Hands-on Machine Learning in Life Science with R — Syllabus

Reza Belaghi

Ali Moazzami

Updated: 2025-08-18

Table of Contents

Welcome Message 2
Instructors 2
Course Overview 2
Prerequisites 2
Learning Outcomes 2
Teaching & Assessment Strategy 3
  Assessment Components 3
Weekly Plan (4 Weeks) 3
  Week 1 — Foundations, R Basics & Logistic Regression 3
  Week 2 — Trees & Ensembles (Merged Module) 3
  Week 3 — Margins & Kernels; Simple Neural Nets 3
  Week 4 — Integration & Presentations 4
Detailed Schedule (Sample Dates) 4
Assignment (Combined) 4
Final Project 5
Software & Packages 5
Policies 5
Appendix: Session Templates 5
  Lecture/Discussion (90–120 min) 5
  Lab (90–120 min) 5

Credits: **5 ECTS** • Subject: **Statistics and Machine Learning** • Level: **Third cycle (postgraduate)** • Delivery: **Online (Zoom: https://slu-se.zoom.us/j/2527490227)** • Marking scale: **Pass / Fail**

Welcome Message

Welcome to **Hands-on Machine Learning in Life Science with R**!
We are excited to have you join this intensive and applied course on statistical and machine learning methods tailored for biological, agricultural, and health data.

This course follows a **flipped classroom** format: you are expected to **review the lecture materials and short recordings before each live session**. Each lecture will be **recorded and made available** for later review.
Please come **prepared for discussion**, bring your **questions**, and engage actively during the synchronous sessions — our goal is to make the sessions collaborative, interactive, and practice-driven.

We look forward to learning with you!
— *Reza Belaghi & Ali Moazzami*

Instructors

- **Reza Belaghi** — reza.belaghi@slu.se

Course Overview

This course provides a hands-on approach to machine learning in life sciences, focusing on the practical application of key algorithms using R. Students will gain basic knowledge of supervised and unsupervised models and learn how to apply, evaluate, and interpret these models in real-world biological datasets.

Prerequisites

- Undergraduate degree in agronomy, biology, chemistry, or a related area
- Basic familiarity with R and introductory statistics

Learning Outcomes

By the end of this course, students will be able to:

- Explain the workflow of supervised and unsupervised machine learning for life-science data
- Implement logistic regression, decision trees, random forests, SVMs, and simple ANNs in R
- Evaluate and tune models (resampling, hyperparameters, performance metrics)
- Interpret results for scientific reporting in agricultural, biological, and medical contexts
- Communicate findings via a short, structured project presentation

Teaching & Assessment Strategy

- **Synchronous sessions:** 2 per week (Lecture/Discussion and Lab), ~1.5–2 h each
- **Assessment:** One combined assignment (supervised + unsupervised) and one final project presentation
- **Preparation:** Short pre-class video/readings for theory; live time is used for Q&A and hands-on labs

Assessment Components

Component	Weight	Due
Assignment	50%	End of Week 3
Final Project Presentation	50%	Week 4 (presentation day)

Pass criteria: Submission quality, code correctness/reproducibility, methodological justification, and clarity of interpretation.

Weekly Plan (4 Weeks)

Suggested dates align to Weeks 47–50 of 2025 (Nov 17 – Dec 12), maintaining the Monday/Friday rhythm.

Week 1 — Foundations, R Basics & Logistic Regression

- **Synchronous 1 (Lecture/Discussion):** Intro to R, data workflow and cleaning, missing data; overview of supervised vs unsupervised ML; logistic regression (model, interpretation, diagnostics).
- **Synchronous 2 (Lab):** Hands-on data cleaning, fitting logistic regression, prediction, and evaluation metrics (confusion matrix, ROC/AUC) in R.

Week 2 — Trees & Ensembles (Merged Module)

- **Synchronous 1 (Lecture/Discussion):** Decision Trees (splits, pruning) → Random Forests (bagging, OOB, importance).
- **Synchronous 2 (Lab):** Implement trees & RF in R; tuning with resampling.

