

MAHÉ BEGNIS

Data Scientist & GenAI Engineer

Lille - mobilité Lille et Paris | mahebegnis9@gmail.com | linkedin.com/in/mahebegnis | github.com/mahebegnis

Deux ans d'alternance en cabinet de conseil en intelligence artificielle, en fin de Master Data & IA. **Machine learning** : prévision de la demande sur 7,17 M de lignes de ventes, 3 fois plus précise que le processus en place. **IA générative et mise en production** : agents RAG connectés aux ERP, serveurs MCP et chaînes d'extraction documentaire, déployés et maintenus sur Kubernetes chez des clients industriels, du retail et de la formation. **Démarche de recherche** : protocoles d'évaluation reproductibles, du jeu de test isolé (AUC 0,996 en classification supervisée) à la revue de littérature PRISMA 2020 assistée par LLM. Large autonomie du cadrage à la recette client.

COMPÉTENCES TECHNIQUES

Machine learning

Apprentissage supervisé : LightGBM, CatBoost, XGBoost, Random Forest, régression linéaire et logistique, SVM à noyau gaussien (RBF), k-NN, scikit-learn. Apprentissage non supervisé : k-means, DBSCAN, clustering agglomératif, ACP. Prévision de séries temporelles, modèles à deux étages pour données à excès de zéros, validation temporelle, analyse d'importance des variables.

IA générative, LLM & NLP

Architectures RAG : découpage, recherche hybride, citation des sources, cloisonnement des corpus. Agents outillés, serveurs MCP, ingénierie et évaluation de prompts, contrôle des hallucinations. Vertex AI Search, ChromaDB. Gemini, Claude, GPT, Mistral. NLP : TF-IDF, sac de mots, Word2Vec (CBOW), Naive Bayes, NER-CRF, similarité cosinus, réseaux récurrents (RNN, LSTM).

Statistiques & préparation des données

Analyse univariée et bivariée : statistiques descriptives, analyse de corrélation, VIF. Tests d'hypothèses : khi-2, Fisher, ANOVA, tests de normalité. Détection et traitement des valeurs manquantes (mécanismes MCAR, MAR, MNAR) et des valeurs aberrantes (Isolation Forest, Tukey, DBSCAN), normalisation des variables (standard, min-max, robuste) et des échantillons (Normalizer), encodage, ingénierie de variables. Métriques : WAPE, MAE, MASE, biais de prédiction, précision, rappel, F1, AUC, score de silhouette.

Données & cloud

Python (pandas, numpy), TypeScript, SQL, PostgreSQL, BigQuery. Google Cloud Platform : Vertex AI, Agent Builder, Cloud Run, Cloud Functions, GCS. Écosystème Microsoft : Azure, Entra ID (inscriptions d'applications, autorisations Graph), Graph, SharePoint. Restitution : Looker Studio, Power BI.

Industrialisation & conformité

FastAPI, Docker, Kubernetes (bases), GitLab, Playwright, etc. Pipelines reproductibles, journalisation structurée, tests de non-régression, mise en production et suivi des performances. RGPD dès la conception : minimisation, pseudonymisation, purge, journal d'audit ; résidence des données en UE. **Certifié SecNumEdu (ANSSI) et atelier RGPD (CNIL).**

EXPÉRIENCE PROFESSIONNELLE

Six and See - Développeur IA / Data Scientist (alternance)

Septembre 2025 - 2027 | Lille | Cabinet de conseil en intelligence artificielle et data

Réalisation de projets clients en **large autonomie** : cadrage technique, développement, mise en production, recette et suivi. Interlocuteur technique direct des clients : avancement, arbitrages, soutenance des projets.

Prévision de la demande par machine learning

L'Incroyable - distribution spécialisée, 40 magasins

- Chaîne complète menée seul : acquisition, nettoyage, analyse exploratoire, ingénierie de variables, entraînement, évaluation, industrialisation sur 7,17 M de lignes de ventes.
- Modèle par famille : LightGBM, perte MAE, 43 variables (moyennes glissantes, stock, calendrier, promotions). WAPE 0,201, MAE 25,5, biais -1,4 %, MASE 0,951, contre 0,621 pour les commandes passées par les magasins. Gain vérifié sur 19/19 familles et 40/40 magasins.
- Modèle par produit : architecture à deux étages (classifieur d'occurrence puis régresseur conditionnel) sur un panel à 76,8 % de zéros.
- Étude comparative documentée CatBoost / LightGBM / XGBoost, découpage temporel strict entraînement / validation / test, jeu de test jamais vu.
- Décision de modélisation défendue auprès du client : prédire la demande latente et non les ventes observées, après mise en évidence d'une boucle de rétroaction induite par le stock.
Python, pandas, LightGBM, CatBoost, XGBoost, scikit-learn, régression logistique

Agent IA de service après-vente

Industrie de la menuiserie

- Conception et développement d'un agent orchestrant en parallèle une base documentaire RAG (111 notices PDF, vidéos transcrites, catalogue produits) et les API ERP temps réel (commandes, devis), avec croisement BigQuery entre commande, catalogue et fiche technique.
- Réponse structurée en blocs : synthèse de commande, diagnostic technique, sources PDF citées, liens de suivi, avec citation systématique des sources.
- Industrialisation de l'intégration des notices : validation de schéma Pydantic, journalisation structurée, manifeste d'exécution. Mise en production sans régression.

