IFQ509 Data Exploration and Mining

Assignment 2A: Project

Page limit: 20–30 pages (including screenshots, tables, diagrams, etc.)

Weighting: 30% (Note: **Assignment 2B: Quiz (https://canvas.qutonline.edu.au/courses/1732/assignments/8321)** **(https://canvas.qutonline.edu.au/courses/1732/assignments/8322)** contains questions relating to **Assignment 2A: Project** **(https://canvas.qutonline.edu.au/courses/1732/assignments/8206)**)

Due date: 11.59pm AEST Wednesday 11 December 2024 (Assignment Week)

Team/Individual: Team

After you have read this information, head over to the **Assignment 2 Q&A** **(https://canvas.qutonline.edu.au/courses/1732/discussion_topics/150397)** discussion board to ask any questions and see what your peers are saying about this assignment.

Assignment overview

For this assignment, you will use Python libraries and code to analyze three case studies and their datasets. The goal is to find insights that could benefit the relevant sectors. You should mainly use the code and libraries covered in the practical sessions. If you use any code or libraries not covered in the course, please explain why.

This assignment supports unit learning outcomes 1, 3, 4, 5 and 6.

Team formation

You are required to work in groups for the report submission. You can continue working with the same group as for Assessment 1 or form a new group.

You may wish to use the **Team assignment discussion board (https://canvas.qutonline.edu.au/courses/1732/discussion_topics/150395)**, which is only accessible to your team and your OLA, for these meetings.

Assignment details

The assignment will be evaluated based on two components. The first component will be assessed from the group report submitted on Canvas. The second component will be assessed through an individual quiz.

This assignment allows you to showcase your Descriptive and Predictive Data Mining skills. You will build various data mining models specific to each case study, such as association mining, clustering, decision trees, logistic regression, and neural networks. This will demonstrate the technical skills and knowledge you've gained from the practical sessions, lectures, and readings.

In this assignment, you will use the following dataset.

- **D1.csv (433 KB) (https://canvas.qutonline.edu.au/courses/1732/files/486127?wrap=1)** **(https://canvas.qutonline.edu.au/courses/1732/files/486127/download?download_frd=1)**
- **D2.csv (356 KB) (https://canvas.qutonline.edu.au/courses/1732/files/486126?wrap=1)** **(https://canvas.qutonline.edu.au/courses/1732/files/486126/download?download_frd=1)**

Work through the following tasks to ensure you complete all necessary parts of this assignment.

Task 1. Case study 1: Association mining to understand movie preferences

IMDB is an online database of movies that records user ratings of movies based on their likings. This dataset **D1.csv (CSV 433 KB) (https://canvas.qutonline.edu.au/courses/1732/files/486127?wrap=1)** **(https://canvas.qutonline.edu.au/courses/1732/files/486127/download?download_frd=1)** consists of 8000 observations, each presenting a user's rating for a film. There is a total of 646 users and 2943 unique movies in this subset of data. More details about the entire dataset can be found on the Kaggle website (n.d). The following table details the attributes in the dataset.

Table 1: D1.csv variable description

Attribute	Description	Data type
userID	Unique identifier of the user	ID
movieID	Unique identifier of the movie	ID
rating	Rating given by the user to the movie	Numeric
timestamp	Date and time when the user rated the movie	Datetime
imbdId	Unique identifier of the movie given by the IMDB	ID
title	Title of the movie	String

Table 1: D1.csv variable description (2024) QUT School of Computer Science

Task: Analyse the dataset 'D1.csv' using the Apriori association mining method to identify frequently watched films. In this dataset, the list of films watched by each user is considered a transaction. All information to complete this task can be found on the following page:

- **5.9 Activity: Association analysis in Assignment 2A: Project (https://canvas.qutonline.edu.au/courses/1732/pages/5-dot-9-activity-association-analysis-in-assignment-2)** .

