DSAIT4000 Data Management and Engineering Group Assignment 1

September 13, 2024

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

   Discuss how you would design a crowd-sourcing task to perform annotations for sentiment analysis. The dataset for annotations is available on brightspace.

3. Part 2 The adult dataset has been split into several datasets. The splits can thus be combined to form the original adult dataset with some extra data columns.
   (a) Implement a function that will enable you to discover which datasets are related to each other.
   (b) Put the different portions of the dataset into the models used in the previous part.
   (c) Combine the data based on these relations found in part (a)
   (d) Select a testing dataset from the full adult dataset from Part 1.
   (e) Combine the datasets through random combinations of the datasets.
   (f) Fit the models chosen from Part 1, with the different combinations of datasets formed in part 2(c) and 2(e)
   (g) Discuss the results of the different combinations.
4. Part 3 From Part 1 and 2, you will have experienced different types of data quality issues. You will also experience data quality issues in various settings. Select one of the scenarios and discuss the effects of data quality and how you would identify and solve the data quality problems.