

Séphora Mitosse

Data Scientist

Data Scientist spécialisée en statistiques appliquées et Machine Learning, avec une expérience significative à l'INSEE sur des enjeux de fiabilité, d'imputation et de valorisation de données socio-économiques. Capable d'intervenir de la phase d'exploration à la mise en production de modèles (ML, NLP), avec un fort sens de la structuration, du reporting et de l'aide à la décision.

@ mitossedes@gmail.com

📍 35700 Rennes

📞 +33 605732129



EXPERIENCE PROFESSIONNELLE

Alternance | Data Scientist

Institut National de la Statistique et des Études Économiques (INSEE)

📅 Septembre 2025 – En cours 📍 Rennes, France

- **Fiabilisation de données** de revenus par la conception d'un protocole méthodologique d'évaluation de la stabilité et de la robustesse des imputations (tests statistiques, validation croisée et temporelle).
- **Amélioration mesurable** de la qualité des données via le développement de méthodes d'imputation avancées par Machine Learning (Random Forest, MICE), démontrant une meilleure stabilité des prévisions que les approches classiques.
- **Optimisation et interprétabilité** des modèles par sélection de variables et régularisation (LASSO, Ridge, Elastic Net), avec production de reportings décisionnels pour validation méthodologique auprès des équipes métiers.

CDD | Data analyst

École des hautes études en santé publique (EHESP)

📅 Décembre 2024 – Août 2025 📍 Rennes-France

- Analyse, préparation et exploitation de données avec Python, R et Excel, dans un contexte d'études en santé publique.
- Production de rapports analytiques automatisés et exploitation de données d'enquêtes (R Markdown, Sphinx).

Stage | Statisticienne

DPSSEL (Ministère de l'Économie)

📅 Mai 2023 – Juillet 2023 📍 Cotonou-Bénin

- Conception et exploitation d'enquêtes statistiques sur la mobilité urbaine et les transports collectifs à Cotonou;
- Nettoyage, contrôle de cohérence et imputation de données sous R

PROJETS

- **Système de recommandation culinaire (NLP/LLM) :** Développement d'un Chatbot intelligent basé sur des données scrapées. Implémentation de techniques de Vector Search (TF-IDF/Embeddings), clustering KMeans et déploiement d'une interface interactive via Streamlit.
- **Détection du churn client :** Comparaison de modèles de Machine Learning avec gestion du déséquilibre de classes, évaluation par métriques adaptées (ROC-AUC, F1-score, recall) et validation croisée afin d'identifier le modèle le plus performant.

FORMATION

Master Mathématiques Appliquées, Statistique
parcours Data Science et IA

Université de Rennes 2 /Rennes, France

📅 Septembre 2024 – En cours

Cours principaux : Machine learning avancé - Deep learning - Big Data - Statistique en grande dimension - NLP - Optimisation pour la science des données.

Licence 3 Mathématiques et Informatique
Appliquées aux Sciences Humaines et Sociales
Parcours Mathématiques-Economie

Université de Bretagne Occidentale (UBO) /Brest , France

📅 Septembre 2023 – Mai 2024

Licence 3 Statistiques parcours Statistiques
Economique et Sectorielle

ENEAM /Cotonou , Bénin

📅 Septembre 2020 – Octobre 2021

COMPÉTENCES

- **Langages et outils** : R (tidyverse, tidymodels), Python (Pandas, PySpark, NumPy, statsmodels), SQL;
- **Machine Learning & Deep Learning** : scikit-learn, TensorFlow, Keras, PyTorch;
- **Cloud & Big Data** : Google Cloud Platform (GCP), Google Cloud Storage;
- **NLP & Generative AI (Expertise)** : LLMs, Transformers (Hugging Face), RAG, embeddings, spaCy, NLTK, classification de textes, analyse de sentiments;
- **Data apps & APIs** : RShiny, Streamlit, Flask (API REST);
- **Visualisation et reporting** : Power BI, Tableau, Matplotlib, Seaborn;
- **MLOps & versioning** : Git, GitHub, GitLab;
- **Soft skills** : rigueur, esprit d'équipe et esprit analytique.

LANGUES

Français Langue maternelle

Anglais : Niveau B2