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

IA génératives adaptées aux données tabulaires pour la génération de données synthétiques

Le contexte

Un des enjeux de la direction Commerce d'EDF est l'amélioration de ses **algorithmes d'IA**, notamment les scores à destination des clients particuliers et prospects, pour certains développés par les équipes R&D.

Par ailleurs, 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.

En réponse à ces besoins et contraintes, au sein d'une équipe Data Science de la R&D d'EDF, le projet ACACIA (Analyse, Connaissance client, Algorithmes pour Commerce et IA) s'intéresse aux algorithmes d'IA générative, spécifiquement appliqués aux données structurées (informations clients particuliers de type contractuelles, typologie logement / foyer, équipements, consommation…), pour générer des données synthétiques.

Générer de telles données, à la fois quasi-réelles, anonymisées et conservables dans le temps, servirait :

- des besoins techniques (pré-entraînement des modèles, data augmentation, …) ;
- des besoins finaux :
    - études statistiques post-hoc (après effacement des données d'origine) ;
    - études longitudinales (par exemple : impact période covid, conception d'offres, élasticité prix) ;
    - facilitation de la publication des travaux de recherche, en fournissant des jeux de données générés non-sensibles.

Des travaux précédents ont déjà permis d'explorer certaines méthodes, notamment des **modèles de diffusion**, et de mettre en place divers indicateurs pour évaluer les critères de **fidélité, d'utilité et de privacy**.

Le stage

L'objectif principal du stage est de générer un échantillon synthétique du portefeuille EDF comprenant à la fois des informations contractuelles et des données de consommation, à un pas hebdomadaire ou plus fin. D'autres cas d'usages seront étudiés. Les étapes seront adaptées au fur et à mesure des découvertes et résultats du stage, et certaines pourront être faites en parallèle.

Realisation d'un Etat de l'Art

- **Lecture et recherche d'articles scientifiques pertinents sur le sujet (données tabulaires - numériques et catégorielles -, comprenant de petites séries temporelles).**
    - Etude des modèles précédemment étudiés (TabDiff, CausalDiffTab).
    - Veille sur d'autres modèles concurrents des benchmarks et d'autres approches pertinentes.
    - D'autres types de données (par exemple, des dates), pourraient être intéressants, et nécessiter d'autres types de modèles.
- **Echange avec les chercheurs et doctorants au sein de la communauté scientifique d'EDF R&D**
    - Echanges autour de la génération conditionnelle de séries temporelles et autres sujets connexes.
    - Présentations des travaux de stage.

Entrainement d'un modele generatif de donnees tabulaires client (Python)

- **Prise en main des modèles précédemment étudiés (TabDiff, CausalDiffTab) et adaptation** à des données plus complètes et plus fines.
- **Implémentation et utilisation d'autres modèles 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 petites séries temporelles (données de consommation) dans un modèle de génération tabulaire.
- **Benchmark des modèles**
    - Comparaison des performances : calcul de métriques pour les critères de fidélité (comparaison des données réelles VS générées), de privacy (sécurité) et d'utilité (performances équivalentes de modèles de prédiction par exemple), et analyse de visualisations associées.
    - Possibilité d'ajouter de nouveaux indicateurs pertinents, et/ou construire une métamétrique (condensant les multiples indicateurs).
    - Contribution à l'extension d'un package interne présentant déjà des visualisations comparatives entre deux jeux de données.

Autres axes – selon l'avancement du stage, un ou plusieurs seront etudies :

- **Use case métier,** à préciser en accord avec les équipes de la direction Commerce (par exemple : utilisation des données synthétiques pour estimer le niveau d'aisance financière).
- **Use case méthodologique :** pré-entraînement de modèles de la R&D avec des données générées par un modèle du stage (par exemple : modèle de détection du chauffage).
- **Amélioration de la frugalité des modèles :**
    - Ajouter ou améliorer un earlystopping dans les modèles implémentés.
    - Evaluer, à titre indicatif, l'empreinte carbone des modèles implémentés.

Profil recherché

La R&D propose ce stage de fin d'études, à destination d'étudiant.es en écoles d'ingénieurs ou Master 2, spécialisé.es en Statistiques / Data Science / Deep Learning / IA. L'étudiant(e) sera amené(e) à mettre en œuvre et/ou acquérir des compétences avec une certaine autonomie.

Compétences requises :

- Deep Learning (CNN, Transformers, Fondation Models) et Machine Learning (utilisés pour certains use cases).
- Lecture et synthèse d'articles de recherche.
- Maîtrise du langage de programmation Python et du framework PyTorch.

Compétences fortement appréciées :

- Modèles génératifs : modèles de diffusion classiques (DDPM, DDIM, DiT…) et adaptés aux données tabulaires (TabDDPM, MTabGen, …), modèles de type GAN ou VAE (TVAE, CTGAN, Tabsyn, …).
- Bonnes pratiques Git.

Informations complémentaires

Ce que le stage vous apportera :

- **La connaissance et la maîtrise de méthodes Data Science à la pointe,** grâce vos travaux et à votre intégration dans une équipe scientifique, où vous collaborerez avec des chercheur.ses et d'autres stagiaires.
- **L'immersion dans le monde de l'énergie et la contribution à des travaux à impact :** vos recherches seront utilisées par les entités métier d'EDF et répondront à des problèmes concrets.

Dates : Stage d'une durée de 6 mois. La date de début est flexible entre février et mai 2026.

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 Senior), mail : laurent.bozzi@edf.fr
Maxime LEPETIT (Chef de projet), mail : maxime.lepetit@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 :

- 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
- 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
- Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, Jure Leskovec (2024, October 17) ***TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation***. arXiv.org. https://arxiv.org/abs/2410.20626
- Zhong Li, Qi Huang, Lincen Yang, Jiayang Shi, Zhao Yang, Niki van Stein, Thomas Bäck, Matthijs van Leeuwen (2025). ***Diffusion Models for Tabular Data: Challenges, Current Progress, and Future Directions***. arXiv.org. https://arxiv.org/abs/2502.17119
- Jia-Chen Zhang, Zheng Zhou, Yu-Jie Xiong, Chun-Ming Xia, Fei Dai. (2025). ***CausalDiffTab: Mixed-Type Causal-Aware Diffusion for Tabular Data Generation.*** arXiv.org. https://arxiv.org/abs/2506.14206