EDF R&D : PROPOSITION DE STAGE DE FIN D'ETUDES (6 MOIS)

IA génératives adaptées aux données tabulaires pour la génération de données synthétiques et l'imputation de valeurs manquantes

Le contexte :

Le département SEQUOIA (Services, Economie, Questions hUmaines, Outils innovants et IA) de la R&D intervient en appui de la Direction Marketing des Clients Particuliers et de la Direction des Systèmes d'Informations et du Numérique de la branche Commerce d'EDF.

Un des enjeux d'EDF Commerce est l'amélioration, l'accélération du déploiement et la pérennisation de ses algorithmes d'IA, notamment les scores à destination des clients particuliers et prospects (scores d'appétence aux services par exemple).

EDF s'engage pour l'utilisation éthique et responsable des données de ses clients, dans un contexte réglementaire fort (RGPD) et avec un objectif de sobriété numérique. **L'utilisation, la diffusion et la conservation** dans le temps de ces données sont en effet strictement restreintes. Cela nécessite un appui méthodologique R&D en continu pour identifier des alternatives innovantes aux contraintes liées à l'utilisation des données personnelles. La génération de données synthétiques est un axe de recherche qui s'inscrit dans ces ambitions EDF.

En réponse à ces besoins, le projet ACACIA (Analyse, Connaissance client, Algorithmes pour Commerce et IA) s'intéresse aux algorithmes d'IA génératives, spécifiquement appliqués aux données structurées (informations clients particuliers de type contractuelles, typologie logement / foyer, équipements, consommation…). Ces données peuvent être catégorielles, ordinales et numériques.

L'IA générative appliquée aux données structurées permettrait de disposer de données à la fois quasi-réelles, anonymisées et conservables dans le temps pour :

- des besoins techniques :
  - du pré-entraînement des modèles ;
  - de la data-augmentation ;
  - de l'imputation de données manquantes ou de la détection d'anomalie ;
- des besoins finaux :
  - des études statistiques post-hoc (après effacement des données d'origine) ;
  - des études longitudinales (par exemple : impact période covid, conception d'offres, élasticité prix) ;
  - et enfin faciliter la publication des travaux de recherche, en fournissant des jeux de données générés non-sensibles.

Des travaux précédents ont déjà permis d'explorer quelques méthodes basées sur des modèles de diffusion. Une première phase d'état de l'art sera à réaliser pour identifier les derniers modèles les plus pertinents, qui seront à évaluer dans une deuxième phase. On s'intéressera notamment aux critères de fidélité, d'utilité et de privacy pour l'évaluation.

Ce champ de recherche étant en constante évolution, la R&D souhaite se tenir au plus près de l'état de l'art et il est possible que de nouvelles méthodes soient mises en avant selon les publications.

Le stage :

Le but du stage est de réaliser un benchmark de méthodes d'IA générative pour les données structurées. L'évaluation des méthodes se fera sur plusieurs jeux de données, issus de la revue de littérature et propres à EDF. Ce travail sera articulé en plusieurs phases, qui seront adaptées au fur et à mesure des découvertes et résultats du stage.

Realisation d'un Etat de l'Art

- **Lecture et recherche d'articles scientifiques pertinents sur le sujet**
  - Etude des modèles de diffusion classiques (tels que DDPM, DDIM, DiT).
  - Emphase sur leurs extensions pour les données tabulaires (comme TabDDPM, MTabGen).
  - Ouverture sur d'autres modèles concurrents des benchmarks (comme TVAE, CTGAN, Tabsyn).

  Une bibliographie non-exhaustive est jointe à l'offre de stage.
- **Echange avec les chercheurs et thésards au sein de la communauté scientifique d'EDF R&D**

Entrainement d'un modele generatif de donnees structurees Client

- **Implémentation en Python et PyTorch des modèles génératifs identifiés**
  - Choix de modélisation (normalisation de variables quantitatives, gestion de valeurs manquantes et aberrantes, etc.)
  - Mise à l'épreuve de différentes approches pour la génération de variables catégorielles (comme le type d'équipement de chauffage, type de contrat, etc.). Génération de données qualitatives et quantitatives dans une même architecture
- **Benchmark des modèles**
  - Comparaison des performances : calcul de métriques pour les critères de fidélité (comparaison des données réelles VS générées) et d'utilité (performances équivalentes de modèles de prédiction par exemple) et création de visualisations associées
  - Contribuer à l'extension d'un package interne présentant déjà des visualisations comparatives entre deux jeux de données.