Task 2: Case Study 2: Clustering COVID-19 data

For **Assignment 1: Problem-solving task (https://canvas.qutonline.edu.au/courses/1732/assignments/8205)**, you were given a COVID-19 dataset of individual responses, including demographic, behaviour and health condition values. This dataset was pre-processed to remove some data errors and variables, resulting in the **D2.csv (356 KB) (https://canvas.qutonline.edu.au/courses/1732/files/486126?wrap=1)** **(https://canvas.qutonline.edu.au/courses/1732/files/486126/download?download_frd=1)** dataset. Your task is to perform clustering on D2.csv and explain the least number of functional clusters discovered.

Table 1: Dataset variable description

No.	Column name	Description
1	Gender	Male, Female or other
2	Age	Age quantile
3	Height	Height of the person in cm
4	Weight	Weight of the person in kg
5	Blood type	Type of the person's blood
6	Insurance	If the person has insurance or not?
7	Income	Type of income (e.g., low, medium, high, or gov)
8	Smoking	Information on how often the person smokes
9	Alcohol	Level of alcohol consumption
10	Contacts count	Number of people the person has contacted

No.	Column name	Description
11	Working	Status of the person's work
12	Worried	On a scale of 1 to 5, indicating how worried the person is
13	Covid19 positive	1 or 0, indicating positive or negative status

Table 2: Variable description of the dataset (2024) created by QUT School of Computer Science

Task: Imagine you are interested in understanding the clusters of responses who stated COVID-19 positive. Complete the activities on the following page.

- **6.10 Activity: Clustering in Assignment 2A: Project. (https://canvas.qutonline.edu.au/courses/1732/pages/6-dot-10-activity-clustering-in-assignment-2)**

Task 3: Case Study 3: Building and Evaluating Predictive Models

Use the COVID-19 dataset D2.csv (363 KB), as in case study 2, for predictive modelling with various techniques. Based on past observations, the objective is to predict whether an individual will test positive for COVID-19 (i.e. 1) or negative (i.e. 0).

Tasks: Build different predictive models, such as decision trees, regression models, and neural networks, on this data set and compare their performances. Your team needs to conduct association mining on this dataset. Complete the activities on the following pages.

- **7.12 Activity 2: Making a Decision Tree for Assignment 2A: Project. (https://canvas.qutonline.edu.au/courses/1732/pages/7-dot-12-activity-2-making-a-decision-tree-for-assignment-2)**
- **8.10 Activity 1: Making a Regression Model for Assignment 2A: Project. (https://canvas.qutonline.edu.au/courses/1732/pages/8-dot-10-activity-1-making-a-regression-model-for-assignment-2)**

- **8.11 Activity 2: Making a Neural Network for Assignment 2A: Project. (https://canvas.qutonline.edu.au/courses/1732/pages/8-dot-11-activity-2-making-a-neural-network-for-assignment-2)**

Task 4: Final remarks

Once your team has built your models, you must consider your final remarks on the predictive modelling you completed. Your team needs to answer the following questions based on all your models and analysis:

- Is there a model you will use in decision-making? Justify your choice. Draw an ROC chart and Accuracy Table to support your findings.
- Based on this analysis, can you summarise the positives and negatives of each predictive modelling method?

Save this work (including screenshots of your models), as it will form part of your final report.

Submission

Your team needs to compile all of the findings of your analysis (both predictive and descriptive) into a final report. The report should **answer each question** of the **three scenarios** and **modelling tasks**. Refer to the **Assignment 2A: Structure guide (https://canvas.qutonline.edu.au/courses/1732/pages/assignment-2-structure-guide)** to create the structure for your report. The report should include responses to the questions set in the scenario. Some answers may require screenshots. Use them as needed, but you may include a table detailing those results. While you may like to go into a lot of detail, you will not have the space to do so. Instead, write down the essential points and attach the important screenshots to show that you have thought the matter through. The report is expected to be about **20–30 pages long**, including screenshots.

