**WIE3007 Data Mining & Warehousing 2023/2024 Semester 1**

**This is a group project. Each group should consist of 4 or 5 members.**

**Assignment Overview:**

Choose a dataset from the UC Irvine Machine Learning Repository (<http://archive.ics.uci.edu/ml/>) or any other source of your choosing. The primary focus of this assignment is to apply data mining methodology in real-world applications, focusing on SAS SEMMA methodology:

**Assignment Instructions:**

1. **Report Writing (10 Marks):**

Below are step-by-step instructions for the student to follow:

Step 1: Select a Dataset

Visit the UC Irvine Machine Learning Repository or another source of your choosing.

Browse through the available datasets and select one that interests you and is suitable for data mining.

Step 2: Understand the Dataset

Read any available documentation or descriptions of the dataset to understand its structure, features, and purpose.

Identify the variables and types of data included in the dataset.

Step 3: Apply SAS SEMMA Methodology

The SAS SEMMA methodology consists of five stages: Sample, Explore, Modify, Model, and Assess.

3.1 Sample

Extract a subset of the dataset for preliminary analysis, ensuring it is representative of the entire dataset.

Define the target variable, if applicable.

3.2 Explore

Perform exploratory data analysis (EDA) to understand the characteristics and patterns within the data.

Use visualizations such as histograms, box plots, and scatter plots to explore the distribution and relationships between variables.

3.3 Modify

Preprocess the data by handling missing values, outliers, and encoding categorical variables.

Perform feature engineering to create new features or transform existing ones.

Normalize or standardize the data if necessary.

3.4 Model

Choose appropriate data mining or machine learning models based on the problem at hand.

Train the models using the modified dataset.

Optimize model parameters to improve performance.

3.5 Assess

Evaluate the performance of the models using appropriate metrics such as accuracy, precision, recall, F1 score, or mean squared error.

Interpret the results and draw conclusions about the dataset and models.

Consider the implications and limitations of the models.

Step 4: Document Your Work

Clearly document each step of the SAS SEMMA methodology, including any decisions made, methods used, and results obtained.

Include visualizations, code snippets, and explanations to support your findings.

Write a conclusion summarizing the main insights gained from the analysis and any recommendations or future work.

Step 5: Submit Your Assignment

Review your work to ensure it meets the assignment requirements and is clear, concise, and error-free.

**2. Presentation Slides (5 Marks):**

Create concise and clear presentation slides summarizing your findings. Each member is required to give a 2-minute recorded presentation on their respective contributions or findings. During the presentation, ensure that your **name** and **picture** are clearly visible. If a member does not present in the recorded session, no marks will be awarded to that individual for this component.

**Hints:**

While the above points are crucial, you are not limited to them. Feel free to incorporate any additional analysis or insights you deem necessary.

**Submission Requirements:**

* A well-structured report covering the aforementioned points.
* Presentation slides summarizing the key findings of your analysis.
* Make sure to cite the data source and any additional resources used.
* This will clarify that the submission should come from the group leader only.

**Evaluation Criteria:**

* Clarity and structure of the report and slides.
* Depth and quality of analysis and exploration.
* Adequacy of graphical representations in exploring the data source.
* Alignment of analysis with the stated goal.

**Deadline:**

Week 14: 17 Jan 2024 , 6pm.