MULTIMEDIA UNIVERSITY

TMA 7021 Assignment

INSTRUCTIONS TO STUDENTS:

1. This assignment carries 30%.

2. If you are going for questions provided by the lecturer, select and answer only **ONE** question from the given questions. Alternatively, you can have your own title, but restricted to using data from data.gov.my only. The assignment is to be completed individually.

3. Your work must consist of Association Rule, one Classification technique and one clustering technique. Use only R and/or GeNie for this assignment. GUI must be in R Shiny.

4. Marks for presentation portion of this assignment are individual.

5. If plagiarism is detected, the assignment is granted 0%.

6. Your report must be produced using R Markdown. Publish your work to Rpub.com.

7. Deadline is week 11, Friday 12pm. No extension allowed.

QUESTION 1: Dengue Prediction

1. Overview of this title:

   In this project, you are requested to create a system that could predict the number of dengue cases in advance and suggest whether an outbreak will happen for a particular place.

2. The project objectives are:

   i. predict the number of cases for next 7 and 14 days.
   ii. Suggest places that could have an outbreak in next 14 days.

   Outbreak for a given location is defined as more than 4 cases in 200m radius.

3. Datasets

   The related data for this assignment are (i) dengue outbreak dataset [LocationHotspot2014.csv, LocationHotspot2015.csv], (ii) weather dataset [RainWind.csv], and (iii) population distribution in Selangor [Mukim_Selangor.pdf]. You may use other related variables/information such as Air Pollution Index, Water level, and many more.

4. Apply data mining techniques

   You need to use R for association rule, one classification and one clustering technique to extract interestingness in the patterns. You can use GeNie to construct Bayesian Networks as well.

5. GUI

   GUI must be in R Shiny.

QUESTION 2: Customer Profile for a Business Type

1. Overview of this title:

   This title aims to study user properties and patterns for a particular business type. You may select from food and beverage, laundry, airlines, and many more.

2. Your tasks involve

   Phase 1:
   Data Collection

   You are requested to collect as many data as possible. Examples of dataset for coin laundry business are:

   - Gender
   - Race
   - Age Range
   - Using washer
   - Number of washers
   - Using Dryer
   - Number of dryers
   - Family with kids
   - Number of kids
   - Types of items washed
   - Car type – mpv, sedan,
   - Car brand – honda, Toyota
   - Time of day
   - Day of week
   - Purchase detergent
   - Purchase softener
   - Etc…

   You are NOT supposed to interview or interrupt the customers. You can only observed from far.

   Phase 2:
   Preparing and preprocessing the data gathered from the questionnaire. Keep your data in CSV format.

   Phase 3:
   Apply Association rule mining, one clustering technique, and one classification technique) to find out and explain your observation on at least FIVE (5) queries.

QUESTION 3: Location Intelligence

1. Overview of this title:

   In this project, you are required to extract the relationships between the characteristics of a location and a particular business. Examples of business type can be food and beverage, laundry, convenience shops, clinics and many more.

2. Dataset

   Once you have identified a particular business type, you can use various sources for your data collection. For instance, the dataset for a laundry shop can be constructed 3 main parts: (1) geographical information, (2) weather information, (3) population information. You can use data from open data Malaysia.

   Attributes for geographical information can be:

   - facing main road
   - fast food
   - convenience shop
   - mamak stall
   - LRT
   - School
   - Hospital
   - Clinic
   - …and many more

   You should have as many variables as possible.

3. Apply Association rule mining, one clustering technique, and one classification technique) to find out and explain your observation on at least FIVE (5) queries.

GRADING SCHEME

The grading will be based on the report you submitted. Points will be assigned as follows:

Report	**Points**
Introduction	5
Data cleaning	25
Applying Data Mining Techniques correctly	20
Results and Discussion	30
Presentation (20%)	**Points**
Clarity in presentation	10
Q & A	10
Total	**100%**

MULTIMEDIA UNIVERSITY

Individual Assignment

TMA 7021

(Data Mining and Analytics)

Put Your Title here

Prepared by

Name, ID, contact number, email

SAMPLE REPORT

1.0 ABSTRACT

Each assignment should have an abstract, to be typed in 12-point, single-spaced and up to 150 words in length. It should contain the dataset that you use, your objective to achieved, method that you have selected and final result. Leave two blank lines after the abstract, and then begin the next section.

