

# Python for Data Science 2: Guided Machine Learning Project

University: Tek-Up  
Group: ING-4-SDIA

Tutor: Haythem Ghazouani

Academic Year 2025-2026

## 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 (SMOTE, GridSearchCV, Experiment Tracking, REST APIs).

## 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, Web Scraping (CNBC), and Exploratory Data Analysis.

- *Tutorial: `tutos/exploration_tuto.pdf`*

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

- ☐ **Week 2: Preprocessing & Imbalance:** Handling class imbalance (SMOTE) and automated hyperparameter tuning (GridSearchCV).
- ☐ **Week 3: Advanced Modeling & MLflow:** Experiment tracking for XGBoost and Random Forest.
- ☐ *Tutorial: `tutos/mlflow_tuto.pdf`*

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

- ☐ **Week 4: Backend (FastAPI):** Service monitoring (Health checks) and Batch prediction endpoints.
- ☐ *Tutorial: `tutos/fastapi_tuto.pdf`*
- ☐ **Week 5: Frontend (React):** Premium dashboard development.
- ☐ *Tutorial: `tutos/react_tuto.pdf`*
- ☐ **Week 6: Containerization (Docker):** Orchestration using Docker Compose.
- ☐ *Tutorial: `tutos/docker_tuto.pdf`*
- ☐ **Week 7: Final Review & CI/CD (Optional):** Project deployment.
- ☐ *Tutorial: `tutos/cicd_tuto.pdf`*

## 4 Evaluation Criteria

1. **Data Pipeline (20%):** Quality of scraping and EDA insights.
2. **ML Excellence (30%):** Correct use of SMOTE, Pipelines, and MLflow logging.
3. **API UI (30%):** Robustness of the FastAPI endpoints and the React user experience.
4. **Deployment (20%):** Successful execution via Docker.

## 5 Workspace Structure

Students must strictly adhere to the following directory structure:

- `code/`: Modular Python scripts (`modeling.py`, `app.py`, etc.).
- `frontend/`: React source code and Vite configuration.
- `tutos/`: Lab instructions in PDF format.
- `data/`: Serialized models and dataset samples.
- `Dockerfile.*`: Container definitions.

## 6 Resources

The reference code, tutorials, and configuration files are available at:  
<https://github.com/haythemghz/Python-for-Data-Science-Project>