K marqueurs moléculaires de type SNP**.

Les approches de ML pourront inclure : différentes architectures de deep learning (transformers, etc. ; architectures de références publiées récemment : GPFormer, PNNIGS, etc.), Monte-Carlo dropout pour estimer l'incertitude autour des prédictions, feature augmentation avec des informations biologiques connues a priori, algorithmes d'explicabilité pour identifier les SNP les plus fortement associés aux différences de performances entre croisements, etc.

Les analyses seront faites sur un serveur de calcul, avec python (Pytorch, Tensorflow, scikit-learn, etc.) et/ou R.

Le travail de l'étudiant.e consistera à :

- identifier les approches à mettre en œuvre sur la base de discussions avec les encadrants et de la littérature
- écrire les scripts des méthodes choisies
- entraîner les modèles et ajuster les hyperparamètres
- réaliser une comparaison quantitative des performances avec les méthodes classiques (implémentées par les encadrants)

Encadrement / contacts :

- Maxime Ryckewaert, chercheur maths-info, équipe GSP, maxime.ryckewaert@cirad.fr
- David Cros, chercheur en génétique, équipe GSP, david.cros@cirad.fr

Durée / dates :

6 mois en 2026 - dates exactes en fonction des impératifs du candidat

Lieu :

CIRAD, **Agropolis**, Montpellier (<https://maps.app.goo.gl/fkPJtxAm1XXFL03d8>)

Gratification :

Oui, environ 600 euros / mois

Références :

Cros et al, 2025. *Optimizing artificial neural network methodologies for enhanced genomic predictions: a case study with oil palm (Elaeis guineensis) data.* <https://hal.science/hal-04870545>

Montesinos López et al, 2022. *Multivariate statistical machine learning methods for genomic prediction.* <https://library.oapen.org/handle/20.500.12657/52837>