MMU®

MULTIMEDIA UNIVERSITY

TDS 3301
GROUP ASSIGNMENT

INSTRUCTIONS TO STUDENTS:

1. This assignment carries **20%.**

2. Answer all questions. The assignment is to be completed in a group of **maximum 3 members**.

3. For Question 1 and 2, the maximum number of pages is 5, excluding the references, using the template given. All reference and citation formats must follow APA.

4. For Question 3, the output is a JupyterLab file, and a Streamlit file. You must host it to Heroku.com or similar hosting.

5. If plagiarism is detected, the assignment will be granted 0% for all members of the group.

6. Deadline for submission is on 24/9/2021, 12pm. Submission is made via Google Classroom. Timestamp will be logged as proof of submission.

7. The project leader only should submit THREE items in a ZIP file: (i) all the research papers you referred to, (ii) a report for Question 1 in PDF, and (iii) JupyterLab and Streamlit Python file. Name your zip file **`<Student ID>_<Project Leader Name>.ZIP`**

QUESTION 1: Applications of Data Mining [5 marks]

Select only <u>ONE</u> domain from the following list:

(i) Healthcare
(ii) Retail
(iii) Domain Name Server
(iv) Banking or Insurance

Discuss the different techniques proposed and used by researchers to tackle challenges in the selected domain. Your findings must be supported by at least 15 journal papers recent than 2015. Regardless of the domains, they must be related to data mining or machine learning or Artificial Intelligence. You should provide conclusion and limitations of current findings and the future work.

Table 1: Challenges and problems solved by researchers

Author	Problem solved 1	Problem solved 2	Problem solved 3	Problem solved *n*
Ali(2018)		x	x	x
Chong & Ting (2019)				x
James et al.(2010)	x			
...	...	...	...	...

Example of discussion:

Table 1 shows the challenges and problems solved by different researchers when dealing with COVID-19. From the table, most of the reports have been centering around vaccine supply and delivery. This can be showed by the number of count where 10 out of 15 papers focused on vaccine supply optimization. The second most challenging work as reported by researchers are predicting the outbreak of an area in next 14 days....

Continued...

Table 2: Data mining techniques used by researchers

Author	Data Mining Technique 1	Data Mining Technique 2	Data Mining Technique 3	Data Mining Technique *n*
Ali(2018)	x			x
Chong & Ting (2019)			x	
James et al.(2010)	x	x		
...	...	...	...	...

Example of discussion:

Table 2 depicts the information about different data mining techniques used in work related to COVID-19. From the above table, decision-tree has been widely accepted in most of the reported research work. Based on the 15 papers, 10 out of them employed decision-tree. The focus areas are vaccine supply chain optimization and forecasting. More recent work has also reported on using geography particularly geospatial analytics to predict the outbreak....

Continued...

QUESTION 2: Tools and Programming Languages for Data Mining [5 marks]

Identify not less than 10 tools/programming languages for data mining. Discuss the features of each one. Table your findings in the table shown below. Provide a short paragraph stating your findings. The findings should focus on the strength and weaknesses of each tool/programming language and also explain under what condition one tool/programming language is preferred over another.

Software/Tool	Open Source	Drag and Drop	…etc	…etc	…etc
…		x			x
…				x	
…		x	x		

Example of discussion:

From the table above, it is clear that X is a good tool for beginner who is totally new to data mining. X has an drag-and-drop GUI that allows users to easily experiment with data mining workflow. Customization of each process can be easily done within the tool environment…

Continued…

QUESTION 3: Python Programming [10 marks]

Study "Open Data on COVID-19 in Malaysia" by the Ministry of Health (MOH), Malaysia via `https://github.com/MoH-Malaysia/covid19-public`. Use only datasets from the categories "Cases and Testing", "Healthcare", "Deaths", and "Static data" for this assignment.

Answer the following questions and prepare your findings using the "Streamlit" package. Upload it Heroku.com. Each analysis must have at least one chart and a short paragraph explaining your findings.

(i) Discuss the exploratory data analysis steps you have conducted including detection of outliers and missing values?

[2 marks]

(ii) What are the states that exhibit strong correlation with (i) Pahang, and (ii) Johor?

[2 marks]

(iii) What are the strong features/indicators to daily cases for (i) Pahang, (ii) Kedah, (iii) Johor, and (iv) Selangor? [Note: you must use at least ***2*** methods to justify your findings]

[3 marks]

(iv) Comparing regression and classification models, what model performs well in predicting the daily cases for (i) Pahang, (ii) Kedah, (iii) Johor, and (iv) Selangor?

Requirements:

1. Use TWO(2) regression models and TWO(2) classification models
2. Use appropriate evaluation metrics.

[3 marks]

End of Pages.