Machine Learning interprétable

Nom : **PULL** Prénom : **Yoann** Année : **M2** Semestre : **9**

Nature : **CM** Volume horaire : **12 H** ECTS / Coef : **2**

Prérequis	Cours de Big Data Analytics I, II, III et IV Cours d'introduction à Python Machine Learning sous Python

Résumé	Ce cours présente les notions de base du Machine Learning interprétable et les principales approches techniques permettant de rendre interprétable des modèles de Machine Learning qui ne le sont pas nativement. Le cours s'articule en 3 parties. La première partie est consacrée aux méthodes de type *model-agnostic* qui permettent d'expliquer le fonctionnement d'un modèle indépendamment de sa structure et qui peuvent s'appliquer à n'importe quel modèle de classification ou de régression. Parmi ces méthodes, nous présenterons les Partial Dependence Plot (PDP), les Individual Conditional Expectation (ICE) et les Accumulated Local E¤ects (ALE). La seconde partie est consacrée aux modèles d'approximation globale et locale (surrogate models) et notamment à la méthode Local interpretable model-agnostic explanations (LIME). La dernière partie sera consacrée aux valeurs de Shapley et aux SHapley Additive exPlanations (SHAP). Les applications seront réalisées sur les logiciels Python et SAS.

Objectifs	L'objectif de ce cours double. Il s'agit tout d'abord de familiariser les étudiants aux enjeux de gouvernance de l'Intelligence Artificielle et du Machine Learning, notamment dans le domaine de la Finance, en faisant un focus sur les notions d'interprétabilité et d'explicabilité des algorithmes. Il s'agit ensuite d'introduire les principaux outils permettant de rendre interprétable des modèles de Machine Learning qui ne le sont pas nativement.

Bibliographie	ACPR (2018), Artificial intelligence: challenges for the financial sector. Discussion papers publication, December 2018. ACPR (2020). Governance of artificial intelligence in finance. Discussion papers publication, November, 2020. Lipton, Z.C. (2018), The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery, Queue, 16(3), 31--57. Miller, T. (2019) Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence, 267, 1-38. Molnar C., Casalicchio G. and Bischk B. (2020), Interpretable Machine Learning: A Brief History, State-of-the-Art and Challenge, arXiv:2010.09337v1. Molnar, C (2019), Interpretable machine learning. A Guide for Making Black Box Models Explainable,. https://christophm.github.io/interpretable-ml-book/.

PLAN