Institute of Business Administration
CSE-464: Introduction to Data Mining
Assignment#03
Due Date: 4th December 2022 (11:55 PM)

Instructions

- You are required to participate in the challenge posted on the Kaggle website. The link to the Kaggle challenge is copied below:
  https://www.kaggle.com/t/edc4603084184567ab968eff85ec0759
- The challenge can be implemented using KINME or Python.
- **This is an individual assignment. Plagiarism in workflow/code or reports will lead to ZERO marks.**
- You are expected to use your proper full name. This is important because many people have similar first names. Any entry with an indistinguishable name/username (like warrior, beast, blackbeauty, etc.) would be removed.
- Your task is to try following clustering algorithms (with different variations/parameters) along with the data preprocessing, cleaning, and transformation techniques to increase the score.
  - K-means Clustering
  - Agglomerative Clustering

  Also, discuss the impact of the following methods for identifying the optimal number of clusters in the data.
  - Elbow Curve
  - Silhouette Plot
  - Dendrogram
- You must also submit a detailed report (Microsoft Word Document) describing the data preparation and algorithms that you attempted and especially identify the things that worked for you. The details are explained below.

Assignment Milestones

The assignment carries 08 marks which will be divided as follows:

1. ***Kaggle Participation:***

   Your Kaggle score + the number of entries spread over the period of the competition and not just on the last day.

2. ***Report (Word Document)***

   The challenge report with the following details:

   - The report must detail the **CRISP-DM approach.** Explain each step of the process in detail.
   - A summary (in the form of a table) of your different attempts performed for the entries that you submitted on the Kaggle with proper reasoning for each attempt/entry.

    - Example: The inclusion of an important categorical column improved the performance of the model from 0.5 to 0.7 because this column in the dataset for the given domain largely impacts the target, and so on.
    - If you made 35 entries on Kaggle then this table must contain 35 rows (1 for each entry)
    - You can add as many columns as you want to distinguish the various attempts. The three major columns include Data preprocessing, Model Details, and Score.
- Lastly, the report must also include your overall findings and insights including but not limited to the following:
    - Which algorithm worked best for the given dataset and why?
    - What is the optimal number of clusters in the data as per your findings and why?
    - What were the overall challenges that you faced while improving the score, and so on?

3. Implementation:

Python notebook or KNIME workflow demonstrating your best entry on LMS. Please note that the submitted file on execution must show the same score as visible on Kaggle.

NOTE:

- **Submit a zip folder containing both the required files (Report & Implementation) on LMS.**
- **Both the modules (Kaggle participation and LMS submission) mentioned above must be submitted within the allotted duration. The submissions with any missing module will not be considered for grading.**