**Business Intelligent and Data Mining**

**2023 Fall Semester - Midterm**

1. Please refer to the attached data file. The data in the file includes information on whether various new and used cars are worth buying, sourced from evaluations on the internet. There are seven variables in the file (buying\_price, maint\_price, persons, doors, lug\_boot size, safety, Suggestion). The definitions and attribute value explanations for each variable are as follows:"

|  |  |  |
| --- | --- | --- |
| **Variable** | **Name** | **Definition** |
| X1 | buying price | Buying Price |
| X2 | maint price | Maintenance Cost |
| X3 | Persons | Seating Capacity |
| X4 | doors | Number of Doors |
| X5 | lug\_boot size | Trunk/Boot Size |
| X6 | safety | Safety |
| Y | Suggestion | Worthiness of Purchase |

Please analyze the following issues based on the data (you can expand as needed) and provide the analytical mechanism (logic and process of the analysis):

* Help choose an appropriate clustering method based on variables (X1 to X3), and explain the meaning of each cluster (e.g., Cluster 1: high buying price, high maintenance cost, low seating capacity).
* After clustering, please predict 'worthiness of purchase' (Y) using the LDA method based on all variables (X1 to X6) for each cluster data and evaluate the prediction results using a confusion matrix.
* Note: Remember to check if SMOTE (Synthetic Minority Over-sampling Technique) is needed!"

P.S. Please save your answers (descriptions, analysis screenshots, etc.) as a Word document, along with the data file, Weka files, or Python files, and place them in the directory 'StudentID\_Name' (e.g., 112578000\_Chih3c). Compress these files into a 'StudentID\_Name.rar' format (e.g., **112578000\_Chih3c.rar**), and send it to me at **chih3c@mail.ntut.edu.tw** before 12:00 AM on November 21th. Thank you and Good luck!