CAB330: Assignment 2

Due date: 24th Oct 2021 (Week 12, Sunday)
Weighting: 25%

Introduction:

The purpose of this assignment is to give you an understanding that data mining methods can be applied to various types of data sets such as record data, transactional data, text data and web logs. This assignment is divided into four parts: Clustering, Association mining, Text Mining and Web Mining. You will use Python with all of the libraries you have learned to use so far.

Task 1: Descriptive Data Mining - Clustering

This forest fire data set '*forestfires.csv*' is from the Montesinho natural park, the Tr´as-os-Montes northeast region of Portugal, where the forest Fire Weather Index (FWI) is used for rating fire danger.

Attribute Information (http://www3.dsi.uminho.pt/pcortez/fires.pdf)

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: 'jan' to 'dec'
4. day - day of the week: 'mon' to 'sun'
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2: 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84

The research on this dataset has noticed that the combination of four weather condition attributes (temperature, rain, relative humidity, and wind speed) is likely capable of predicting small fires. This task requires creating segments and finding the minimum number of weather condition segments to allow the Bureau of Meteorology to advertise small fire alarms to the residents.

Your task is to conduct k-mean clustering on this data set and find and describe the **optimal number of effective clusters**. Answer the followings in relation to this data and analysis.

1. Can you identify data quality issues in this dataset such as unusual data types, missing values, or very skewed attributes?
2. Which variables were included in your analysis and what were their roles and measurement level set? Justify your choice.
3. Build a default clustering model with K=3 and answer the following questions:

   a. How many records are assigned into each cluster?
   b. Plot the cluster distribution using pairplot. Explain key characteristics of each cluster/segment.

4. What is the effect of using the standardization method on the model above? Does the variable normalization process enable a better clustering solution?
5. Using elbow method and silhouette, find the optimal K. What is the best K? Explain your answer.
6. How the outcome of this study can be used by decision makers? Given an application where this clustering outcome can be used by the Bureau of Meteorology.

Task 2: Descriptive Data Mining - Association

A food market store is interested in determining the associations between items purchased by its customers. The store has chosen to conduct a market basket analysis of items purchased. This data set, "*retail_row.csv*", includes over 200,000 transactions made over the past three months.

Attribute Information

1. LOC: Point of sale device identification number (1 to 10).
2. TRANS_ID: Unique transaction identification number for a given sale. A sale may include several products and thus the same transaction id may occur over several rows.
3. TRANS_DATE: Date of transaction
4. PRODUCT: Products Purchased in [Yoghurt, Jam, Shampoo, Bread, Egg, Milk, Tea, Cordial, Peanut butter, Dishwashing liquid, Cereal, Coffee, Conditioner, Butter, Sugar, Jelly, Cheese]
5. QUANTITY: Quantity of this product purchased (always set to 1 by a point of sale device)

Your task is to conduct association analysis on this data set, and answer the followings in relation to this data and analysis.

1. Can you identify data quality issues in this dataset for performing association analysis?
2. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
3. Conduct association mining and answer the following:
   a. What is the highest lift value for the resulting rules? Which rule has this value?
   b. What is the highest confidence value for the resulting rules? Which rule has this value?
   c. Plot the confidence, lift, and support of the resulting rules. Interpret them to discuss the rule-set obtained.
4. The store is particularly interested in products that individuals purchase when they buy "Yoghurt".
   a. How many rules are in the subset?
   b. Based on the rules, what are the other products these individuals are most likely to purchase?
5. How the outcome of this study can be used by decision makers?

Task 3: Text Mining

A leading company is planning to start online review analysis to understating patients' opinions on their products (drugs). The company has collected a dataset, "*drugReview.csv*", which includes patient reviews about benefits on specific drugs along with related conditions.

Attribute Information:

1. urlDrugName (categorical): name of drug
2. benefitsReview (text): patient review on benefits
3. rating (numerical): 10 star patient rating
4. sideEffects (categorical): 5 step side effect rating
5. effectiveness (categorical): 5 step effectiveness rating

Perform text mining on this dataset to determine clusters of drugs based on similar topics that can be obtained from the patients' reviews on benefits and answer the followings in relation to this data and analysis.

1. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
2. Based on the ZIPF plot, list the top 10 terms that will be least useful for clustering purpose.
3. Did you disregard any frequent terms? Justify their selection.
4. Justify the term weighting option selected.
5. What is the number of input features available to execute clustering?
   (FYI: Note how the original text data is converted into a feature set that can be mined for knowledge discovery.)
6. State how many clusters are generated? Name each cluster meaningfully according to the terms that appear in the clusters?
7. Identify the first six high frequent terms (that are not stop words) in the start list?
8. Describe how these clusters can be useful in the online service for recommending drugs based on their actual benefits to patients.

Task 4: Web Mining

For an e-commerce business, the website structure and site plan were established with the efficiency and usability in mind, but its effectiveness was not verified. Only basic statistics have been produced through simple report and query techniques, but they provide no means for sophisticated web site analysis and predictions. Your task is to determine the user browsing patterns of the website and analyze those patterns to provide recommendations to improve the website.

