LE/EECS 4412 – Data Mining

Winter 2022 – Section M

Course Project

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

Part 2: Phase 1

Objectives

The purpose of this activity is to understand the basic characteristics of the data set and apply some fundamental data mining techniques on the data.

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 Part2 submission link provided. The name of submission on the link must also be the same as your file name above.
- **If you use any tools such as Python, R, Excel etc., then upload all related material (code, scripts, worksheets) as a single file zip file to the submission link. Submission of this material is compulsory to get the grade in this part.** Name this file as `100131001-100131002.zip`.
- 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 February 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 task 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 information.**
- **Add necessary explanation to make sure that TA can understand different parts of your document.**

Task 1: Describing the Data

- Use a suitable graph/chart/visualization technique to describe your data set. You must describe all (or at least 5 object types and 5 dimensions per object type if the object types and dimensions are more than 5) object types and their dimensions in their original form (without any preprocessing). Multiple object types or dimensions can be combined in one graph/chart/visualization if that makes sense.
- Give a brief justification to support each of the description technique that you have selected.

Task 2: Basic Statistical Analysis

- From your data set, select one dimension of each kind: Nominal, Ordinal, Interval, Ratio.
- For each selected dimension:
  - perform two basic statistical operations/tests and
  - describe the results using a suitable technique
  - describe any data preprocessing that you have performed
  - submit the data for these four dimensions in the form of CSV file; name it as `100131001-100131002—T2.csv`.

Task 3: Standardization and Normalization

- Select two dimensions, one Interval and one Ratio; dimensions from Task2 may be reused.
- For each selected dimension, perform:
  - Z-score standardization and
  - Min-Max normalization
- For each selected dimension:
  - describe the two modified versions (Z-score and Min-Max) along with the original version using a suitable technique to observe/indicate if there is any change in the nature/trend/behavior.
  - Comment whether a technique is better than the other or not, with justification.
- Submit the data for the two dimensions, original and modified, in the form of CSV file; name it as `100131001-100131002—T3.csv`.

Task 4: Principal Component Analysis

- From your data set select the object type with most dimensions. If the number of dimensions is more than 10 in this object type, then you many select any 10 dimensions and leave the additional.
- Using a suitable tool/language, perform the Principal Component Analysis and select two best PCs.
- Use a suitable graph/chart/visualization technique to describe the two PCs.
- Describe any data preprocessing that you require for this task.
- Submit the data for the two PCs in the form of CSV file: name it as `100131001-100131002—T4.csv`.

Task 5: Similarity Measurement

- From your data set select the object type with most numeric dimensions. For this task we will use only numeric dimensions of that object type; so, ignore the other dimensions. If the number of

numeric dimensions is more than 5 in this object type, then you many select any 5 numeric dimensions and leave the additional.
    - Submit your filtered data, the data for the selected dimensions, in the form of a CSV file: name it as `100131001-100131002—T5Data.csv`.
- Compute the Euclidian distance of each object from the other. In your report show the distance of 10 randomly selected objects from each other in a table.
- Compute the Cosine distance of each object from the other. In your report show the distance of 10 randomly selected objects from each other in a table.
- Compute the Mahalonobis distance of each object from the other. In your report show the distance of 10 randomly selected objects from each other in a table.
- Our objective is to draw the three distances using line graph to compare their behavior. Explain why or why not the inter-observation distances computed above are suitable for this comparison.
- Submit the three distance matrices in the form of three CSV files: name them as `100131001-100131002—T5EU.csv`, `100131001-100131002—T5CO.csv`, `100131001-100131002—T5MA.csv`.

Task 6: Classification

Classification is a supervised process and hence require labeling.

- Using a distance approach of your choice from Task5, split your data observations (only for the object type that you used in Task5) into at least three classes and label them. Clearly describe any preprocessing performed, your labeling process and rationale for your actions.
    - Submit your data, the selected dimensions and the class label, in the form of a CSV file: name them as `100131001-100131002—T6Data.csv`.
- Divide your observations into two groups, training, and test.
- Create a decision tree classifier for your object type. Use this classifier on your test data observations and describe its performance and accuracy.
- Create a rule-based classifier for the same object type. Use this classifier on your test data observations and describe its performance and accuracy.
- Compare and comment on the performance of the two classification models for your data.

Task 7: Progress on your Objectives

- Is/are any of your initial five questions is/are or may be addressed at this point? Describe.
- Do you need to restructure or reformulate your questions? If yes, then provide the updated set of problems. If no, then explain why not.

Selection of tools

- To perform the tasks described in this part, you are free to use any languages and tools of your choice and comfort.
- Any source codes and tool specific documents must be provided with the submission document.
- Any installation and usage instructions must also be provided in the document.

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