Coursework 1 Brief

Overview

Coursework 1 is designed to develop your practical skills in machine learning through structured, comparative experiments on a selected benchmark dataset using **scikit-learn**. You will demonstrate your understandings of dataset investigation, analysis design, performance evaluation, and overfitting prevention while focussing on **experiment design and justification rather than raw accuracy**.

This coursework series prioritises:

- Rigorous **experiment design**
- Clear **justification of modelling choices**
- Meaningful **performance analysis** over raw accuracy

Task

Apply either **linear regression** or **logistic regression** to your chosen benchmark dataset.

Submission

1. Report (maximum 5 pages) that explains the pipeline including
    - Dataset exploration & pre-processing
    - Model application with training and validation
    - Accuracy & computational efficiency analysis
    - Visualisation and communication of results

    NB The report must use the **provided template**.
2. Jupyter notebook with
    - Clarity and reproducibility.

Submission Format

- All reports must be submitted in PDF, using the **provided template**.
- Each report must be accompanied by a **Jupyter Notebook**, showing your workflow and outputs clearly.
- Ensure your notebooks are clean and reproducible.
- Individual submissions need to be titled “ENG4200_CW<#>_<student ID>.pdf” for report and “ENG4200_CW<#>_<student ID>.ipynb” for Jupyter notebook

Assessment Criteria

	Report				Jupyter notebook
Grade Band	**1. Dataset Investigation & Pre-processing (20%)**	**2. Model Application with Training and Validation (20%)**	**3. Accuracy & Computational efficiency Analysis (20%)**	**4. Visualisation & Communication of Results (20%)**	**5. Clarity & Reproducibility (20%)**
A Excellent	Comprehensive and insightful data exploration, identifying key trends, distributions, and potential issues. Clear rationale for preprocessing decisions, including effective handling of missing, noisy, or imbalanced data.	Model is applied correctly with a well-structured training and validation pipeline (e.g. appropriate train/test split or cross-validation). Demonstrates strong understanding of the model's assumptions and limitations, with evidence of thoughtful experimentation.	Evaluation includes relevant and well-justified accuracy metrics. Analysis reflects deep understanding of computational efficiency (e.g. runtime, memory usage, scalability). Critical reflections are supported by empirical evidence (e.g. performance on varying data sizes or tuning runs).	Visualisations are clear, purposeful, and effectively support the narrative (e.g. coefficient plots, learning curves, confusion matrices). Results are communicated with clarity and precision, with insightful interpretation in the context of the problem.	Excellent clarity and structure; notebook is fully reproducible, clean, and professionally documented.
B Very Good	Very good data exploration that identifies most key trends and potential issues. Preprocessing decisions are appropriate and clearly explained, including reasonable treatment of missing, noisy, or imbalanced data.	Model is applied correctly with a sound training and validation pipeline (e.g. appropriate split or CV used). Very good understanding of model assumptions and tuning, with evidence of thoughtful experimentation.	Evaluation uses appropriate accuracy metrics, with clear justification. Some attention is paid to computational performance, including basic runtime analysis. Very good reflections are provided with a minor lack of depth or empirical support.	Visualisations are generally clear and relevant, supporting most key points in the analysis. Results are communicated with clarity, with minor lack of context and scope.	Very good clarity and structure; notebook is reproducible and well-documented with minor issues.
C Good	Basic data exploration is present and good, identifying some relevant trends or issues, but lacks depth or completeness. Pre-processing steps are applied but may be incomplete or insufficiently justified. Shows good understanding but limited critical analysis.	The model is implemented and trained, but the training/validation process may be overly simplistic, missing important steps like proper splitting or cross-validation. Demonstrates a good understanding of the model, but with limited explanation or minimal experimentation.	Evaluation includes at least one relevant accuracy metric, but justification may be weak. Computational efficiency is mentioned, but not clearly analysed or supported by evidence. Analysis remains mostly descriptive with limited critical reflection.	Visualisations are present and in a good quality but in some cases lack clarity or some aspects. Communication is understandable but sometimes lack structure, depth, or interpretive insight.	Good clarity and structure; notebook is functional and generally reproducible.

	Report				Jupyter notebook
Grade Band	**1. Dataset Investigation & Pre-processing (20%)**	**2. Model Application with Training and Validation (20%)**	**3. Accuracy & Computational efficiency Analysis (20%)**	**4. Visualisation & Communication of Results (20%)**	**5. Clarity & Reproducibility (20%)**
D Satisfactory	Minimal data exploration is present, with only superficial analysis or missing key observations. Pre-processing steps are incomplete, inappropriate, or poorly explained. Shows limited understanding of how data preparation affects model performance.	Model is implemented but with significant issues (e.g. training on test set, lack of validation). Some core concepts (e.g. model assumptions, parameter settings) are misunderstood. Little to no experimentation is performed.	Accuracy metrics are used but not well justified or interpreted. No meaningful attempt is made to assess computational efficiency. Evaluation is shallow and lacks evidence, relying mostly on defaults or assumptions.	Visualisations are unclear, irrelevant, or missing. Communication is confusing or overly simplistic, with poor organisation and limited explanation of results. Interpretation is descriptive or incorrect.	Satisfactory clarity; notebook is functional but lacks structure and consistent documentation.
E Weak	Little or no evidence of data understanding or preparation. Major issues in handling of missing/noisy data. Approach lacks rationale and may reflect serious conceptual gaps.	Model is applied in a fundamentally incorrect or incomplete manner. No proper training/validation is demonstrated. Critical errors show a lack of basic understanding of model use.	Evaluation is either missing or incorrectly applied. No justification or analysis is provided. Little to no investigation of computational aspects. Some evaluation appears flawed.	Report is poorly written, disorganised, or lacks coherence. Visualisations are mostly absent or meaningless. Communication fails to convey key ideas or model findings.	Weak clarity; notebook is unclear, incomplete, or non-functional.