

Arushi Sagar

Ghaziabad, Uttar Pradesh, India | ✉ arushisagar18@gmail.com | ☎ +91-7488358904

🐙 GitHub | 🔗 LinkedIn

Education

KIET Deemed to be University

Expected Graduation: 2028

B.Tech in Computer Science & Engineering (AI & ML) | CGPA: 8.27

Technical Skills

Programming: Python, C++, SQL

Machine Learning: Supervised Learning, Feature Engineering, Model Evaluation, XGBoost, SHAP

Deep Learning & GenAI: PyTorch, Transformers, NLP, LangChain, LLM Applications, RAG

Data & Visualization: Pandas, NumPy, Plotly, Streamlit, Folium

Backend & APIs: FastAPI, REST APIs, Pydantic

Databases: MySQL, SQLite

Tools & Cloud: Git, GitHub, VS Code, Google Colab, Power BI, AWS

Projects

PipelineGuardian – LLM-Orchestrated ML Pipeline Auditor

Jul 2026

GitHub | Live Demo

- Built a notebook/script auditor catching data leakage, reproducibility gaps, and validation mismatches missed by standard linters, using AST-based static analysis paired with targeted LangChain/Groq Llama 3.3 70B reasoning for context-dependent checks.
- Shipped as CLI, FastAPI backend, and web UI; validated deterministic detectors against synthetic fixtures with automated true-negative test coverage.

Tech Stack: LangChain, FastAPI, Python AST, Pydantic, Pandas, Groq, Pytest, nbformat

Transformer from Scratch – Neural Machine Translation

Jun 2026

GitHub

- Implemented the original Transformer architecture in PyTorch from scratch – multi-head attention, positional encodings, masking, encoder-decoder stack – with no HuggingFace abstractions; achieved 35.25 BLEU (beam search, k=4) on Multi30k German-English translation.
- Ran a 12-configuration ablation study isolating attention heads, network depth, and FFN size; built attention visualization tooling to inspect cross-attention alignment across layers.

Tech Stack: PyTorch, NumPy, SacreBLEU, Multi30k

InfraUrban – Urban Heat Intelligence Platform

Jan 2026 – Feb 2026

GitHub | Live Demo

- Trained an XGBoost heat-risk model across 24 zones in 6 Indian cities using a time-based forward split (train 2021–24, test on unseen 2025) to prevent temporal leakage; achieved 61.9% RMSE improvement over a naive lag-1 baseline.
- Built local SHAP explainability per zone and a rule-based policy simulator projecting risk reduction under infrastructure interventions; deployed as a 4-page Streamlit dashboard.

Tech Stack: XGBoost, SHAP, SQLite, Streamlit, Folium, Plotly, Open-Meteo API

Problem Solving & Open Source

- Solved 100+ DSA problems on LeetCode spanning arrays, hashing, trees, graphs, and dynamic programming.
- Merged a feature PR to render-engine (Python static site generator), adding include/exclude glob filtering for static site map entries with test coverage; iterated through 4 rounds of maintainer review.

Certifications

AWS Certified Data Engineer – Associate | AWS Certified Machine Learning – Associate | AWS Certified Cloud Practitioner – Foundational