CS105 Group Project Guideline

1 Objective

The group project is intended to

- give you an opportunity to work on a real-world dataset;
- let you apply the theoretical and practical concepts covered in class (and potentially beyond via independent readings and research);
- promote collaboration between team members to mimic business settings.

2 Summary of Milestones

The project component accounts for 20% of the final grade. Further breakdown and a list of major milestones is given below.

Milestone	Weight	Deadline	Deliverable	Description
Team formation	-	29 Jan 2023	eLearn signup	4-5 members
Part I: EDA	8%	19 Mar 2023	Jupyter Notebook	Answers to the issues outlined in each part, with reproducible and well-documented codes
Part II: Modeling	12%	14 Apr 2023		
Peer evaluation	-	15 Apr 2023	Survey	Used for moderation of scores

2 Datasets / Plagiarism Warning

You should work on **only one of the datasets** from the list below. These datasets and their descriptions are available on eLearn in the Project folder.

1) Climate dataset
2) Credit dataset
3) Employee dataset

While these datasets are derived from publicly available sources, we have modified parts of the data for the purpose of the project. That means you should not search for and refer to studies on these datasets online, as the datasets we released here are different versions from what are publicly available.

Any plagiarism suspected will be investigated to the fullest extent, and if confirmed, will be subject to disciplinary actions. Sanctions may include getting zero for the project, failing the course, suspension and expulsion, depending on the nature, severity and circumstances of the plagiarism.

3 Project Content

Your project is divided into two parts: Exploratory Data Analysis (EDA) and Modeling.

Part I: EDA

Submit a single *Jupyter Notebook* based on the given template, which must contain

- Point-by-point answers of the following issues;
- If applicable, corresponding codes that are reproducible (i.e., they produce output consistent with your answers), and well documented (in the form of comments and markdown cells).

You need to address the following issues. The percentages below add up to 100%, but they contribute to 8% of the overall grade of the course.

1. Overview of dataset [15%]
    a. Summarize the background of the dataset.
    b. State the size of the dataset.
    c. For each variable, describe what it represents and its data type (numerical or categorical).
2. Data pre-processing [35%]
    a. For each variable, determine the percentage of missing data. For any column with missing data, describe how you resolve the issue. Clearly state any assumption you made.
    b. For each variable, identify outliers (if any) and describe how you resolve the issue. Clearly state any assumption you made.
    c. For categorical variables, perform the necessary encoding.
3. Exploratory analysis and visualization [50%]
    a. For each variable, provide relevant summary statistics.
    b. For each variable, provide an appropriate visualisation depicting the distribution of its values, and summarize any key observation(s) you made.
    c. Perform bi-variate analyses on the variables. You do not need to analyse every pair; only focus on the pairs you believe are worth investigating and explain your choices. For each pair, describe the relationship between the two variables. Use appropriate statistical methods and/or visualization.

Part II: Modeling

Submit a single *Jupyter Notebook* based on the given template, which must contain

- **Include Part I on EDA as well** to produce a self-contained notebook. You may fine-tune Part I findings and/or codes based on feedback received on Part I submission;
- Point-by-point answers of the following issues;
- If applicable, corresponding codes that are reproducible (i.e., they produce output consistent with your answers), and well documented (in the form of comments and markdown cells).

You need to address the following issues.

1. Problem formulation
    a. Formulate one regression problem and one classification based on the dataset.
    b. State which problem (regression or classification) you would be investigating and why.
    c. Clearly specify the dependent variable you are predicting, and its significance.
2. Model training
    a. Describe the steps taken to split the dataset into train and test sets
    b. State the model(s) you will train on, and explain your choice(s). Please limit yourself to no more than **three** models—Grading is based on the validity and soundness of your model, rather than the quantity.
    c. For each model, perform the training, and report the trained parameters and the training scores, if applicable.
3. Model evaluation and selection
    a. For each model, predict the response variable on the test set.
    b. Describe the metric you use to evaluate your model(s). Report the test scores for each model.
    c. If you trained more than one model, identify the final model you would choose for the prediction task, and explain your choice.
4. Findings and conclusion
    a. Interpret what your model is implying, and summarize any insight you have drawn from the project. Explain if it is consistent with intuition, and if not, provide a plausible justification.
    b. Share any lesson you have learned from the project.
5. Team contribution
    a. Describe the contribution of each member, including both the tangible (e.g., implementation and writing) and intangible (e.g. generating ideas, planning).
6. References
    a. List any sources you have cited.

4 Grading criteria

Your project will be graded on the key factors listed below.

- Completeness in answering most, if not all, points.
- Validity and soundness of your approach, explanation and interpretation.
- Code reproducibility.
- Documentation and presentation.

5 Peer Evaluation

At the end of the term, a peer evaluation will be conducted via eLearn. In the event of any unusual peer evaluation (i.e., a member receiving very low score with detailed comments supplied), the instructors reserve the right to conduct further investigation and moderate the scores accordingly based on the investigation, if necessary. Your evaluation will be anonymous and confidential to all other team members or students, although the teaching team will be able to see your identity in order to conduct any investigation and/or moderation. More details will be released towards the end of the term.