

ML Forge (MLFlow: Bridging Data to Deployment)

1. Project Description

Our project, **ML Forge**, introduces a unified Machine Learning Lifecycle platform — **MLFlow** — designed to streamline the process from raw data ingestion to model deployment. Enabling data scientists, analysts, and engineers to go from CSV to deployed API in minutes, not weeks.

ML Forge provides a single, intuitive, and fully automated pipeline.

It automates data cleaning, exploratory data analysis (EDA), model training, evaluation, and deployment through a seamless interface.

The main goal is to reduce complexity, improve data quality, and speed up the journey from data to actionable insights.

2. Key Features

- One-Click Data Ingestion: Supports CSV, Excel and JSON (Tabular formats)
- Intelligent Auto-Cleaning: Handles missing values, duplicates and outliers.
- Interactive EDA Dashboard: Pandas Profiling, Rich visualizations (Plotly), correlation heatmaps, distribution analysis, statistical summaries, usage of Gemini API to generate interactive charts and Interactive Chatbot.
- AutoML-Powered Model Training: Trains 7+ state-of-the-art models (including XGBoost, Random Forest, Neural Networks) with hyperparameter tuning (Random, Grid, Optuna) for different problems(classification, regression, clustering).
- Smart Model Selection: Automatically picks the best-performing model based on problem type and metrics
- Built-in Evaluation Suite: Accuracy, F1, RMSE, MSE, MAE, MAPE, R2-Score, Confusion Matrix, Classification Report, Cross-validation
- One-Click Model Deployment: Saves best model + preprocessing pipeline + FastAPI inference server
- Production-Ready API: Deployed model accessible via REST API with Swagger UI
- Full Reproducibility: All artifacts (model, encoder, feature names, metadata) saved and versioned by the help of MLFlow.

Why ML Forge? Traditional ML workflows are fragmented: → Data cleaning in Jupyter → EDA in another tool → Training in Colab → Deployment manually ML Forge unifies everything into a single Streamlit-powered interface with zero setup.

3. Target users

- Data Scientists & ML Engineers
- Business Analysts (non-coders)
- Startups & SMEs without dedicated MLOps teams
- Educational institutions teaching ML

4. Tech stack

Layer	Technology
Frontend	Streamlit
Backend	Python, FastAPI
Core ML	Scikit-learn, XGBoost, Optuna
Data Processing	Pandas, NumPy
Visualization	Plotly, Seaborn, Matplotlib

Deployment	Pickle + FastAPI + Uvicorn + Docker
Version Control	Git & GitHub
Project Management	Trello

5. Group Members & Roles

Member Name	Role / Responsibility
Mohamed Adel	Data Loading Function (responsible for reading and importing datasets in CSV, Excel, and JSON formats).
Eyad Sherif	Data Cleaning Function (handles missing values, duplicates, outliers, and feature preprocessing).
Ahmed Mohamed	Exploratory Data Analysis (EDA) Function (creates data visualizations and statistical insights).
Salim Mohamed	Evaluation Function (assesses model performance using metrics like accuracy, F1-score, and error rate).
Ahmed Akram	Deployment & UI Function (deploys the model as an API and builds a user-friendly interface).

6. Team Leader

Ahmed Akram — Project Manager and Evaluation Function Lead.
Responsible for overall coordination, task integration, and progress tracking.

7. Milestones & Deadlines

Milestone	Description	Deadline
M1: Planning & Research	Architecture & UI design	Week 1
M2: Data Module Completion	Develop and test data loading & cleaning modules.	Week 2
M3: EDA Visualization	Implement EDA and visualization functions.	Week 3
M4: Model Training	Create and test training pipeline.	Week 4
M5: Evaluation Metrics	Implement model evaluation metrics and testing.	Week 5
M6: Deployment & UI	Build API and user interface for deployment.	Week 6
M7: Integration & Testing	Combine all modules and finalize report.	Week 7

8. Project Status

Fully Functional MVP All 7 milestones completed successfully on time Live demo available at: <https://mlforge.streamlit.app>

9. Summary

This proposal presents **ML Forge**— an integrated ML lifecycle platform combining **data engineering, analysis, model training, evaluation, and deployment** in one unified workflow. Our modular design enables teamwork, efficiency, and scalability, ensuring a smoother path from **data to production**.