SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

ANL305
Association and Clustering

Tutor-Marked Assignment

July 2024 Semester

TUTOR-MARKED ASSIGNMENT (TMA)

This assignment is worth 20% of the final mark for ANL305 Association and Clustering.

The cut-off date for this assignment is **26 September 2024, 2355hrs.**

Note to Students:

Compose your report using Microsoft Office Word, and save either as .doc or **.docx (preferred).**

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

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.

If your course involves programming, you are urged to read the following articles as well:
https://wiki.cs.astate.edu/index.php/Plagiarism_in_a_Programming_Context

https://www.turnitin.com/blog/plagiarism-and-programming-how-to-code-without-plagiarizing-2

Global obesity has been rising steadily over the past few decades, becoming a significant public health concern. According to the World Health Organization (WHO), 2.5 billion adults (18 years and older) were overweight in 2022. Of these, 890 million were obese. Obesity is a complex health condition characterized by excessive body fat accumulation, leading to increased risks of chronic diseases such as diabetes, heart disease, and certain cancers. Studying the impact factors of obesity is crucial for understanding its multifaceted causes. By identifying these factors, researchers and healthcare professionals can develop targeted prevention and intervention strategies, promote healthier lifestyles, and inform public health policies.
Imagine you are a Data Analyst in the Healthcare Department. You are involved in a project which aims to study the factors related to obesity. You are given a dataset named *ObesityDataset.csv*, which contains survey information collected from 2111 respondents. The data dictionary of the dataset is shown in Table 1.

FIELD	DESCRIPTION
Gender	Gender: Male, Female
Age	Age: integer value between [14, 61] inclusive
Height	Height: numeric value in meters [1.45, 1.98]
Weight	Weight: numeric value in kilograms [39.0, 173.0]
family_history_with_overweight	Has a family member who suffered or suffers from overweight? • yes, no
FAVC	Do you eat high caloric food frequently? • yes, no
FCVC	Do you usually eat vegetables in your meals? • Never, Sometimes, Always
NCP	How many main meals do you have daily? • one, two, three, more than three
CAEC	Do you eat any food between meals? • Always, Frequently, Sometimes, No
SMOKE	Do you smoke? • yes, no
CH2O	How much water do you drink daily? • Less than 1 Litre, Between 1 and 2 L, More than 2 L
SCC	Do you monitor the calories you eat daily? • yes, no
FAF	How often do you have physical activity (in a week)? • Do not have, 1 or 2 days, 3 or 4 days, 5 days or more
TUE	How much time do you use technological devices such as cell phone, videogames, television, computer and others? • 0-2 hours, 3-5 hours, More than 5 hours
CALC	How often do you drink alcohol? • Always, Frequently, Sometimes, Never
MTRANS	Which transportation do you usually use?

	• Automobile, Bike, Motorbike, Public Transportation, Walking
NObeyesdad	Obesity level • Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III.

Table 1: Data Dictionary of the Obesity Dataset

Question 1

Based on the given background and dataset details, identify a potential business problem. State the business objective and data mining objective and appraise the suitability of using association rule mining for solving the proposed business problem.

(7 marks)

Question 2

Import the data into the SPSS Modeler and construct Apriori models using the following settings (*Data preparation is NOT required for Q2 before performing the following data and model settings*):

- Data Node:
  - Measurement setting: click "Read Values" under the tab "Types". Keep the default "Measurement" setting unchanged.
  - Role setting: *Nobeyesdad* = Target, non-Continuous fields = Input, Continuous fields = None
- Apriori Node:
  - Parameter setting 1: default setting
  - Parameter setting 2: your own customized setting

Answer the following questions:

a) For parameter setting 1, provide a screenshot of the generated rules, including the total number of rules and the first three rules. Discuss the modeling results and findings using parameter setting 1, with reference to the business problem you have proposed in Q1.

(15 marks)

b) Describe your own parameter setting 2 in detail and provide a screenshot of the generated rules to show the total number of rules and the first three rules. Compare the generated rules with those from parameter setting 1 and discuss any new insights or findings.

(28 marks)

Question 3

Suppose you would like to explore the numeric fields of the dataset and then construct other model(s), answer the following questions to complete the work:

(a) Import the data file into a new data node and apply the following Data Node settings:

- Keep the default measurement setting unchanged.
- Set the Role of *Nobeyesdad* = Target and all other fields = Input.

Report distribution screenshots and use up to 50 words to describe the distributions of the three numeric fields: *Age, Height, Weight*.

(5 marks)

(b) Transform all the three numeric fields into categorical variables using the "Binning" node in the SPSS Modeler. You can select the "*Equal-width*" method or the "*Tiles (Equal-count)*" method to complete the binning work. Report the distribution of the transformed fields.

Note: *Each of the newly generated fields should have* ***no more than four categories*** *(aka bins). You may use the same binning method (or a mixture of different binning methods) for the fields. It is also acceptable to propose an alternative binning method if you think the binning methods provided by the IBM SPSS Modeler are not suitable for some fields. Justify your choices briefly.*

(10 marks)

(c) Link a "Type" Node to the "Binning" node and report the screenshot of role setting. Construct new Apriori model(s) using the transformed dataset. Analyze and discuss your modeling results, including:

- Apriori node setting
- The number of generated rules
- Important rules
- A summary of the findings
- A discussion on the impacts of the three new fields, comparing these to the findings in Q2

Note: *to illustrate the modelling results, you may only need to show those rules that may contain useful patterns. It is not necessary to display all the generated rules.*

(25 marks)

(d) Propose ***two (2)*** deployment suggestions to answer the business problem proposed in Q1. Discuss the potential limitations of the study and propose one improvement.

(10 marks)

Report writing

Your writing should be succinct but not at the expense of excluding relevant details. Highlight only the points that are relevant to your discussion. Use plain and simple language. Some questions may not come with absolutely right or wrong answers. For such questions, you have the liberty to express your views about the problem. However, your points have to be supported by evidence and good reasoning. It's the quality and not the length that counts. Make sure you follow the report guidelines and style specified in this assignment.

Make sure you indicate your name and student number on the cover page of the report.

The topics in the main report should be presented in the order according to the sequence of the tasks/questions listed in the assignment; that is, in the order of (a), (b), ..., etc. You can have several sub-sections within a section if you deem appropriate.

The report must be self-contained. It is important to include all relevant tables and figures in the report as evidence to support the answers given.

The followings are some details of report format:

- Length: **should not exceed 9 pages** (including the relevant graphs, tables, references, screenshots and appendices (if any), but excluding the cover page) **Note: Deduct 5 marks for each extra page. Deduct 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
- Include less relevant details in the Appendix
- Good overall presentation of the report

---- END OF ASSIGNMENT ----