Introduction to ML in Orange

MZ & JK

Introduction

What is ML

Why ML

Introduction

Problems ML can solve

Problem	How ML Helps
Disease diagnosis	Classifies samples (e.g., gene expression, medical images) as diseased or healthy.
Drug discovery	Predicts how molecules will interact, speeding up screening of drug candidates.
Gene function prediction	Identifies unknown gene roles based on expression patterns and networks.
Protein structure prediction	Predicts 3D structure of proteins (e.g., AlphaFold).
Genotype-to-phenotype prediction	Predicts traits like disease risk or plant yield from genetic data.
Image analysis	Counts cells, detects tumors, segments microscopy images.
Omics integration	Finds patterns across transcriptomics, proteomics, metabolomics data.

Introduction

- Know Your Task, Know Your Data

- What do I want to predict or classify?
- Is this a classification task (e.g., predicting blood type)?
- Or a regression task (e.g., predicting protein concentration)?
- Do I want to group data (clustering), or reduce complexity (PCA)?

Introduction

Why Orange?

Feature	Benefit
No coding required	Drag-and-drop interface is perfect for non-programmers.
Visual workflow	Helps you understand each ML step — from data import to model evaluation.
Great for teaching	Makes abstract concepts (like overfitting or classification trees) easier to grasp.
Add-ons for bioinformatics	Includes tools for gene expression analysis, clustering, and PCA.
Instant results	See predictions, accuracy, and visualizations immediately.

The workshop includes a little bit of theory followed by hands-on work

- Supervised learning
  - Classification and Regression
  - Generalization, Overfitting and Underfitting
  - Supervised ML algotithms
    - K-Nearest neighbour
    - Linear models
    - Decision trees

Introduction to Machine Learning

Supervised

- You train the model on labeled data — meaning you already know the correct answers.

Unsupervised

- The model doesn’t have labels — it tries to find patterns or structure on its own.

Classification vs. Regression Under Supervised Learning

Type	What It Does	Output Type	Example in Biotech
Classification	Sorts into categories	Discrete labels (e.g., yes/no, type A/B/C)	Is this sample tumor or normal?
Regression	Predicts numerical values	Continuous numbers	What is the protein concentration?

Clustering under unsupervised learning

- Group similar data points together,
- Based only on their features,
- Without knowing in advance what the “correct” groups should be.

Generalization, Overfitting and Underfitting

Goal: Find the sweet spot where the model is complex enough to learn the pattern but not too complex to memorize noise.

Concept	What Happens	Result on Test Data	Example
Underfitting	Too simple	Low accuracy	Using 1 gene to predict complex trait
Overfitting	Too complex	Low accuracy	Fitting all noise from small gene dataset
Good Generalization	Just right	High accuracy	Using 50 relevant genes, cross-validated

Summary

Core ML Concept

- What is a model
- Training and testing
- Generalization, Overfitting and Underfitting

  - *"A model is like a smart guesser. We show it many examples (training), and then ask it to guess something new (testing)."*

Supervised ML algotithms

- K-Nearest neighbour
- Linear models
- Decision trees

Decision tree