Group Project Stage 3

Due: 11:59 PM on Sunday at the end of week 12 (Nov 2nd)

Value: 5% of Total Mark

Note: Get started your project ASAP. Discuss with your tutors and make use of Ed to ask questions.

1 Purpose

Stage 3 builds on the insights and collaboration developed in Stage 1&2. It introduces students to practical machine learning (ML) modelling and interpretation. Each group will design and compare different predictive models. Every member must include their own contribution in the report.
This stage emphasises:

- Model development using Python
- Evaluation and interpretation of predictive models
- Report of findings to an audience of peers with a data science background

2 Group Formation

- Groups must have 3–4 members and remain from Stage 1&2 unless changed with tutor approval.
- Any group may reuse, refine, or change their Stage 1&2 dataset.

3 Project Work for Stage 3

This assignment should cover the followings:

3.1 Common Prediction Target

- Choose one attribute from a dataset to predict.
- The attribute can be nominal or quantitative.
- The dataset may come from Stage 1&2 or be newly selected.
- Preliminary analysis should suggest the target is meaningfully predictable.

3.2 Evaluation metrics

- Select at least one metric to evaluate prediction accuracy.
- Justify why this metric is appropriate.
- For higher marks, more than one metric should be used.

3.3 Data preparation

- Split the dataset into training, validation, and test sets (e.g. 75/15/10).
- No part of the test set may be used in model training or tuning.

3.4 Develop Predictive Models

- Use Python (e.g. scikit-learn) to train at least 3 families of predictive models on the training set. Note that at least 3 families of models must be used to answer the same research question/topic.
- Apply any necessary pre-processing, if needed.
- Tune hyperparameters using only the validation data.
- Explain modelling decisions.

3.5 Model Evaluation

- Evaluate each model's performance on the test set.
- Report at least one metric (more for higher marks)
- Interpret the performance and reflect on effectiveness or limitations.

3.6 Conclusion

- A comparison of the different approaches used
- Observations on which models worked better and why
- Any dataset-related insights or modelling challenges
- Reflections on metric choice and model limitations

4 Stage 3 Submission

- **Submit on Canvas:**
  - **One PDF** containing the group report.
- **Upload a compressed `.zip` archive** with the following structure:
  - **Predictive model subfolders** (one per model), each containing:
    1. Python code used to train and evaluate the model.
    2. Any cleaned datasets and/or transformation artifacts.
  - **Shared content** folder (if applicable) for common resources.

Only one student per group needs to submit both files.
Submission Links:
Report(.pdf) submission: [Report submission 1902]
Code & Data(.zip) submission: `[Code & Data submission 1902]`

5 Marking

There are three components to be assessed. All are group-marked, and each member will receive the same score unless specific differences in contributions are explicitly noted.

Component	Weight (%)	Full Marks Criteria
Predictive modelling (max 4 pages)	40%	1) Choose at least 3 different families of predictive models. 2) Clearly explain the model training process using Python (e.g. scikit-learn). 3) Show explicit code in the report. 4) Demonstrate appropriate preprocessing and/or hyperparameter tuning. 5) Use at least 3 families of predictive models. 6) For full marks: justify why the chosen modelling approach suits the task.

Model Evaluation (metric reporting + insight) (max 2 pages)	35%	1) Compute at least one suitable evaluation metric per model on test data. 2) Clearly show the metric computation code. 3) For full marks: interpret results (e.g., overfitting/underfitting) and relate outcomes to dataset characteristics and model choices. Multiple metrics expected for higher marks.
Conclusion (max 1 page)	25%	1) Summary of findings in relation to your research topic. 2) Clearly communicates relative strengths, weaknesses, and trade-offs of different models. 3) For full marks: Report includes thoughtful discussion of modelling limitations, metric limitations, and reflects a shared understanding of machine learning principles. 5) Clear presentation

6 Late Submission

As announced in the unit outline, late work (without approved special consideration or arrangements) suffers a penalty of 5% of the maximum marks, for each calendar day after the due date. That is, we subtract 0.25 marks per day from what you would otherwise get for the work. No late work will be accepted more than 10 calendar days after the due date.