LE/EECS 4412 – Data Mining

Winter 2022 – Section M

Course Project

Submission Deadline: January 28, 2022, before 23:59

Part 1: Project Introduction

Objectives

The purpose of this activity is to introduce the topic, team, and initial scope of your project.

Submission Requirements

- Please submit your results/output in a PDF file. Submit your code in a text (.txt) file.
- The name of the files must be your YorkU student number(s) such as `100131001-100131002.pdf`.
- File must be uploaded to the Phase1 submission link provided. The name of submission on the link must also be the same as your file name above.
- Your diagrams, if applicable, can be hand drawn or digital. In both cases, the images must be clear.
- If the files do not open properly or the content is not clear, then you will be awarded zero.
- **Deadline is January 28, 2022 before 23:59. Late submission is permitted with 10% penalty per day only up to 7 days after the original deadline.**
- **Your submissions will be verified using Turnitin (or some other suitable tool) for originality. 40% or more similarity will be awarded zero in the assignment and reported to the department. We may report similarity less than 40% if it is of significant nature.**

Warning: Please add the necessary headings and labels to your report so that the TAs can understand the different parts properly. Anything that we cannot understand will be awarded zero.

General Instructions

- **Please add the statement given in Appendix A in the start of your report. Each team member must add her own statement.**
- **Please make sure that your document is easy to understand; clearly add the question number, part number, captions and foot notes wherever required. If the TA can't locate the answer, then it is your responsibility.**
- **Make sure that images/screenshots are clear. If the image is big then split it into multiple parts. Clearly write their purpose. You can add multiple images even if the question statement doesn't say so to make sure that your answer is easy to comprehend.**
- **Highlight the significant parts of each image so that TA can easily identify the required answer.**
- **Add necessary explanation to make sure that TA can understand different parts of your document.**

Task 1: Team Description

The project in this course is a group project. The recommended size of the team is 3 students but in case of non-availability, 2 member teams are also acceptable. However, there will be no extra credit or work compensation for a 2-member team.

Please provide the names, student IDs and eClass emails of all the group members. All members must be reachable through the email address provided.

Task 2: Problem or Dataset Identification

In order to practice the concepts and techniques that we will learn in this course, each team will select a dataset. Each team is free to select any data set with following restrictions:

- Each team must work on a different data set.
- There must be at least 3 entity or object types.
- Each entity/object type must have at least 3 attributes/dimensions.
- There must be at least 25 instances per each entity/object type.

Task 3: Describe the Dataset

Please provide a detailed description of your dataset. Following is a minimum list of items that must be in your description:

- Source, purpose and problem domain/background
- Number/count of Entity/Object types, their dimensions, their instances etc.
- Purpose or semantics of each entity type, attribute etc.
- Nature/type of each entity type, attribute etc.

Task 4: Purpose or Scope of the Project

Each team is required to apply Data Mining techniques to mine useful information from the Dataset they have selected. Describe at least 5 possible questions that you want to answer in this project or 5 useful information that can be discovered. Certainly, it is early to be precise and you may eventually reshape/redefine the questions but there should be an initial target in the minds of the team. Clearly explain with the help of suitable examples what your anticipated targets are.

Appendix A

The following statement must be added in the beginning of your report. Each team member must submit her own independent statement. The signature can be electronic or you can add a scan of statement with handwritten signature.

I ***student_name*** student ID # ***student_id*** acknowledge that I have contributed at least 30% time and effort to the preparation of this report and work discussed herein.

Student_Signature

Appendix B: Dataset Resources

Below are some popular sources to look for the datasets:

- http://kdd.ics.uci.edu/
- http://archive.ics.uci.edu/ml/index.php
- http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html
- https://cds.cdm.depaul.edu/resources/datasets/
- https://www.cs.cmu.edu/~awm/15781/project/data.html
- https://www.springboard.com/blog/data-science/free-public-data-sets-data-science-project/