Project Proposal: Cloud-Native Auto-Scaling ML API (Free Deploy on Render)

Background: Machine learning inference endpoints are often deployed on fixed infrastructure, causing inefficiencies under variable load. Container orchestration and serverless platforms enable dynamic scaling; however, many academic exercises do not demonstrate fully automated, cost-free deployments suitable for students. Motivation: Demonstrate an end-to-end, zero-cost cloud deployment of an ML inference API to illustrate cloud-native scalability, logging, and CI/CD. This is useful for students and developers who want practical MLOps experience without cloud bills. Plan: Build a FastAPI inference service around a pre-trained RandomForest on the Iris dataset, containerize it, and deploy to Render's free web service. Validate auto-scaling by load testing and collect logs/metrics from the Render dashboard. Expectations: A live, public API endpoint, simple load-test evidence of scaling (cold-starts/scale-to-zero), and a one-page report describing architecture, testing, and lessons learned.