Introduction to Data Science

Midterm Project

Project :

Apply data preparation steps (which can be applied) and calculate descriptive statistics for the given data set. In this project, we are going to use a modified version of Diabetes Prediction Dataset which can be downloaded from the Teams. The original dataset can be found in the following link where the dataset description is available as well (you may need to log-in to download the dataset).

https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Project Deliverables

- Submit the implemented R program (R file or Text file) and updated **Text** dataset in the 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 **April 26, 2025 (you must submit the assignment before 6:59 AM).**
- At the beginning of the report, write a short note about the dataset. You will get the dataset details from the above link provided for 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 the balanced data set.
- Split the dataset for Training and Testing.
- Compare the central tendencies (mean, median, mode) of Age across different groups of Gender and interpret the results.
- Compare the central tendencies (mean, median, mode) of Age across hypertension and interpret the results.
- Compare the Spread (Range, IQR, Variance, Standard Deviation) of Age across different groups of Gender and interpret the results.