- Tableau de bord de pilotage Looker Studio (satisfaction, typologie des demandes) et formation du développeur interne du client à la maintenance et à l'évolution de la solution.
Vertex AI Agent Builder, Gemini 2.5 Flash, Cloud Run, BigQuery, GCS, Python
- Serveur MCP - système d'information conversationnel du pôle formation**
Six and See - outil interne, en production
- Système d'information de l'activité de formation certifiée Qualiopi, piloté en langage naturel par le dirigeant : clients, sessions, inscriptions, boîte mail, indicateurs de conformité, en remplacement d'une gestion manuelle.
 - 64 outils métier exposés à un modèle de langage via le protocole MCP, dont 14 générateurs produisant directement au format Word les documents réglementaires exigés par la certification : contrats, convocations, attestations, feuilles de présence, certificats de réalisation.
 - Modèle de données PostgreSQL de 11 tables, architecture en couches séparant le routage de la logique métier : migration du transport STDIO vers HTTP sans modifier la logique.
 - Sécurisation par jeton à comparaison en temps constant, puis Resource Server OAuth 2.0 (RFC 9728) adossé à Keycloak.
- Mise en production sur Kubernetes : image Docker multi-étages non privilégiée, base PostgreSQL managée, TLS automatisé, audit d'image sans secret embarqué.*
TypeScript, Node.js, PostgreSQL, MCP SDK, Docker, Kubernetes, OAuth 2.0 / Keycloak

Structuration technique et gouvernance des développements

- Six and See - chantier interne transverse, en cours*
- Conception, avec le second développeur de l'équipe, du dépôt GitHub partagé de l'entreprise : organisation des dépôts, chaîne de traitement, outillage et contrôles à l'entrée (détection de secrets par gitleaks, analyse de sécurité automatisée).
 - Instauration d'une revue de sécurité systématique à chaque lot livré, formalisation de la méthodologie data et des critères d'arbitrage des outils et frameworks.
 - Deux applications déployées et maintenues en production sur le cluster Kubernetes OVH, et contribution à son évolution : ajout d'un nœud de secours et d'un nœud dédié à la montée en charge.
- GitHub, Kubernetes (OVH), Docker*

Autres réalisations chez Six and See

- **Diagnostic Flash IA** (produit interne, en production) : service d'auto-diagnostic de maturité IA, FastAPI, PostgreSQL, génération et envoi automatisés d'un rapport PDF, 28 contrôles automatiques dont 17 bloquants, déployé sur Kubernetes.
- **Organisme de formation** (preuve de concept) : classification et extraction documentaire sous contrainte RGPD (43 classes, 108 documents réels), précision 0,967, rappel 0,947, F1 0,957 ; BigQuery en région UE, données sensibles.
- **Fédération professionnelle** (confidentiel) : agent juridique sur corpus de commande publique et étude comparative de trois architectures RAG sur protocole d'évaluation reproductible.
- **Groupe de propreté** (confidentiel) : extraction OCR de bons de travaux manuscrits par modèle de vision, évaluée contre un jeu de référence annoté à la main.
- **INHNI** (formation professionnelle) : automatisation de la production des contrats pédagogiques, service Python et application web de validation, intégration SharePoint via Microsoft Graph, RGPD dès la conception.

STAGES

Stagiaire technicien - Observatoire de la Cybersécurité de l'Océan Indien

- Mai - juillet 2025 | La Réunion*
- Contribution au développement d'une plateforme d'alerting : collecte de données de cybersécurité via les API Shodan et Censys, stockage PostgreSQL, visualisation temps réel, capteurs IoT.
- Azure, Python, PostgreSQL, Git*

Stagiaire - Groupe Les Flamboyants

- Décembre 2023 - janvier 2024 | La Réunion*
- Participation au déploiement d'une solution de détection et réponse sur postes (EDR Kaspersky) en milieu hospitalier.

FORMATION

2025 - 2027	Master Informatique, parcours Data & Intelligence Artificielle - École du Numérique, FGES, Université Catholique de Lille (diplôme délivré par l'UPHF). En alternance chez Six and See. <i>Modélisation statistique, Machine learning, Deep learning, NLP, Acquisition et nettoyage des données, Analyse prédictive</i> <i>Projet ML (ClinVar) : classification de variants génétiques. Pathogénicité, AUC 0,996 sur un jeu de test isolé, 8 modèles comparés, écart validé par test de DeLong ; désaccords entre laboratoires, AUC 0,79.</i> <i>Projet de recherche : revue systématique PRISMA 2020 sur les jumeaux numériques pour l'optimisation énergétique des bâtiments. 704 références, présélection assistée par LLM puis validation manuelle, 45 études incluses ; article IEEE de 29 pages et pipeline Python publiés sur GitHub.</i>
2023 - 2025	ESIROI , La Réunion - 2e année de cycle préparatoire intégré, puis 1re année du cycle ingénieur informatique. Réorientation vers une spécialisation Data & IA en alternance.
2022 - 2023	Prépa des INP - 1re année de cycle préparatoire.

Certifications : SecNumEdu - ANSSI (2025) | Atelier RGPD - CNIL (2025)

DISTINCTIONS, LANGUES ET CENTRES D'INTÉRÊT

- 1re place nationale :** Nuit de l'Informatique, défi Ubisoft (2024)
- Intervenant :** conférence sur la sécurité et l'éthique de l'IA, devant un public de DPO et de RSSI (2025), sous la bannière de l'OCOI
- Langues :** Français, anglais courant (certification CLES B2)
- Centres d'intérêt :** sciences du vivant et astronomie, musculation, cinéma, voyages