MULTIMEDIA UNIVERSITY

TDS 3301 DATA MINING
Group Project

INSTRUCTIONS TO STUDENTS:

1. This project carries 30%.
2. This project is a group project with a maximum of 4 members in a group.
3. If plagiarism is detected, the assignment is granted 0%.
4. The maximum number of pages is 15, including the front cover.
5. Deadline for submission is on 15/1/2023, 12pm. Submission is to be made via Google Classroom. Timestamp will be logged as proof of submission. Marks will be deducted for late submission.
6. You should use external dataset to supplement your analysis work.
7. Your work must consist of at least TWO feature selection methods, ONE association rule mining algorithm, at least TWO classification models, at least TWO regression models, and ONE clustering technique.
8. There shall be a project presentation session in Week 14. There will be individual presentation marks. Make sure the report has clearly indicated the contribution of each member. Member without contribution will be granted 0%.
9. Report must be prepared using the template given.
10. 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 Examples of question, *but **not limited to***: - What are the extra data points that I can include? - How to visualize the content? - What data transformation to the dataset? - Do I need to perform data imbalance treatment? - How about outliers and missing values? - Relationships between variables? - …etc	5
Feature Selection Examples of question, *but not limited to*: - What are the suitable feature selections techniques to use? - How should I obtain the optimal feature set?	5
Model Construction and Comparison Examples of question, *but not limited to*: - Why is a particular model used? - How to visualize the output of the model? - How to validate and compare models? - What is the impact of SMOTE and non-SMOTE datasets? Is the difference statistically significance? - What is the impact of features towards the modeling? - How did you perform hyper-parameter tuning? - Does stacking ensemble classifier work better? - …etc	10
Deployment Examples of question, *but not limited to*: - How to connect to your web API? - Hosted streamlit? - Do I have real-time prediction and visualization? - How should I improve the performance? What strategy? - Features such as download report, email…etc - …etc	5

Continued…

Assessment on the Presentation & Individual Contribution component.

Presentation & Individual Contribution Examples of question, *but not limited to*: - Did I answer all the questions asked? - Do my contribution in the group correctly captured and stated in the report? - Has the streamlit been hosted and loaded successfully? - ...etc	5

Continued...

QUESTION: Customer Customers

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 consider 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

End of Question