Python for Data Science 2: Guided Machine Learning Project

University: Tek-Up

Group: ING-4-SDIA

Tutor: Haythem Ghazouani

Academic Year 2025-2026

Contents

1 Introduction

This 7-week module is designed to immerse students in a professional Machine Learning workflow. Unlike introductory courses, this module focuses on the **end-to-end lifecycle**: from raw data acquisition to a production-ready containerized application.

Module Objective

To master advanced Python for Data Science by implementing a reproducible, scalable, and deployable Machine Learning pipeline integrating MLOps best practices.

2 Project Synopsis & Teams Choice

The **Bank Customer Churn** example used in the tutorials serves as a **Project Synopsis** (a comprehensive reference implementation).

Important: Dataset Validation

Students must work in teams. Each team is free to choose their own dataset and objective (e.g., Fraud Detection, Recommendation Systems, Energy Consumption Prediction).
The choice MUST be validated by the tutor before the end of Week 2.

3 Detailed 7-Week Roadmap

3.1 Phase 1: Foundations & Data (Week 1)

- ☐ **Week 1: Setup, Scraping & EDA:** Environment configuration, Git workflow, Web Scraping, and Exploratory Data Analysis.

3.2 Phase 2: The ML Pipeline (Weeks 2-3)

- ☐ **Week 2: Preprocessing & Feature Selection:** Building robust pipelines for encoding, scaling, and handling missing data.
- ☐ **Week 3: Advanced Modeling & MLflow:** Ensembling/Boosting algorithms and experiment tracking.

3.3 Phase 3: Integration & Deployment (Weeks 4-7)

- ☐ **Week 4: Backend (FastAPI):** Serializing models and exposing them via REST API endpoints.
- ☐ **Week 5: Frontend (React):** Creating a user-friendly interface to interact with the model.
- ☐ **Week 6: Containerization (Docker):** Orchestrating the full stack (API + Front) using Docker.

☐ **Week 7: Deployment & Final Review:** Final project submission and live demonstration.

4 Evaluation Criteria

Projects will be assessed based on the following:

1. **Code Quality:** Modularization, PEP8 compliance, and documentation.
2. **Experiment Tracking:** Proper use of MLflow to compare at least 3 model architectures.
3. **Robustness:** Handling edge cases in the API and Frontend.
4. **Reproducibility:** Successful deployment via `docker-compose up`.
5. **Creativity:** Originality of the chosen dataset and feature engineering.

5 Workspace Structure

Students must strictly adhere to the following directory structure:

- `/`: Root containing `README.md`, `.gitignore`, and `docker-compose.yml`.
- `cours/`: Theory notes and project guide (PDF).
- `tutos/`: Lab instructions (PDF).
- `code/`: Modular Python scripts (`train.py`, `app.py`, etc.).
- `data/`: Dataset samples (strictly no large files in Git).