SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

ANL305
Association and Clustering

Group-based Assignment

July 2024 Semester

GROUP-BASED ASSIGNMENT

This assignment is worth 20% of the final mark for Association and Clustering.

The cut-off date for this assignment is **22 October 2024, 2355hrs.**

This is a group-based assignment. You should form a group of **4 members** from your seminar group. Each group is required to upload a single report via your respective seminar group site in Canvas. Please elect a group leader. The responsibility of the group leader is to upload the report on behalf of the group. Those submitting individually will be given a 10 marks deduction.

It is important for each group member to contribute substantially to the final submitted work. All group members are equally responsible for the entire submitted assignment. If you feel that the work distribution is inequitable to either yourself or your group mates, please highlight this to your instructor as soon as possible. Your instructor will then investigate and decide on any action that needs to be taken. It is not necessary for all group members to be awarded the same mark.

Up to 25 marks of penalties will be imposed for inappropriate or poor paraphrasing. For serious cases, they will be investigated by the examination department. More information on effective paraphrasing strategies can be found on https://academicguides.waldenu.edu/writingcenter/evidence/paraphrase/effective.

Note to Students:

You are to include the following particulars in your submission: Course Code, Title of the GBA, SUSS PI No., Your Name, and Submission Date.

For this GBA, it is mandatory that questions are not divided among group members. Each member must independently address and work on a question before engaging in group discussions of the question. In the event of a peer evaluation, each member is expected to submit their own original answers along with justifications for their individual contributions.

All peer evaluation requests must be submitted to the school at least three working days before the GBA due date. Late requests will not be considered.

Please keep a copy of your report before submission, and check whether your submission has been uploaded onto Canvas successfully (you can check if the Turnitin system can read your submission). **Failed or unreadable submissions will be marked as zero (0)**.

Use of Generative AI Tools (Allowed)

The use of generative AI tools is allowed for this assignment.

- You are expected to provide proper attribution if you use generative AI tools while completing the assignment, including appropriate and discipline-specific citation, a table detailing the name of the AI tool used, the approach to using the tool (e.g. what prompts were used), the full output provided by the tool, and which part of the output was adapted for the assignment;

- To take note of section 3, paragraph 3.2 and section 5.2, paragraph 2A.1 (Viva Voce) of the Student Handbook;

- The University has the right to exercise the viva voce option to determine the authorship of a student's submission should there be reasonable grounds to suspect that the submission may not be fully the student's own work.

- For more details on academic integrity and guidance on responsible use of generative AI tools in assignments, please refer to the TLC website for more details;

- The University will continue to review the use of generative AI tools based on feedback and in light of developments in AI and related technologies.

This assignment is an application project that is open-ended and there is no fixed answer. You are to select one specific topic from the ***four (4)*** candidate datasets provided below. Identify a potential business problem and solve it using association analysis and (or) clustering techniques. You should choose the business problem that is interesting and possible to analyse in a meaningful manner. Write a report to present your study and findings.

Four candidate datasets:

1. **Bike-Share Demand**

   Rental bikes have recently been introduced in many urban cities to improve last mile mobility and comfort. Ensuring that rental bikes are available and accessible to the public at the right time is crucial for reducing waiting times. Consequently, maintaining a stable supply of rental bikes is a major concern. Understanding the demand for bikes at different times of the day is essential for achieving this stability.

   The dataset includes weather information, the number of bikes rented per hour, and date information. You can also give reasonable assumptions and preparation on the dataset to facilitate your study.

   Data source: https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand

2. **Click Through Rate**

   Many organizations rely on email campaigns for effective user communication and product promotion. Email campaigns, featuring various Calls To Action (CTAs), aim to maximize the Click Through Rate (CTR). CTR is a key success metric for online campaigns, which is calculated by dividing the number of clicks on CTAs by the total email recipients. A higher CTR indicates a more successful campaign.

   This dataset provides the information of over 1800 email campaigns with detailed CTR values. You can give reasonable assumptions and preparation on the dataset to facilitate your study. More details about the dataset can be found via:

   Data source: (Note: use the train dataset for your project)

   https://www.kaggle.com/datasets/shibumohapatra/clickthrough-rate-ctr/data?select=train_F3fUq2S.csv

3. **Online Learning Review**

   COVID-19 significantly impacted education, shifting from traditional classrooms to online learning formats, bringing flexibility and convenience to some extent. Even in the post-pandemic era, many students still choose online learning, which has forced educational institutions to adapt quickly, embracing digital platforms and proposing new formats like hybrid learning. However, online learning may not be suitable for all learners.