2.0 INTRODUCTION

Type the introduction of your manuscript in 12-point, single spaced. Do not bold the text and be sure that your text is fully justified. All paragraphs should be indented ½ inch. Please do not place any additional blank lines between paragraphs.

Introduction should describe the dataset that you use, preliminary examination on the data (how many attributes, if its qualitative or quantitative type, no missing values or outliers that can be identified etc.). It must also contain the objective that you want to achieve at the end of your mining process.

Tables and figures must be numbered separately. Table and figure captions should be 12-point boldface, aligned center. Initially capitalize only the first word of each figure caption and table title. Bold the header of each column. For table contents, use left alignment for texts and centre alignment for numbers. Set the vertical alignment of each cell to center.

TABLE 1. Title of the first table in boldface type

Column 1	Column 2	Column 3	
		Sub-Column	Sub-Column
Text	Text	Number	Number
Text	Text	Number	Number
	Text	Number	Number

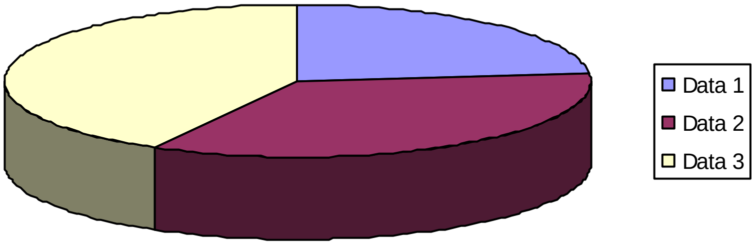

FIGURE 1. Title of figure

3.0 DATA CLEANING

State the reason why you have to clean the data, and what method that you use. Any changes to the data should be depicted in detail here, eg: replacing missing or outlier values with its new value, transformation of data or removing any attributes. You can use the data cleaning sub processes as sub-header, eg: 1. Cleaning ->identifying missing and outliers and how do you replacing it, 2. Data Selection → identifying which attribute should be maintained or removed and how do you judge that, as well as 3. Data Transformation-> why you need to transform data, transformation method that you use and final data range that you wish to have. Eg : all dimension should be in categorical data type and values range [1-10].

Type your main text in the same format as the introduction. The header title, just like the abstract and introduction title should be in 12-point, boldface type. Make sure that you insert page number at the bottom of each page, aligned right.

Long lists of notes or bibliographical references are generally not required. The method of citing references in the text is 'name date' style, e.g. 'Mallach (2002) claimed that …'. When citing from more than one reference, separate the authors' names by a semicolon, for example '… (Mallach 2002; Alahakoo et al. 2002).'

4.0 THE SELECTED DATA MINING TECHNIQUES

Identify 3 techniques/methods that you can use to achieve your objective. You should know the strength, weakness and the requirement that you must meet, in order to use those techniques/ methods. Explain each technique.

Long lists of notes or bibliographical references are generally not required. The method of citing references in the text is 'name date' style, e.g. 'Mallach (2002) claimed that …'. When citing from more than one reference, separate the authors' names by a semicolon, for example '… (Mallach 2002; Alahakoo et al. 2002).'

4.1 ASSOCIATION RULE MINING *(this is just an example only)*

Explain about this technique in detail of how it is applied in your case.

4.2 CLASSIFICATION TECHNIQUE 1 *(this is just an example only)*

Explain about this technique in detail of how it is applied in your case.

4.2 CLUSTERING TECHNIQUE 1 *(this is just an example only)*

This is where you write in the detailed ***process*** on techniques that you applied on your chosen dataset.

5.0 RESULTS & DISCUSSION

Write your result, and presenting you output in the most appropriate form. It could be in text or graphical format.

6.0 CONCLUSION

Type your conclusion here. The header title should be in 12-point, boldface type. This whole manuscript should not exceed 20 pages (minimum 10 pages).

7.0 REFERENCES (IF RELEVANT)

Use **APA format** for referencing. This section is compulsory as well. Don't forget to cite your dataset source (you can get it from you dataset desc file, and any paper or book that you use as a reference.

End references should be listed in alphabetical order. Use 12-point font, fully-justified. Indent the subsequent line(s) ½ inch from the left.

Alahakoo, D., Halgamuge, S. K. and Srinivasan, B. (2002). Dynamic self-organizing maps with controlled growth for knowledge discovery. *Journal of IEEE Transactions on Neural Networks,*11(3), 601-604.

Mallach, E. G. (2002). *Decision support and data warehouse system.* Boston: McGrawhill.