



IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom



Data Science Toolkit

Lina Fahed

Laurent Brisson

1 - Learning Objectives

Providing you with all necessary elements to:

- ✓ **Understand** and **communicate** with Machine Learning Engineers and Data Scientists
- ✓ Successfully **interview** for positions in data science and ML engineering
- ✓ Obtain your **first role** in data science or ML engineering
- ✓ Build **production-ready** machine learning systems

Helping you discover:

- ✓ **Industry-standard** methodologies and concepts (CRISP-DM, MLOps)
- ✓ **Professional tools** used in companies (Python, scikit-learn, MLflow, Git)
- ✓ How to lead end-to-end **ML projects**
- ✓ Software engineering **best practices** for data science workflows

2 - Pedagogical Approach

No traditional lectures, but:

- ✓ Minimal **readings** (companion to "Fundamentals of Machine Learning" course where theory is covered)
- ✓ Intensive **hands-on workshops** focused on practical implementation
- ✓ **Real-world** case studies (air quality prediction, spam detection)
- ✓ Collaborative programming project with **Git workflows**

Assessment

✓ Projects:

- First project (Air Quality): Pair programming to implement a structured pipeline
- Second project (NLP): Group collaboration with independent problem-solving

✓ Knowledge assessment quiz: Individual evaluation at course completion

✓ Competencies & Grade calculation:

- Air Quality project validates competencies (max grade C)
- NLP project enables grades A-B based on quality
- Quiz may upgrade or downgrade final result

3 - Competency-Based Assessment

Three Core Competencies

Software Engineering Best Practices

- **BC01** - Apply professional development principles to data science workflows
- Ensure code quality, maintainability, and effective team collaboration

End-to-End ML Pipeline Implementation

- **BC05** - Design and implement complete, production-ready ML pipelines
- Use modern MLOps practices and industry-standard tools

Data-Driven Communication

- **BC02** - Transform technical analysis into actionable business insights
- Effective visualization and statistical analysis for stakeholders

4 - Course Structure

Three progressive sessions

Session 1: Data Analysis Fundamentals

- ✓ Python **environment setup** and **basic data manipulation** with Pandas
- ✓ **Data visualization** and **version control** with Git workflows
- ✓ Practical: Complete sales analysis project with professional documentation

Session 2: Production ML Pipelines

- ✓ Transform notebook code into structured, **testable pipeline architecture**
- ✓ **Hyperparameter optimization**
- ✓ **MLOps introduction** with MLflow experiment tracking

Session 3: Natural Language Processing

- ✓ **Text preprocessing** and **spam detection** across different domains (SMS vs Email)
- ✓ Production **NLP pipeline** with cross-domain performance analysis
- ✓ **Group Project**: Text classification system with comprehensive evaluation

5 - Resources

Where to start?

The course website provides comprehensive resources:

- ✓ Session activities with progressive weekly learning paths
- ✓ Practical tutorials covering environment setup and Git workflows
- ✓ Assessment guidelines explaining evaluation criteria and deliverables

The Moodle space will be used for submitting deliverables.

Prepare indicated content before attending sessions

- ✓ Session 1: Environment setup before the 3rd activity “Setup a Data Science project”
- ✓ Session 2: Read “Towards Production-Ready ML” before the 2nd activity
- ✓ Session 3: Read “NLP Lecture” before the 1st activity

