UECS3213 UECS3453 UECS3483 DATA MINING ASSIGNMENT	
Programme(s)	Bachelor Of Science (Honours) Applied Mathematics with Computing
	Bachelor Of Engineering (Honours) Electronics (Computer Networking)
	Bachelor Of Science (Honours) Software Engineering
Trimester	June 2023
Course Lecturer	Dr. Fatimah Audah Md. Zaki
Submission Date	28 August 2023 (Week 11)
Submission Platform	WBLE

Student Information					
Group no.	**Student Name**	**Student ID**	**Programme Code**	**Signature**	**Final mark**

Course Learning Outcomes Assessed
CLO2: Create programming solutions using data mining techniques for given problem
CLO3: Evaluate performance of data mining solutions for a given problem
CLO4: Construct a data mining project as a team

<u>**Data Mining Assignment: Analyzing a Real-World Dataset**</u>

Overview

In this assignment, you will analyse a real-world dataset and apply various data mining techniques to gain insights and answer specific research questions. Your task is to analyse the data, pre-process it, develop data mining models, write a report and present it.

Dataset

For this assignment, you will be choosing your own dataset to work with. The dataset should be related to a topic of your interest, and it should be publicly available. Follow the instructions below to choose your dataset:

1. Identify a topic of interest. This can be anything from sports to politics to health to social media, etc. Think about what you're interested in and what kind of data might be available on that topic.

2. Search for datasets related to your topic. There are many sources of publicly available datasets that you can use, including:
   - Kaggle: https://www.kaggle.com/datasets
   - UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php
   - Google Dataset Search: https://datasetsearch.research.google.com/
   - Data.gov: https://data.gov/
   - KKMNOW: https://data.moh.gov.my/
   - GitHub: https://github.com/

3. Evaluate the dataset. Once you've found a dataset that you're interested in, take some time to evaluate it. Consider the following questions:
   - Does the dataset have enough data to be useful for your project?
   - Is the data of high quality? Are there any missing values or outliers?
   - Are the data types appropriate for your project?
   - Are there any privacy or ethical concerns with the data?

4. Choose your dataset. Once you've evaluated your options, choose a dataset that is appropriate for your project and that you're interested in working with.

5. Confirm your dataset choice with your lecturer. Before you start working with your dataset, confirm your choice with your lecturer to ensure that it is appropriate for the project and that you'll have the support you need.

Tasks

1. **Data Understanding (15 marks)**
   - Identifies the dataset and describes its source.
   - Summarizes the variables in the dataset.
   - Identify the objective of analyzing the dataset, and potential issues or limitations with the dataset.

   (Maximum 500 words, approximately 2 pages)

2. **Data Description (15 marks)**
   - Identify 3 questions to be addressed using the dataset.
   - Use visual methods to find the key features, patterns, and trends in the data (minimum 3 visualizations).
   - Elaborate on the findings in the report.

   (Maximum 500 words, approximately 3 pages)

3. **Data Preprocessing (15 marks)**

   Demonstrate understanding of data preprocessing techniques by performing any of the following, as needed, according to the chosen dataset: (minimum 2)

   a. Perform data cleaning, transformation, discretization, and normalization.
   b. Select and justify data sampling techniques.
   c. Remove any irrelevant data or outliers.
   d. Feature selection and feature engineering.

   (Maximum 500 words, approximately 2 pages)

4. **Data Mining Methods (40 marks)**
   - Develop data mining models to achieve the objective in Task 1. (Minimum 3 models)
   - Evaluate the data mining models to ensure that the model is accurate and reliable.
   - Perform hyperparameter tuning or revisit the preprocessing task to improve the models' performances.
   - Discuss the results.

   (Maximum 500 words, approximately 2 pages)

5. **Presentation (15 marks)**
   - Presents the analysis in a clear, organized, and engaging manner.
   - Uses appropriate language, tone, and style for the intended audience.
   - Follows appropriate formatting and structure for the presentation.

Deliverables

Your submission should include the following:

- The cover page.
- The marking rubric.
- A Jupyter notebook (or equivalent) with your code, comments, and visualizations.
- A report based on the outlined tasks, summarizing your findings, including a discussion of your approach, key insights, and any limitations or areas for further exploration. Submission file format: PDF

Report format:

- **Paper size:** A4
- **Font type & size:** Times New Roman, 12 pt
- **Line spacing:** 1.5 lines
- **Margins:** Top = Bottom = Left = Right = 1"
- **Full justify**
- **Page numbering** at the bottom of each page, except the cover page

Submission Guidelines

- Submit your assignment as a compressed file (zip or tar) that includes the Jupyter notebook and report. Name the file using your group number (e.g., Group1.zip). Submit your assignment through the course's WBLE. The deadline for submission 28 August 2023. Late submissions will not be accepted.

Marking Rubric

The assignment will be graded out of 100 points. The final score will be 30%, which is 10% for each CLO. The marks are based on the following criteria, as outlined in the tasks:

CLO	Criteria	Poor	Satisfactory	Excellent	Marks
4	**Data Understanding (15 marks)** • Identifies the dataset and describes its source (5m) • Summarizes the variables in the dataset (5m) • Identify the objective of analyzing the dataset, and potential issues or limitations with the dataset (5m)				
2	**Data Description (15 marks)** • Identify 3 questions to be addressed using the dataset.(5m) • Use visual methods to find the key features, patterns, and trends in the data (minimum 3 visualizations). (5m) • Elaborate on the findings in the report.(5m)				

2	**Data Preprocessing (15 marks)**
	• Demonstrate understanding of data preprocessing techniques by performing any of the following, as needed, according to the chosen dataset: (minimum 2)
	  a. Perform data cleaning, transformation, discretization, and normalization.
	  b. Select and justify data sampling techniques.
	  c. Remove any irrelevant data or outliers.
	  d. Feature selection and feature engineering.				

3	**Data Mining Methods (40 marks)** • Develop data mining models to achieve the objective in Task 1. (Minimum 3 models) (5m) • Evaluate the data mining models to ensure that the model is accurate and reliable. (5m)				

	• Perform hyperparameter tuning or revisit the preprocessing task to improve the models' performances. (10m) • Discuss the results. (20m)				
4	**Presentation (15 marks)** • Presents the analysis in a clear, organized, and engaging manner (5m) • Uses appropriate language, tone, and style for the intended audience (5m) • Follows appropriate formatting and structure for the presentation (5m)				