**DATA MANAGEMENT**

**INDIVIDUAL ASSIGNMENT**

**LEARNING OUTCOMES**

1. Evaluate the various data types, data storage systems and associated techniques for indexing and retrieving data.
2. Design feature engineering techniques to transform transactional data into meaningful inputs in order to create a predictive model.
3. Propose a suitable approach to designing a data warehouse to store and process large datasets.

**DATA MANAGEMENT**

The machine learning pipeline involves several tasks before the development of a predictive/descriptive models. The inevitable and vital process includes preparing and understanding the data. Moreover, the performance of the predictive/descriptive model depends on the choice of pre-processing techniques.

For the assignment, you are required to prepare and explore the given dataset. It is imperative to explain and justify the pre-processing, transformation, and feature engineering techniques that have been chosen. Your analysis should be deep and in detail, also it must go further than what has already been covered in this course.

The assignment should involve a number of experiments, and a detailed exploration and analysis of the results using **SAS Studio**.

**You need to do the following tasks:**

**PART 1**

**Feature Engineering**

Several Data Mining/Machine Learning algorithms are designed to work with qualitative or quantitative data and very few algorithms support mixed data. Hence, this task requires you to transform with an appropriate method(s) and proper justification to be provided. In addition to that, the **metadata** should be created for each dataset. Feature engineering itself can be divided in 2 steps:

* Variable transformation.
* Variable / Feature creation.

In this section , you need to summarize feature engineering task and provide the interpretation of work related to feature engineering task that you have done in SAS Studio.**[1000 words].**

**Documentation Format:**

* + Typeface: Times New Roman. Boldface, italic & lines can be used for emphasizing and to enhance readability.
  + Font size: 12 (except titles and headings).
  + Margins: 1” from the left, right, top & bottom of the edges of the A4 paper.
  + Spacing: 1.5 lines between texts of a paragraph.
  + Alignment: Justify.
  + Headers and footers can be used all pages must be numbered accordingly.
  + Standard cover page as available in the learning management system

**PART 2**

1. **Related Works**

In this section, you are supposed to research and present the other works related to the application domain.

**Initial Data Exploration –**

Data exploration is an approach similar to initial data analysis, whereby a data analyst uses visual exploration to understand what is in a dataset and the characteristics of the data, rather than through traditional data management systems.

This section should contain the following task.

* Indicate the type of each attribute (nominal, ordinal, interval or ratio).
* Identify the values of the summarising properties for each attribute including frequency and spread e.g. value ranges of the attributes, frequency of values, distributions, medians, means, variances, and percentiles. Wherever necessary, use proper visualisations for the corresponding statistics. Summary / descriptive stats
* Using SAS explore your dataset and identify the variables any outliers, missing values, and outliers treatment.

1. **Data Pre-processing**

Investigate the required method(s) to handle the incomplete, noisy and inconsistent data.

Report each of the applied techniques with detailed explanations. Show your results and justify your approach.

**NOTE: Easiest way to handle dirty data is through removing the feature(s) / instance(s). Choosing this method will be award ZERO for pre-processing.**

1. **Exploratory Data Analysis (EDA) – graph**

This task requires you to perform an analysis on the datasets generated during your feature engineering. **Exploratory Data Analysis** (**EDA**) can be defined as the numerical and graphical examination of **data** characteristics and relationships before formal, rigorous **statistical** analyses are applied. You are evaluated based on the approaches undertaken to get familiar with the dataset.

1. **Hypothesis**

Formulate a minimum of **FIVE (5) hypotheses** based on the dataset (cleaned dataset or transformed dataset) with required analytical variable(s). Interpret the hypotheses with the query resulted from **Query and visualization using SAS Studio.**

**Deliverables**

The deliveries include:

* A report, which structure should follow the tasks of the assignment.
* **SAS** program (Initial Data Exploration, Data Pre-processing, and Dataset Transformation) and queries with an individual file for each task.

Your report should include the following:

**Abstract** – A self-contained, short, and powerful statement/brief that describes your work. It may contain the scope, purpose, results, and contents of the work. **[150 to 200 words]**

**Introduction** - The purpose of your report. Background information about the topic. You also have to place some brief details of your methods applied for the study. Include an outline of the structure of the report. **[500 to 800 words]**

**Related Work** - Carefully structure your findings. It may be useful to do a chronological format where you discuss from the earliest to the latest research, placing your research appropriately in the chronology. Alternately, you could write in a thematic way, outlining the various themes that you discovered in the research regarding the topic. **[800- 1000 words]**

**Method** - This section should contain detail exploration of the dataset, pre-processing, feature engineering, EDA and Hypothesis. **[No limit]**

**Discussion -** For each of the task include a section title in your report. Finally, you need to summarize your findings, and this summary section should **NOT** be a narrative of your tasks, but a summarized informative section of what is your findings of the data. This section should provide detail interpretation of the work along with the supporting related works. **[500 to 1000 words]**

For example, it should include details like specific characteristics (or values) of some attributes, important information about the distributions, relationship or association that exist between variables found that should be investigated more rigorously, etc.

**Conclusion** – In this section, you need to state your position about what you gained in this assignment that can contribute to other readers.