Mark as done

Coursework Description

The coursework is a data science challenge. You will be presented with a tabular dataset in the form of a CSV file. Your assignment is as follows: (1) Train the best possible regression model to predict “outcome” (the first variable in the dataset) as a function of the remaining features. (2) Write a brief report on your data processing and model training/selection pipeline.

Model evaluation [50 points]

Objective: To attain the lowest possible generalization error on a held-out test set sampled from the same distribution as the training set.

You will be provided with an example script that implements a trivial baseline model. The binaries you submit will be loaded into this script, producing predictions on a test set of 1,000 samples.

Performance will be evaluated by out-of-sample accuracy. Top marks will go to any predictor that is within 10% of the Bayes optimal model (this is known, as the dataset is simulated).

Any model that successfully executes (i.e., produces predictions on the test set) gets 10 marks. Further points are awarded in accordance with model performance. Errors at runtime will result in deductions.

Report [50 points]

Objective: Write a two-page report in LaTeX using Overleaf that summarises your pipeline for data processing as well as model training/selection.

Subtasks:

1. **Exploratory data analysis [10 marks].** This includes visualization, cleaning, preprocessing, etc. It is possible that some features may need to be transformed or excluded.

2. **Model selection [10 marks].** Which algorithm(s) were selected for this task and why? The only way to be sure which methods work best is to compare them directly.

3. **Model training and evaluation [15 marks].** How does the model perform? What hyperparameters were chosen and why?

4. **Code supplement [15 marks].** In addition to the two-page report, you must include either an appendix or a link to a GitHub repository so that assessors can evaluate the implementation.

Assessment criteria include the clarity, accuracy, and creativity of the report. Late work will not be accepted without explicit approval.

Last modified: Thursday, 22 January 2026, 5:10 PM

◀ CW2 Late Submission - MCF approved only

Jump to...

CW1 – P3 training data ▶