Use cases - selon l'avancement du stage, un ou deux cas d'application seront etudies :

- **METHODOLOGIQUE : adaptation d'un modèle génératif pour la complétion de valeurs manquantes**
  - **Adaptation des modèles génératifs identifiés pour des cas d'imputation**
    - Adaptation des travaux en introduisant des masques aléatoires sur les données d'entraînement.
  - **Benchmark des modèles**
    - Comparaison des performances d'imputation avec celles d'autres modèles classiques (par exemple MissForest).

- **METIER : Utilisation d'un modèle génératif pour la simulation d'impact d'un changement d'appareil sur les consommations annuelles ou mensuelles**
  - **Benchmark des modèles**
  - **Lien possible avec un projet connexe**
    - Traitant d'IA générative appliquée aux données de séries temporelles
    - Traitant d'évaluation d'impact / inférence causale

Informations complémentaires :

La R&D propose ce stage de fin d'études, à destination d'étudiants en écoles d'ingénieurs ou Master 2, spécialisés en Statistiques / Data Science / Deep Learning / IA. Une appétence pour la recherche sera appréciée.

Compétences : L'étudiant(e) sera amené(e) à mettre en œuvre et/ou acquérir des compétences en :

- Modèles génératifs (modèles de diffusion DDPM, VAE, Diffusion Transformers)
- Deep Learning (CNN, Transformers)
- Machine Learning (Random Forest, GBM, SVM)
- Lecture et synthèses d'articles de recherche

Le langage de programmation sera Python ; les développements seront réalisés avec le framework PyTorch. De bonnes pratiques sur Git seront valorisées. Une expérience démontrée avec des modèles de Deep Learning et ces outils serait particulièrement appréciée.

Dates : Stage d'une durée de 6 mois. La date de début est flexible entre février et mai 2025.

Lieu du stage : EDF Lab Paris-Saclay – Recherche et Développement, 7 Bd Gaspard Monge, 91120 Palaiseau. Le stagiaire pourra bénéficier de mesures de télétravail en fonction du niveau d'autonomie.

Contacts :
Laure CAREL (Ingénieure Chercheuse), mail : laure.carel@edf.fr
Laurent BOZZI (Ingénieur Chercheur Expert), mail : laurent.bozzi@edf.fr
Alice DUQUENNE (Cheffe de projet), mail : alice.duquenne@edf.fr
Merci d'envoyer un C.V et une lettre de motivation sur ces trois e-mails.

Horaires : 35 h / semaine.

Indemnité : en fonction des formations.

Bibliographie :

Les articles de recherche suivants sont pertinents pour les travaux du stage :

- Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2022). **TabDDPM: Modelling Tabular Data with Diffusion Models**. *arXiv.org*. https://doi.org/10.5555/3618408.3619133
- Qian, Z., Cebere, B., & Mihaela, V. D. S. (2023, January 18). ***Synthcity: facilitating innovative use cases of synthetic data in different data modalities***. arXiv.org. https://arxiv.org/abs/2301.07573
- Stekhoven, D. J., & Bühlmann, P. (2011). **MissForest—non-parametric missing value imputation for mixed-type data**. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597
- Villaizán-Vallelado, M., Salvatori, M., Segura, C., & Arapakis, I. (2024, July 2). ***Diffusion models for tabular data imputation and synthetic data generation***. arXiv.org. https://arxiv.org/abs/2407.02549
- Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019, July 1). ***Modeling Tabular data using Conditional GAN***. arXiv.org. https://arxiv.org/abs/1907.00503
- Zhang, H., Zhang, J., Srinivasan, B., Shen, Z., Qin, X., Faloutsos, C., Rangwala, H., & Karypis, G. (2023, October 14). ***Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space***. arXiv.org. https://arxiv.org/abs/2310.09656
- Zhao, Z., Kunar, A., Birke, R., & Chen, L. Y. (2022, April 1). ***CTAB-GAN+: Enhancing Tabular Data Synthesis***. arXiv.org. https://arxiv.org/abs/2204.00401
- Lin, X., Xu, C., Yang, M., & Cheng, G. (2024). **CTSyn: A Foundational Model for Cross Tabular Data Generation**. arXiv preprint arXiv:2406.04619.