Introduction to Data Science

Midterm Project – Spring 24-25

Project :

Apply suitable data preparation steps and calculate descriptive statistics for the given dataset. I have provided a dataset in MS Teams for this project.

Project Deliverables

- Submit the implemented R program ("ids_mid_project_group_XX.r" file) in MS Teams. Replace "XX" in the filename with your group number such as for group-1, the program file name will be "ids_mid_project_group_01.r." During the VIVA session, you will bring this program, and we may ask you to execute it.
- Submit the report ("ids_mid_project_group_XX.pdf" file) in MS Teams. See the instruction section below for the report details.

Instructions

- **The submission deadline for all deliverables is April 26, 2025,11:59 PM.**
- **Comments are not allowed in the submitted R program**.
- **I will announce the project VIVA schedule in MS Teams.**
- **Please do not copy content from any sources. It will be strictly handled.**
- At the beginning of the report (after the cover page), write a short note about the dataset. You will get the original dataset details from the link provided (for the dataset) in MS Teams.
- 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.
- The following topics can be focused on when thinking about the project. **Note that the project is not limited to these topics, which are mentioned to get an idea of 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. Do not remove them.
    - We can see missing values using an appropriate technique.
    - Detect outliers in the dataset and use the appropriate approach to handle those values.
    - We can convert attributes from numeric to categorical and categorical to numeric where necessary. Do it at most for two attributes.
    - We can apply the normalization method to any continuous attribute.
    - We can find and remove duplicate values.
    - We can apply some filtering methods to filter the data.
    - Detect invalid data in the dataset and use the appropriate approach to handle those values.
    - We can convert the imbalanced dataset into a balanced dataset.
    - Split the dataset for Training and Testing.
    - Compute the central tendencies (mean, median, mode) and interpret the results. Do it for any two numeric and two categorical attributes.
    - Compute the spread (Range, IQR, Variance, Standard Deviation) and interpret the results. Do it for any two attributes.