Module 01: Opening / ML Basics — Cheat Sheet

ML for Innovation Research — Prof. J. Osterrieder

Three ML Paradigms

Supervised Learning
Learn a mapping $f: X \to y$ from labelled data.
Examples: classification, regression.

Unsupervised Learning
Find structure in data *without* labels.
Examples: clustering, dimensionality reduction, topic modelling.

Generative AI
Models that generate new text, images, or code.
Examples: GPT-4, Claude, DALL-E.

Key Terms

Term	Meaning
Feature (x)	Input variable (column)
Label (y)	Target variable to predict
Training set	Data used to fit the model
Test set	Held-out data for evaluation
Overfitting	Model memorises noise
Hyperparameter	Setting chosen *before* training

The ML Workflow

1. **Research Question** — What do you want to learn or predict?
2. **Data Collection** — Surveys, APIs, databases, web scraping.
3. **Preprocessing** — Clean, encode, scale, handle missing values.
4. **Model Selection** — Choose algorithm suited to the task.
5. **Training & Tuning** — Fit model, tune hyperparameters.
6. **Evaluation** — Measure performance on unseen data.
7. **Interpretation** — Explain results in domain context.

Supervised vs. Unsupervised at a Glance

Supervised	**Unsupervised**
Needs labels	No labels needed
Predict outcomes	Discover patterns
Accuracy measurable	Quality is subjective
Classification, regression	Clustering, PCA, LDA

Python Libraries You Will Use

Library	Purpose
`pandas`	DataFrames, data wrangling
`scikit-learn`	ML models, metrics, pipelines
`matplotlib`	Static plots and charts
`seaborn`	Statistical visualisations
`nltk`	Tokenisation, sentiment, stopwords
`numpy`	Numerical arrays

Quick Starter Code

```
import pandas as pd
from sklearn.model_selection import train_test_split

df = pd.read_csv("data.csv")
X = df.drop(columns=["target"])
y = df["target"]
X_train, X_test, y_train, y_test = \
    train_test_split(X, y, test_size=0.2,
                     random_state=42)
```

Tip: Always set `random_state` for reproducibility.