The data set you will use is a web server log file, "*server_logs_raw.txt*", which is an original text file that needs to be processed with the steps required for web usage mining as explained in the practical. Please note that a '-' in a field indicates missing data.

Your task is to pre-process the given dataset and apply a suitable data mining operation, such as classification, clustering, or association mining, to the raw log data set. Answer the followings in relation to this data and the analyses that you have chosen.

1. Pre-process the ***log*** data to identify useful attributes based on columns in the text file such as IP_Address,Timestamp, Request, Status Code or Referrer.
2. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
3. Apply a data mining task on the processed dataset. Explain the rationale behind selecting the data mining task/method.
4. Discuss the results obtained. Discuss also the applicability of findings of the method. You should include only a high-level managerial kind of discussion on the findings. It should not just be an interpretation of results as shown in results.

Marking Guide:

Distribution of Marks (Total: 25 marks)

We would mark your data mining projects in the Week 13 practical class. You should be prepared to show your final diagrams and results panels to your marker. The marker will ask each individual student questions and will assign individual mark (~15%).

In data mining, there is hardly ever a single solution. Also many times, there is no correct or wrong solution. You may find that your project partner may have different solution as yours. Your group should decide on a single project that you would like to be marked. Submit the report discussing the final project components.

The marks are distributed as follows:

Task 1 (7 marks)
Task 2 (5 marks)
Task 3 (7 marks)
Task 4 (6 marks)

Instructions

1. The assignment is due on 24th October 2021.

2. You should submit the assignment report via Blackboard Assignment.

3. This assignment will be **marked in the practical class in Week 13**. We will check the code, plots and results, along with the assignment report, to assign you marks. The entire team should be present to show the project result and answer the questions raised by marker. We will ask questions to each student, and will assign about 15% of total marks as per individual performance.

4. The datasets required for this assignment can be found on Blackboard with the file named as **Asm2-data.zip**. It includes four datasets:

    a. *forestfires.csv* to perform clustering
    b. *retail_row.csv* to perform association mining
    c. *drugReview.csv* to perform text mining
    d. *server_logs_raw.txt* to perform web mining

5. Name the case-study report as **asm2_[groupName].doc.**The word file should include a cover page with Student ID number and full name (as in QUT-Virtual) for all students, along with the group name. Combine this file with your **team contract** and your **source code** and name the compressed file as **asm2_[groupName].zip.** Submit the zip file on **Blackboard (under assessment panel Assignment 2).**

6 . The **project report** should be divided into four parts according to each task, each part starting from a new page. There is no need of including introduction, summary, conclusion or references in the report. The report should just include responses to the questions set in the case-study. Some answers may require screen shots. Answer the questions in the case study for each model appropriately and succinctly. If a case-study step is about conducting a process, you do not have to provide an explanation or a screen shot. Include the final screen shot when you added all kinds of nodes in a particular analysis. However if a question such as "Examine the results of clustering/association mining" is asked, you then need to explain what, why or why not? While you may like to go into extreme detail about, you will not have the space to do so. Rather, write down the important points and attach the important screen dumps, to show that you have thought the matter through.

7 . This is a group assignment. The team size is three. You can continue the same group as in case study 1. If you have formed a new group after assignment 1, please notify the lecturing staff. They will remove you from the existing group. In this case, you need to register your new team at Blackboard.

8. The group is to be ARRANGED and MANAGED by you. As in real life, the performance of the individuals in the team shall be judged by the performance of the team together, so choose your partners carefully.

9. Of course, the work your group hand in must be your own; no collaboration or borrowing from others groups is permitted. Read the Assessment Policies on Blackboard or QUT Website.

Assignment Criteria Sheet

Criteria	Comments and scoring
Non Submission of all components/ evidence of plagiarism	0
Has demonstrated a task with a working model with /without submission and demonstrates the ability to run the program and add some components.	1-5
Has demonstrated a task with a working model having a data source, and diagram with substantial but incorrect implementation of at least one of the components. Questions were poorly answered.	6-11
Has implemented all tasks with at least two being substantially correct. Shows some understanding of concepts with some success in applying knowledge. Only basic questions were answered.	12
Has implemented all four tasks: One mining task is fundamentally correct, with substantially correct work flow diagrams which may contain minor errors. Response to questions shows fundamental understanding of terms and concepts.	13-15
Has fundamentally correct implementation of all tasks i.e. selection of correct variables in data, correct allocations, understanding, and explanation of clusters, findings association rules, finding clusters in text data with good term features, and application of an appropriate data mining operation to the log data. Shows competency in applying text mining. Many questions have been reasonably answered. Demonstrate a good understanding of the methods and terms used in clustering, association mining, text mining and web mining, during written and verbal analyses. Some minor errors are allowed. Written application is required to be of reasonable standard.	16-18
Has implemented all of the requirements above with very few errors. A strong focus on application of tools, and evaluation and interpretation of results is evident.	19-21
All of the criteria above are met, extensive model generation and analyses have been conducted to produce exceptional outcomes. Have applied principles learnt in lectures to enhance the results.	22-25