

Mourad DO-REGO

Data Scientist & Générative AI

Tél : 07 52 06 64 48 | mouwahiddorego@gmail.com | linkedin.com/in/mourad-do-rego

PROFIL

Data Scientist passionné par l'apprentissage et l'application de nouvelles techniques pour produire des modèles d'IA performants et à fort impact. Maîtrise de Python et de ses bibliothèques de data science ainsi que de SQL pour l'extraction et l'optimisation de requêtes. Solide connaissance des algorithmes et structures de données appliqués à la science des données : modélisation prédictive, machine learning, Gen AI, LLM, NLP. Capacité à formuler des problèmes analytiques, synthétiser les résultats et produire des représentations visuelles convaincantes assorties de recommandations concrètes.

PROJETS

- **Biomedical question answering RAG & LLMs : Score global RAGAS 0.792** (Python, LangChain, ChromaDB, Groq/Llama 3.3 70B, FastAPI, Streamlit) [github](#) : collecte de 652 abstracts PubMed via API (2000 à 2025) sur la santé mentale des jeunes diplômés, 836 chunks indexés dans ChromaDB (embeddings *all-MiniLM-L6-v2*, 384 dimensions), pipeline RAG avec stratégie de fallback, évaluation RAGAS-style, API REST FastAPI + dashboard interactif déployé sur Streamlit.
- **Prévision de consommation énergétique : MAPE 0.74% sur données réelles PJM (145k obs., 2003–2018)** (Python, XGBoost, LSTM, FastAPI, Streamlit, Docker) [github](#) : feature engineering (35 features : lags, moyennes mobiles, données météo Open-Meteo, encodage cyclique), comparaison 5 modèles (Naïf → Régression Linéaire → LSTM → XGBoost), intervalles de confiance calibrés via Conformal Prediction, prévision itérative H+24, API REST FastAPI sur Render, dashboard monitoring Streamlit Cloud.
- **Optimisation des prix e-commerce : +18.7% de CA via élasticité prix et Machine Learning** (Python, scikit-learn, Power BI) [github](#) : analyse de 108k commandes, calcul d'élasticité prix par régression log-log, imputation MICE, Random Forest, optimisation scipy, simulation Monte Carlo 10k itérations, dashboard Power BI.
- **Prédiction du churn client télécom : Recall 81.6%, gain net estimé 54 500€** (Python, XGBoost, FastAPI, Streamlit, Docker) [github](#) : EDA sur 7k clients, feature engineering, XGBoost avec optimisation du seuil de décision, explicabilité SHAP, API REST FastAPI dockerisée, dashboard interactif déployé sur Streamlit Cloud.
- **Analyse de sentiment NLP : 85% d'accuracy sur 568k avis** (Python, NLTK, scikit-learn, Streamlit) [github](#) : pipeline NLP end-to-end, nettoyage texte, SVM LinearSVC avec cross-validation, extraction de topics TF-IDF + LDA, dashboard interactif Streamlit.

EXPÉRIENCES PROFESSIONNELLES

Data Scientist - Machine Learning & Big Data (GCP)

Groupe Carrefour France

03/2025 – 09/2025

Mondeville, France

- Conception et déploiement d'un pipeline d'automatisation end-to-end mis en production (-35% délais de traitement).
- Requête et exploitation de données massives du datalake via BigQuery pour créer de tables consolidées et historisées.
- Optimisation de requêtes SQL avancées pour alimenter des dashboards de KPI Loker Studio. (-65% à -99% taux d'erreurs).
- Collaboration transverse avec les équipes métiers pour traduire les besoins en solutions analytiques adaptées.
- Déploiement de cas d'usage IA en production avec suivi de la performance et documentation des traitements.

Data Scientist – Analyse de fraude

Société TEST S.A

05/2024 – 06/2024

Paris, France

- Analyse de données comportementales pour identifier des profils à risque.
- Construction de score de détection de fraude et segmentation de populations.
- Production de reporting analytique pour la prise de décision.

COMPÉTENCES TECHNIQUES

Python & data science : pandas, scikit-learn, NumPy, Matplotlib, Seaborn, XGBoost, LightGBM, TensorFlow, LangChain, ChromaDB.

LLM & GenAI : RAG (Retrieval-Augmented Generation), Groq API (Llama 3.3 70B), embeddings sémantiques, évaluation RAGAS.

SQL : Rédaction et optimisation de requêtes avancées (BigQuery, Redshift), extraction de données multi-sources, contrôle qualité.

Algorithmes & structures de données : Arbres de décision, NLP, séries temporelles, clustering, optimisation, recherche vectorielle.

Visualisation & communication : Power BI, Looker Studio, Streamlit, R Shiny, Excel.

Déploiement & MLOps : FastAPI, Docker, Render, Streamlit Cloud, Git, CI/CD.

Autres : R, SAS, GCP, Spark, ETL, Data Warehouse, Biopython/PubMed API.

FORMATION

Master Méthodologie et Statistique en Recherche Biomédicale

Université Paris Saclay

2025 – 2026

Paris, France

Master Mathématiques Appliquées, Statistiques

Université de Caen Normandie

2023 – 2025

Caen, France

Licence 3 Mathématiques et Informatique Appliquées

Université de Caen Normandie

2022 – 2023

Caen, France

LANGUES

Français – Courant | Anglais – Professionnel