The report should be submitted using the process outlined in the **following Submission details overview** accordion.

Note: Every team member needs to **submit** a copy of the assignment to the submission point on this page.

References

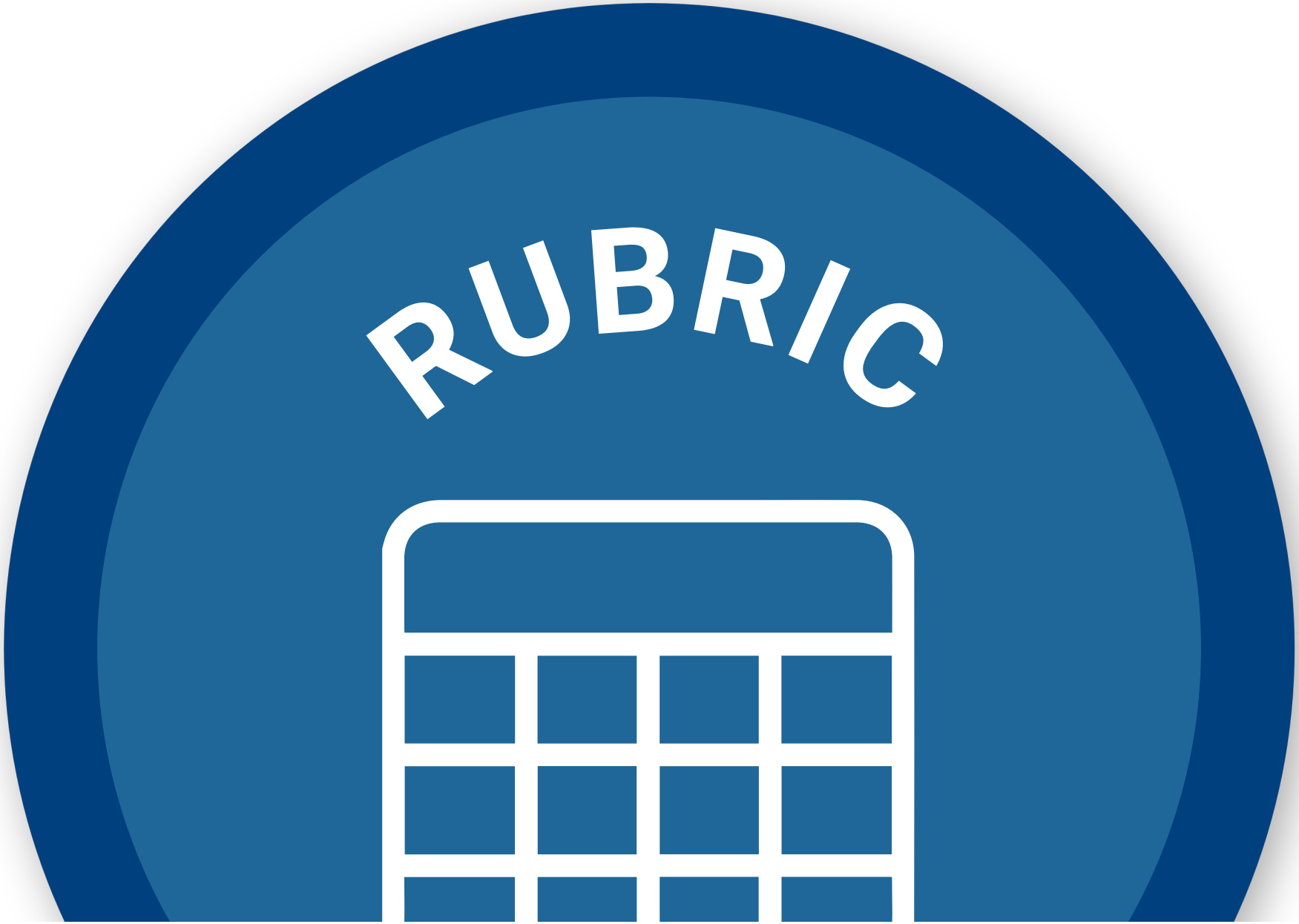
Kaggle Inc. (n.d.). The movie dataset. Delaware, USA. ***https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset*** ***(https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset)***

