Proposition de stage 2022
« Automated Data Cleaning - Expérimentation de méthodes et de solutions »

Descriptif

Contexte R&D :

La R&D d'EDF a pour missions principales de contribuer à l'amélioration de la performance des unités opérationnelles du Groupe EDF, d'identifier et de préparer les relais de croissance à moyen et long terme. Dans ce cadre, le département SEQUOIA (Services, Economie, Questions hUmaines, Outils innovants et IA) est un département pluridisciplinaire (sciences des données et IA, offres et services innovants aux clients, sciences humaines et sociales, économie et stratégies) qui fournit un appui à l'élaboration et au portage des offres, des services et des outils aux Directions opérationnelles du Groupe EDF. Au sein de ce département, le stage sera rattaché au groupe SOAD (Statistiques et Outils d'Aide à la Décision) qui compte une vingtaine d'ingénieurs chercheurs spécialisés en Data Science, Data Engineering, Informatique Décisionnelle et Text Mining. Cette équipe a pour mission de construire et mettre en œuvre les méthodes d'analyse, de fouille et d'enrichissement de données volumineuses d'origines multiples, structurées ou complexes. Le(a) stagiaire sera amené(e) à interagir et évoluer dans un cadre collaboratif avec d'autres chercheurs travaillant sur des problématiques communes au Groupe EDF.

Contexte technique :

Les données volumineuses auxquelles EDF doit faire face sont en grande partie des séries temporelles (données de consommation, données de production, données de capteurs issues de nos centrales, données issues d'objets connectés, données météo, etc.) mais aussi des données tabulaires (stockées dans des bases de données historiques) et des données textuelles (issues des comptes rendus d'intervention, du patrimoine documentaire de nos installations, etc.), entre autres. Ces données peuvent parfois contenir des erreurs (ex : erreurs de saisie, valeurs aberrantes, valeurs dupliquées, etc.) et engendrent ainsi des problèmes de fluidité des interactions, de productivité et de compétitivité à long terme d'un modèle. La gestion de la qualité des données s'avère ainsi un enjeu majeur pour les entreprises dans tous secteurs d'activités confondus. Nous cherchons ainsi à identifier et/ou développer une méthode d'optimisation de ces tâches de préparation de données dans le processus d'apprentissage automatique.

Une des pistes déjà identifiée et en cours d'explorer par notre équipe SOAD est une méthode d'optimisation par l'apprentissage par renforcement avec Q-Learning[1]. Il pourrait être une des méthodes candidates à tester par le stagiaire.

Objectifs et contenu du stage :

L'objectif général du stage sera donc de participer à l'identification et à la mise en œuvre d'un algorithme ou modèle IA capable de proposer la ou les meilleures solutions de nettoyage de données dans le processus de l'apprentissage automatique. Cet algorithme doit être assez générique et réutilisable pour différents modèles (régression, classification, clustering, …) pour les données structurées en premiers temps et non-structurées en deuxième temps. Pour cela, le stagiaire devra :

- Réaliser une revue de littérature des méthodes (open source ou payantes) permettant la gestion (identification et correction) de la qualité de données selon la nature des données (tabulaire, texte, série temporelle) ;

- Expérimenter les méthodes identifiées dans la littérature et/ou développer son propre algorithme pour répondre à la problématique ;

- Évaluation sur un jeu de données fourni par EDF (données textuelles, séries temporelles, …) d'une (ou des méthodes) pertinente(s) basée(s) sur les méthodes en apprentissage automatique permettant l'identification des erreurs et leur correction.

[1] Laure Berti-Equille 1,2. 2019. Learn2Clean: Optimizing the Sequence of Tasks for Web Data Preparation. In Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313602

Profil souhaité

Formation : M2, Ecole d'Ingénieur ou équivalent dans les domaines de compétences : Data Science, Machine Learning et gestion de données

Compétences requises :

- Bonne connaissance en programmation Python
- Bonne connaissance en modélisation de données par les méthodes d'apprentissage automatique
- Bonne connaissance en apprentissage par renforcement sera un plus

Qualités recherchées :

- Esprit scientifique ;
- Esprit critique et de synthèse ;
- Force de proposition ;
- Ouverture sur de nouvelles problématiques ;
- Autonome ;
- Savoir prendre du recul face à un problème.

Conditions matérielles

- Unité d'accueil : Département SEQUOIA, Groupe SOAD (Statistique et Outils d'Aide à la Décision)
  EDF Lab Paris-Saclay, 7 boulevard Gaspard Monge, 91120 Palaiseau.
- Début souhaité : T1 2022.
- Durée : 6 mois
- Rémunération : prévue en fonction de la formation
- Pour candidater : transmettre par mail un CV et une lettre de motivation à : somsakun.maneerat@edf.fr