DSAIT4000 Data Management and Engineering Group Assignment 1

September 5, 2025

Jie Yang, Rihan Hai

This is the first assignments for the DSAIT4000 Data Management and Engineering. The groups will be made up of **5 students**, made on Brightspace by the students. The deadline will be **01/10/2024 23:59:59 sharp**.
More detailed descriptions can be seen in the available notebook.
The tasks can be done in the notebook.
You must submit :

- **A PDF containing all of your discussions.**
- **The executable code that was written to complete the assignment.**

---

1. Part 1 The adult dataset is provided.
   (a) Load the dataset.
   (b) Clean the dataset, remove null values or empty cells
   (c) Choose 4 classifiers to classify the data against the target variable
   (d) Validate the model using various methods
   (e) Create several copies of the original dataset by changing the target variable's proportions.
   (f) Evaluate the effects of the perturbations on the models.
   (g) Discuss how you would reduce the impact of wrongly labeled data or correct wrong labels.

2. Part 2

   Your group will receive a large set of datasets. All of these datasets are in raw form, which forms a data swamp. *Note: All groups will have different data swamps.* A data swamp is not very useful. dataset

   (a) In the lecture, you would have been taught various data discovery methods and metrics. Please implement two methods for each relation as defined above.

   Note (not graded): You may experiment to verify implementation correctness for a sanity check. Evaluate data discovery results on the adult dataset by creating your own partitions to re-integrate the datasets.

(b) Report the results of the algorithms in the following form: "Dataset 1, Dataset 2, Relation". Compare the results as needed or report any issues you encounter during the process.

As a consequence of running the algorithms on the results, there may have been a lack of relations; or possibly no relations were discovered., If any of you have sampled the datasets and reviewed them, you may have noticed some data quality issues.

(c) In conclusion, some cleaning is needed. The errors can exist anywhere, so you may need to do some exploration. Clean the datasets to the best of your knowledge and then re-run the discovery algorithms. Also, report the relations that you have been able to discover. It would be great if you could highlight the difference between the results before and after cleaning.

(d) After doing the cleaning and discovery, you now have some new results. What have you found to be the case? Are the results consistent between the different methods? Which cleaning method or error caused the most data problems? Write a short report on the impact of data quality on the results of data discovery. (Max 400 words)