Week 3 — Margins & Kernels; Simple Neural Nets

- **Synchronous 1 (Lecture/Discussion):** Support Vector Machines (margins, kernels, C/γ) and ANN basics (layers, activations, overfitting).
- **Synchronous 2 (Lab):** SVM classification in R (kernels, tuning); simple MLP (e.g., via `keras` or `nnet`).
- **Deliverable:** *Assignment due end of Week 3.*

Week 4 — Integration & Presentations

- **Synchronous 1 (Workshop):** Reception of the previous sessions and reproducible pipelines (`renv`, `targets`/`drake`), reporting, and error analysis.
- **Synchronous 2 (Presentations):** Final Project Presentations (20 min talk + 5 min Q&A per student/team).
- **Deliverable:** Slides + short report (2–4 pages) with methods, results, interpretation, and reproducible code repo.

Detailed Schedule (Sample Dates)

Week	Day	Date	Time	Type	Topic
1	Mon	2025-11-17	13:00–15:00	L1	ML workflow; metrics; unsupervised overview; Logistic Regression
1	Fri	2025-11-21	13:00–15:00	C1	Lab for L1: data cleaning, logistic regression, prediction & model evaluation
2	Mon	2025-11-24	13:00–15:00	L2	Trees → Random Forests theory
2	Fri	2025-11-28	13:00–15:00	C2	Trees/RF lab in R (tuning, importance)
3	Mon	2025-12-01	13:00–15:00	L3	SVMs + ANN basics
3	Fri	2025-12-05	13:00–15:00	C3	SVM + simple MLP lab
4	Mon	2025-12-08	13:00–15:00	L4	Repetition, Reproducibility; reporting; model critique
4	Fri	2025-12-12	09:00–15:00	C4	Final presentations

Assignment (Combined)

Goal: Demonstrate competence in both unsupervised and supervised pipelines using one life-science dataset (or two related datasets).

Data preparation & preprocessing: Preprocess data; handle missing data; train/test split; cross-validation; evaluate the fitted model.

Supervised: Choose at least two models among (Tree/RF, SVM, simple ANN). Define a clear target, conduct CV-based tuning, compare metrics (AUC/accuracy/F1, etc.), and explain trade-offs.

Reporting: Short, reproducible report (Rmd → HTML/PDF) with code appendix, figures, and clear interpretation for non-ML audiences.

Submission: R project folder with data dictionary, Rmd, and knitted output.

Final Project

Small open-ended project using a dataset of interest (agriculture, biology, or medicine). Present: problem, data, methods (at least two families), metrics, key findings, and limitations. Deliver slides and a brief write-up (2–4 pages).

Software & Packages

- **Core:** `tidyverse`, `tidymodels`, `rsample`, `yardstick`, `parsnip`, `workflows`
- **Models:** `rpart`, `randomForest` or `ranger`, `e1071` (SVM), `nnet` or `keras` (ANN)
- **EDA/plots:** `GGally`, `factoextra`, `cluster`, `pcaMethods`
- **Reproducibility:** `renv`, `targets/drake`, `janitor`, `here`, `styler`

Policies

- **Attendance:** Live sessions are interactive and required for a Pass.
- **Academic integrity:** Follow SLU and department guidelines. Plagiarism or unauthorized collaboration will not be accepted.
- **Accessibility:** If you need accommodations, contact the instructors early.

Appendix: Session Templates

Lecture/Discussion (90–120 min)

- 10 min: Agenda & framing
- 40–60 min: Key theory & demos
- 20–30 min: Q&A + checkpoint quiz/discussion
- 20–30 min: Mini-exercise or coding walk-through

Lab (90–120 min)

- 10 min: Setup & objectives
- 60–80 min: Guided coding + experimentation
- 10–20 min: Wrap-up; rubric-aligned checklist for deliverables