DLSS 2026 – Assignment 1

Predicting Household Wealth with Multilayer Perceptrons

Giordano De Marzo — 30.04.2026

Submission deadline: 20.05.2026 at 23:59

Overview

In this assignment you will use a Multilayer Perceptron (MLP) to tackle a regression task on real-world survey data. Your objective is to predict **household wealth**, specifically the household's total net income decile, from a set of socioeconomic, demographic, and lifestyle variables drawn from the **European Social Survey (ESS), Wave 11**. This is a complex, real life task, so do not expect perfect results. Your model should be able to correctly identify a trend, but imperfect predictions are expected.

The maximum number of points is **30**. Bonus tasks may award additional points, but the final score is capped at 30.

Dataset

A pre-filtered dataset containing the relevant variables is available for download from the course page on my website. The dataset contains one row per respondent. The **target variable** is *hinctnta*, which encodes the household's total net income as a decile (1 = lowest 10%, 10 = highest 10%).

To understand what each variable measures and how its values are coded, you can consult two resources:

- The **ESS codebook** (`codebook.html`), provided alongside the dataset on the course page. It documents every variable's question wording, response scale, and the numeric codes used for missing answers, refusals, and not-applicable cases.
- The **official ESS data portal**: `https://ess.sikt.no/en/datafile/242aaa39-3bbb-40f5-98bf-bfb1ce53d8ef`, where you can browse the full Wave 11 documentation.

Important – ESS missing-value codes: The ESS uses numeric sentinel values to encode non-responses, refusals, and not-applicable answers. These sentinels vary by variable; consult the codebook to identify the correct codes for each column. They are *not* real observations and must be handled during preprocessing. Rows where the *target* is missing should be dropped entirely.

Tasks

The assignment is worth **30 points** divided equally between the code and the written report.

Component	Points
Code	15
Task 1: Data import, exploration and preprocessing	5
Task 2: MLP building and training	5
Task 3: MLP evaluation and comparison with ML model	5
Report	15

Task 1 Data Import, Exploration and Preprocessing (5 code points)

- Load the dataset. Use the codebook to understand the variables and identify which values represent missing or non-applicable responses.
- Inspect the data: print its shape, examine the columns, and report the extent of missing values after sentinel replacement.
- Visualise the distribution of the target variable *hinctnta* and explore the features through appropriate plots and summary statistics.
- Preprocess the features appropriately. Handle missing values; a standard approach is to fill missing numerical values with the median and missing categorical values with the mode, though you are free to explore alternatives.
- Split the data into train, validation and test sets.

Important – ISCO-08 variables: The variables *isco08* (respondent's occupation) and *isco08p* (partner's occupation) are coded using the International Standard Classification of Occupations (ISCO-08), a four-digit hierarchical scheme. For this assignment, truncate both variables to their first two digits before encoding them, which corresponds to the sub-major occupational group and reduces dimensionality considerably. Note that *isco08p* will have substantial missingness for respondents without a partner or whose partner is not employed. Think carefully about how to handle this during preprocessing.

Task 2 MLP Building and Training (5 code points)

- Build and train **at least 3 MLP variants**. Vary at least two design choices across your models, for example the number of layers and neurons, the activation functions, the learning rate, the batch size, or the regularisation strategy.
- Use a loss function and output activation appropriate for a regression task.
- Track and plot the learning curves (training loss vs. validation loss) for each model.

Task 3 Model Evaluation and Comparison (5 code points)

- Evaluate all MLP variants on the validation set and select the best model. Justify your choice.
- Train a standard machine learning model (e.g. a Random Forest or a Gradient Boosting regressor) as a baseline.
- Evaluate both the best MLP and the baseline on the **test set**. Report appropriate metrics and include figures or tables to support your comparison. Show a comparison between the predicted and ground truth wealth.

Report (15 points)

Write a short scientific report of **at most 3 pages** (excluding bonus tasks). The report should be self-contained: a reader who has not seen this assignment sheet should be able to understand what you did and why. A possible structure is the following:

- **Title, Name and Matriculation Number**
- **Introduction:** briefly introduce the prediction task and your overall approach.
- **Results:**

  - *Data:* describe the dataset, the preprocessing decisions you made, and any relevant findings from your exploratory analysis.
  - *Models:* describe the architecture and training setup of each MLP and the baseline.
  - *Training:* show and discuss the learning curves.
  - *Evaluation:* compare the models on the test set using appropriate metrics and visualisations.
- **Conclusions:** summarise your findings and reflect on any limitations or open questions.

Bonus Tasks

Bonus tasks are optional. You may use up to an additional half page in your report for each bonus task you complete. The final grade is **capped at 30 points**.

Bonus Task 1: A Different Modelling Approach (2 bonus points)

The target variable *hinctnta* takes only 10 distinct integer values. This raises a genuine modelling question about whether regression is the most appropriate framing for this problem.

- Discuss the nature of the target variable and argue whether regression is the right approach, whether an alternative framing would be more suitable, or whether the answer depends on the use case.
- Train an MLP under your proposed alternative framing using an appropriate architecture, loss function, and output layer.
- Compare this model with your regression MLP. Discuss the trade-offs in terms of predictive performance, interpretability, and practical suitability.

Bonus Task 2: Model Ensemble (1 bonus points)

- Combine your trained MLPs into an ensemble. Choose a combination strategy appropriate to your task framing.
- Evaluate the ensemble on the test set and compare it to the best individual MLP and to the baseline. Discuss whether and why ensembling helps.

Submission Guidelines

- Submit your work on the course GitHub as a Colab notebook (`.ipynb`) and a PDF report. More information will be provided before the deadline.
- Your notebook must be clean, well-commented, and fully reproducible. Re-running it from top to bottom should reproduce all results.
- Use a fixed random seed wherever randomness is involved to ensure reproducibility.
- Include a `README.md` if needed.