COLLINS NYATUNDO NYAGAKA

AI & Machine Learning Engineer

Nairobi, Kenya | cnyagakan@gmail.com | linkedin.com/in/collinsnyagaka001 | github.com/CollinsNyatundo

PROFILE

AI and machine-learning engineer in Nairobi building ML models that hold up under honest evaluation and LLM systems that are safe to run: a legal research platform that retrieves from 61,000+ passages of Kenyan law, a fraud-detection model scoring in 1.4 ms (p95), and an AML/KYC monitoring prototype for Kenyan financial institutions. Focused on fintech, legaltech and agritech in Kenya.

TECHNICAL SKILLS

Machine Learning	Python, SQL, scikit-learn, LightGBM, XGBoost, CatBoost, PyTorch, Optuna, SHAP, imbalanced learning (SMOTE), probability calibration, anomaly detection (Isolation Forest, River)
AI Engineering	RAG, hybrid (vector + full-text) search with reranking, retrieval evaluation, agent and tool loops, multi-agent orchestration, DSPy, LangChain, MCP; Voyage AI, Gemini, NVIDIA NIM, OpenAI, OpenRouter, vLLM
Data & Serving	FastAPI, Express (TypeScript), PostgreSQL / Supabase (pgvector, row-level security), Redis, Apache Kafka, Airflow, Kedro
MLOps	MLflow, Docker, GitHub Actions CI, OpenTelemetry, gVisor sandboxing, pytest

EXPERIENCE

Backend AI Engineer (Intern) · *FlyRank AI* — Oct 2026 – Present · Remote

- Building AI backend services: RAG, structured outputs, tool-calling agents and evaluation, with grounding and validation.

ICT Officer · *Office of the Director of Public Prosecutions* — Jul 2026 – Present · Kenya

- Support the security, reliability and backup of institutional systems and databases, and train users on them.

Independent AI & ML Engineer — May 2022 – Present
Freelance and contract work — Remote, Nairobi

- **ComplianceGuard (AML/KYC):** built the backend of a transaction-monitoring prototype for Kenyan financial institutions: Kafka ingestion, 35 typology rules, a composite risk score from seven scoring engines, Isolation Forest and streaming (River) anomaly detection, and SAR/STR/CTR reporting.
- **Ulti-IDE:** built an AI development environment in which agents write, run and test code inside gVisor-sandboxed containers. Added a four-tier LLM gateway (OpenRouter, NVIDIA NIM, local vLLM, semantic cache) and a self-repair loop that turns failing traces into tests and candidate patches.
- Built ETL pipelines in Python, Airflow and Kedro that prepare high-volume transaction data for modelling.

Data Scientist — Sep 2024 – Feb 2025
Lish AI Labs — Nakuru, Kenya

- Built and validated statistical and machine-learning models in Python and SQL with the risk and product teams for quarterly KPIs, and trained 10+ operators to act on the results through Tableau and Python dashboards.

SELECTED PROJECTS

LegalAI Kenya: legal research workspace for advocates — legalai-kenya.vercel.app

- Multi-tenant RAG platform over 61,683 passages of Kenyan law (543 Acts, 990 judgments): Voyage embeddings, pgvector and full-text search fused with reciprocal rank fusion, and Voyage reranking. It refuses to answer from weak matches, flags citations to sections not in the corpus, and generates with Gemini and NVIDIA NIM with model fallback.
- Raised hit@10 on 300 unseen litigation questions from 0.03 (previous keyword search) to 0.30 on the live index; the gain held on a further 300 untuned questions, and 34 of 40 off-topic queries are refused.
- Firm data isolated by Postgres row-level security, verified by 37 cross-tenant attack probes; role-gated citation approval, rate limits, strict CSP and an SSRF allow-list.

Real-Time Credit Card Fraud Detection — github.com/CollinsNyatundo/creditcard-fraud

- Trained a LightGBM model on 284,807 transactions (0.17% fraud) using SMOTE plus undersampling, 79 features (50 engineered) and Optuna tuning that rejected configurations too slow to serve. All preprocessing ran on chronological splits, so no future data leaked into training.
- Time-ordered holdout: F1 0.80, precision 0.87, recall 0.75, ROC-AUC 0.98; calibration lifted PR-AUC from 0.74 to 0.77, reported with bootstrap confidence intervals and a power analysis.
- Model inference at **1.4 ms p95** in-process (10 ms target); served via FastAPI, Redis and PostgreSQL in Docker, with SHAP explanations.

Production AI Reference Template — github.com/CollinsNyatundo/production-ai-template

- FastAPI reference for multi-tenant LLM apps: identity-derived tenant scope, a step-limited ReAct tool loop, circuit breakers, retrieval with reranking, permission-gated tools and offline trajectory evaluation, with tests and security scans in CI.

Machine Learning Systems Reference MCP Server — github.com/CollinsNyatundo/machine-learning-systems-plugin

- A FastMCP server that makes an adapted edition of V. J. Reddi's open textbook *Machine Learning Systems* (CC BY-NC-SA 4.0) searchable from AI assistants: 7 read-only BM25 tools over SQLite FTS5, 5 resources and 4 prompts, with CI on Ubuntu and Windows.

GEO-AI Cropland Mapping Team Lead — cropland-mapping.streamlit.app

- Led a four-person team mapping smallholder farms in a Zindi food-security competition; a stacked ensemble (XGBoost, LightGBM, CatBoost, Random Forest, Vision Transformer) raised land-cover accuracy by 12%.

EDUCATION & CERTIFICATIONS

Bachelor of Business and Information Technology, JKUAT — 2019 – 2023

Certificates: Data Science, Moringa School (2025) · Data Engineering, Sankhyana (2025) · IBM Data Science (2024)