Introduction to Data Science

Midterm Project

Project :

Apply the appropriate data preparation steps to your chosen dataset and compute the relevant descriptive statistics. For this project, each group must work with its own selected dataset and modify it according to the project requirements. You may choose any dataset that fits the task.

Project Deliverables

- Submit the implemented R program (R file or Text file) in Teams. During VIVA session, you will bring this implemented program and we may ask you to execute the program.
- Submit the report in the Teams. See the instruction section below for the report details. **Please bring the printed copy of the submitted report during the VIVA session.**

Instructions

- The submission deadline for all deliverables is **December 2, 2025, by 7 am.**
- At the beginning of the report, write a short note about the dataset.
- For each implemented code segment in the R program, provide the code and its output along with their description in the report. In the description part, only write the content (do not write unnecessary content) that is sufficient to understand the code and its output.
- **Comments are not allowed in the R program**.
- The following topics can be focused to think about the project. **Note that the project is not limited to these topics which are mentioned to get an idea about how to proceed with the project.**
    - If there are any missing values in the dataset, we should apply all applicable methods from the available options to handle the missing values.
    - We can see missing values on a graph.
    - Detect outliers in the data set and use the appropriate approach to handle those values.
    - We can convert attributes from numeric to categorical or categorical to numeric.
    - We can apply the normalization method for any continuous attribute.
    - We can find and remove duplicate rows.
    - We can apply some filtering methods to filter the data.
    - Detect invalid data in the data set and use the appropriate approach to handle those values.
    - We can convert the imbalanced data set into a balanced data set.
    - Split the dataset for Training and Testing.
    - Calculate descriptive statistics and interpret the results for the numerical variables according to the target classes of your chosen data set.

- Compare the mean values of a selected numerical variable across two distinct categories of an appropriate categorical variable from your dataset.

- Examine and compare the variability (e.g., IQR,standard deviation, variance, or range,) of another numerical variable across the different categories or levels of a chosen categorical variable within your dataset.