   This data collection aims to explore learners' satisfaction with online learning through a review. By understanding the factors that impact effective online learning, learners and educators can make informed decisions about choosing the appropriate learning format and enhancing the quality and effectiveness of online learning in the future. More details about the dataset can be found via:

Data source:

https://www.kaggle.com/datasets/sujaradha/online-education-system-review

4. **Environmental Air Pollution**

Environmental air pollution involves the presence of harmful substances in the atmosphere, adversely affecting human health and ecosystems. Many countries have made significant efforts to understand the cause factors and combat pollution.

This dataset introduces large scale spatial-temporal data involving the major actors in urban air pollution. It combines multiple sources for obtaining the information essential for studying urban air pollution - the pollutants themselves, traffic in the city, pollution from power generation industries and meteorological factors. It is a rich dataset covers a daily level information for over 50 cities in the United States and over a 2-year period. You can work on a sampled subset, e.g., study the pollution condition of one city only, or study the pollution condition in a short period like one month or one quarter.

Data source:

https://www.kaggle.com/datasets/mayukh18/deap-deciphering-environmental-air-pollution

Your report should include (but not limited to) the following key points (90 marks):

(a) Introduction

Refer to the provided data and (or) any other related literature, describe the research background of the chosen project. Identify a business problem and discuss why it is interesting or important. Appraise the potential usefulness of association rule mining and (or) clustering techniques to solve the proposed problem.

(15 marks)

(b) Data Understanding and Data Preparation

Perform Data Exploration to understand the data. Describe and discuss the data characteristics (e.g., data size, data measurement types, data quality issues, etc.). Apply data preparation (if any, and examples include data sampling, treating missing values, data transformation, etc.) so that the data can be analysed using the identified data mining method(s). Clearly explain the rationale of your data preparation.

***Note**: In addition to the IBM SPSS Modeler, other tools and software are also permitted for data preparation.*

(20 marks)

(c) Modelling

Construct an association rule mining model and (or) implement a clustering solution using the IBM SPSS Modeler or other open-source tools; provide the following details:

- Explain your choice of the data mining method used for modelling
- Report your design decisions by explaining the parameter settings used
- Analyse and interpret the generated rules and (or) the generated clusters
- Evaluate the analysis results

(40 marks)

(d) Conclusion and Discussion

Summarise the constructed model(s) and discuss how the association and (or) clustering results address the identified business problem in Part (a). Propose at least ***two (2)*** deployment suggestions/recommendations that could help address the proposed business problem in Part (a).

(10 marks)

(e) Appendix

Provide a screenshot of your IBM SPSS Modeler streams with proper annotations and any other support material (e.g., Python code file).

(5 marks)

Writing (10 marks)

To write your report, you can refer to the structure of academic research papers. One possible framework includes several key sections:

- Introduction (introduce the background of the selected dataset or research field, propose business problem, and discuss possible solutions)
- Data description and data preparation (if any)
- Modelling and result interpretation
- Summary and Deployment/Recommendation
- Reference and Appendix

Your writing should be succinct but not at the expense of excluding relevant details. You should provide enough details/descriptions of your study in the report so that your work can be properly assessed. Make sure **you follow the report guidelines and style specified below**. The report must be self-contained. It is important to include all relevant tables and figures in the report as evidence to support your study and conclusion.

The followings are some details of report format:

- Length: **should not exceed 15 pages** (including the relevant graphs, tables, references, screenshots, and appendices (if any), but excluding the cover page). **Deduct 5 marks for each extra page. Deduct a maximum of 10 marks for**

excessive page count.

- Font Style: Times New Roman
- Font size: 12;
- Line spacing: 1.5
- Margins: 1" for the top, bottom, right and left
- Include the page number on each page

Some further suggestions:

- Ensure minimal grammatical and typographical errors
- Write clearly in plain English
- Write appropriately to the context
- Cite appropriate sources
- Provide a reference or bibliography at the end of the main report
- Good overall presentation of the report

---- END OF ASSIGNMENT ----