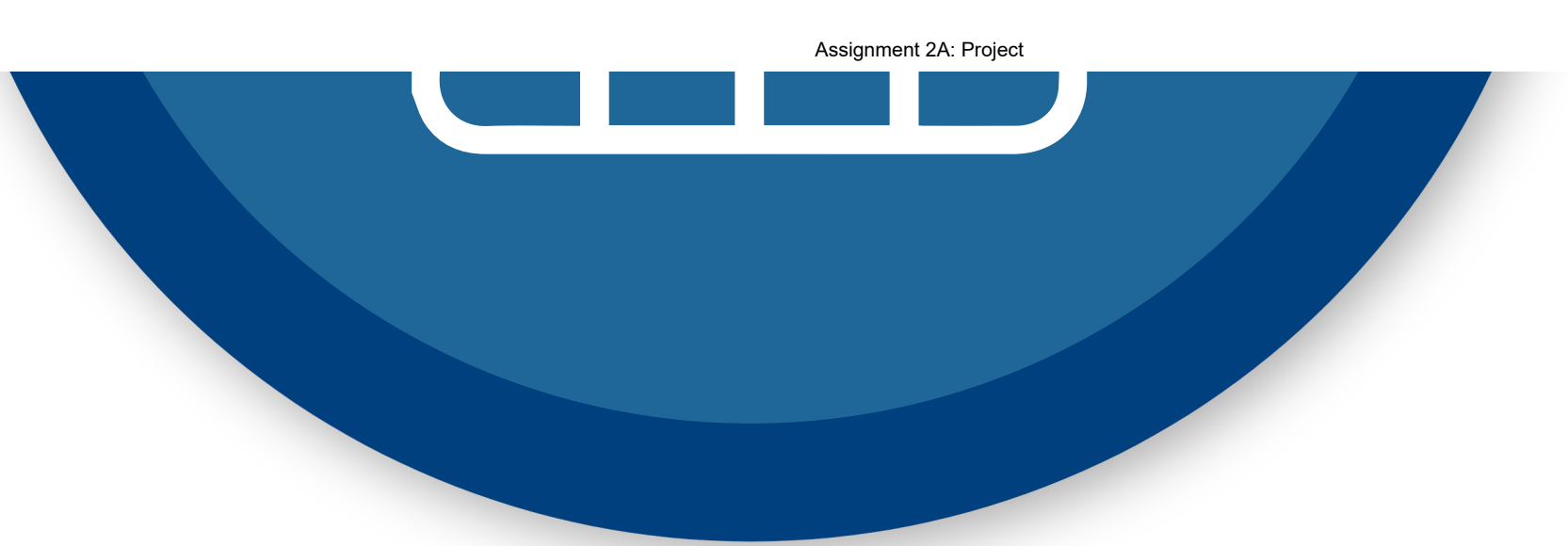
(https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset) Supporting resources

- **QUT cite|write: APA (https://www.citewrite.qut.edu.au/cite/qutcite.jsp#apa)** .
- **Assignment 2A: Structure guide (https://canvas.qutonline.edu.au/courses/1732/pages/assignment-2-structure-guide)** .
- **QUT cite|write: Writing a report (https://www.citewrite.qut.edu.au/write/writing-well/report.html)** .
- **QUT cite|write: APA (https://www.citewrite.qut.edu.au/cite/qutcite.jsp#apa)** .

Assignment criteria

1. Approach.
2. Implementation.
3. Findings.
4. Documentation.


RUBRIC



Rubric

Select the '**Rubric**' icon for more detail about this assignment's marking criteria.

▶ Assessment declaration

▶ Submission details overview

▶ Request Assignment Extension