MULTIMEDIA UNIVERSITY

TDS 3301 DATA MINING
Group Project

INSTRUCTIONS TO STUDENTS:

1. This project carries 30%. Choose ONE (1) question only.
2. This project is a group project with a maximum of 3 members in a group.
3. If plagiarism is detected, the assignment is granted 0%.
4. The maximum number of pages is 10, including references, using the template given. References and citations must be using APA format only.
5. Deadline for submission is on 22/10/2021, 12pm. Submission is to be made via Google Classroom. Timestamp will be logged.
6. You should use external dataset to supplement your analysis work.
7. Your work must consist of one association rule mining algorithm, at least two classification models, at least 2 regression models, and one clustering technique. You must provide visualization to your findings and analyze them accordingly.
8. There shall be a project presentation session in Week 13 or Week 14.
9. Project leader should submit THREE items in a ZIP file: (i) a report in PDF, and (iii) Datasets used, and (iii) a Jupyter Lab and Streamlit file. Name your zip file **<ID>_<Project Leader Name>.ZIP**

Your report should consist of the following items:

Item	Marks
Exploratory Data Analysis and Data Pre-Processing Examples of question, *but not limited to*: - Can I add extra data point? - Can I do feature engineering? - How to visualize the content? - Do I need to preprocess the dataset? - Do I need to perform data imbalance treatment? - How about outliers and missing values?	10
Feature Selection Examples of question, *but not limited to*: - Justify why that feature selection technique(s) selected? - How should I obtain the optimal feature set?	5
Appropriate Machine Learning Techniques Used Examples of question, *but not limited to*: - Discuss on the suitability of the techniques to the data - How do you obtain the optimal accuracy? What tweaking is needed?	10
Deployment. Examples of question, *but not limited to*: - Visualization - Web Services (e.g. hosted at Heroku.com)	5

Continued...

QUESTION 1: Profiling Customers in a Self-Service Coin Laundry Shop

Figures above show customers visiting a self-service laundry shop. Customers with different attires visited the shop. As a data scientist, you are expected to provide insights to the owner. You have the freedom to provide any useful insights.

Study the dataset carefully. You should extend from the following examples:

(i) Is there any relationship between basket size and race?

(ii) What types of customers will likely to choose Washer No. 2 and Dryer No. 3?

(iii) Did weather information impact the sales?

(iv) ...etc

Continued...

QUESTION 2: Malaysia COVID-19 Cases and Vaccination

Study the open data on COVID-19 in Malaysia by the Ministry of Health (MOH), Malaysia at https://github.com/MoH-Malaysia/covid19-public and vaccination data by Malaysia's National Covid-19 Immunisation Programme via https://github.com/CITF-Malaysia/citf-public.

You are required to use both datasets to find useful and important insights by combining the two datasets.

Examples of question, which you should expand from them, and not limited to the following:

(i) Has vaccination helped reduced the daily cases? What states have shown the effect of vaccination?

(ii) What state(s) require attention now?

(iii) Does the current vaccination rate allow herd immunity to be achieved by 30 November 2021? You can assume that herd immunity can be achieved with 80% of population has been vaccinated.

(iv) What clustering algorithm is suitable to find different clusters from the dataset? How do you quantify the risk?

(v) Is there any correlation between vaccination and daily cases for Selangor, Sabah, Sarawak, and many more?

(vi) How many different clusters in the movement data as shown in the MySejahtera data?

(vii) ... etc

End of Pages.