OFFRE DE STAGE

Utilisation des forêts aléatoires dans un but explicatif

ORGANISME : ARVALIS – Institut du végétal
3, rue Joseph et Marie Hackin
75116 PARIS

LIEU D'ACCUEIL : ARVALIS Institut du Végétal
Service des Innovation digitales, matériels et méthodes
Equipe statistiques
91720 – BOIGNEVILLE

TITRE DU SUJET :

Utilisation des forêts aléatoires dans le but d'identifier des variables causales.

SUJET :

Les performances prédictives des modèles « boites noires » telles les forêts aléatoires n'est plus à démontrer. Cependant, de plus en plus, ce type de modèle est utilisé également dans un but explicatif, en vue de mettre en évidence des relations de cause à effet entre variables. L'utilisation des modèles « boites noires » dans un but explicatif est un sujet d'actualité.

En particulier, les forêts aléatoires semblent des bonnes candidates pour expliquer un phénomène. En effet, elles sont faciles à mettre en œuvre et proposent des sorties, comme l'importance des variables et les partials plot, qui peuvent être interprétées dans but explicatif. Cet usage reste cependant à valider.

ARVALIS Institut du végétal souhaite évaluer la pertinence de l'utilisation des forêts aléatoires pour identifier des relations causales. Ce type d'étude est généralement réalisée en simulant des données avec un modèle théorique connu, et en analysant ensuite les données simulées au moyen d'une forêt aléatoire afin de vérifier si cette dernière permet de retrouver le modèle génératif des données. Nous proposons d'utiliser un Réseau Bayésien comme modèle théorique. Ce type de modèle, basé sur un DAG (Directed Acyclic Graph), permet de simuler différents biais de causalité, ce qui permettra d'étudier le comportement d'un modèle de forêts aléatoires face à ces biais. Plusieurs modèles théoriques seront testés. Ces modèles seront construits à dire d'experts. Ils simuleront des phénomènes bien connus des ingénieurs d'Arvalis (par exemple la teneur en mycotoxines des grains de blé à la récolte) afin que les données simulées ressemblent le plus possible aux jeux de données réels analysés habituellement. Différents types de variables (qualitatives, quantitatives), différents schémas relationnels, et différents tailles d'échantillons seront simulés.

Le but du stage est de définir des règles de bonnes pratiques pour l'utilisation des forêts aléatoires dans un but explicatif.

Une interaction forte avec des experts en modélisation et statistique permettront au stagiaire de poser les bonnes hypothèses de travail afin de mener à bien son stage.

OBJECTIF DU TRAVAIL :

L'objectif du stage est :

1. Rédiger un guide de bonnes pratiques de l'utilisation des forêts aléatoires dans un but explicatif.

PROFIL REQUIS :

Elève ingénieur (ou équivalent) – mémoire de fin d'études ou année de césure, en spécialisation statistique ou modélisation (les candidats d'autres spécialisations avec un goût prononcé pour la statistique et la modélisation peuvent aussi postuler).
Attrait pour la transmission des connaissances de façon vulgarisée.
Fort attrait pour la programmation sous R (expérience requise).
Intérêt pour l'agronomie est un plus.
Permis (VL) et véhicule personnel souhaitables.

DUREE : 6 mois (à partir de Mars 2021)

INDEMNITE DE STAGE : Indemnités de stage en vigueur (environ 600€)

LIEU du STAGE : Station expérimentale de Boigneville (91) (limite frontière Loiret).
Aide à la recherche de logement, restauration d'entreprise.

RESPONSABLE(S) :
Emmanuelle HERITIER (statisticienne, Arvalis)
François PIRAUX (statisticien, Arvalis)

CANDIDATURES (CV & REFERENCES + LETTRE DE MOTIVATION) A ADRESSER A :
e.heritier@arvalis.fr et f.piraux@arvalis.fr