Introduction to Data Science

Midterm Project

Project 1:

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 Loan Approval Classification 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/taweilo/loan-approval-classification-data

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 **December 14, 2024** (you must submit the assignment before **11:59 PM**).
- 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.
    - We can convert the imbalanced data set into the balanced data set.
    - We can find and remove duplicate values.
    - We can apply some filtering methods to filter the data.
    - We can convert attributes from numeric to categorical or categorical to numeric.
    - We can apply the normalization method only for one attribute.
    - If any invalid data/outliers exist in the data set, use the appropriate approach to handle those values.

Project 2: Data Preprocessing Steps for Text Data

This project aims to develop a comprehensive and efficient data preprocessing for text data to enhance the performance of natural language processing (NLP) models. Preprocessing text data is a critical step in any NLP project, as it involves cleaning, transforming, and structuring raw text into a format that models can interpret and learn from effectively. This project will cover a range of preprocessing techniques tailored to text data, addressing challenges like noise, inconsistency, and redundancy. You need to collect text data from any sources using web scraping and perform the following data preprocessing steps.

Key Steps in Text Data Preprocessing:

1. **Text Cleaning**
2. **Tokenization:**
3. **Normalization**
4. **Stop words Removal**
5. **Stemming and Lemmatization**
6. **Handling Contractions**
7. **Handling Emojis and Emoticons**
8. **Spell Checking**