Project Report

Group Name: Team Sachin and Nishi

October 6, 2024

Team Member's Details

- **Group Name**: Team Sachin and Nishi
- **Name**: Nishi Gandhi
- **Email**: gandhi.nis@northeastern.edu
- **Country**: USA
- **College/Company**: Northeastern University
- **Specialization**: Data Science

Problem Description

The project aims to analyze healthcare data to predict patient outcomes, specifically focusing on whether a patient will continue treatment or not. The goal is to use machine learning models to identify factors influencing treatment persistence.

Data Understanding

The data consists of historical clinical records and demographic information of patients. The dataset contains approximately 3,605 features, each representing medical encounters, treatments, comorbidities, and other health-related factors.

What type of data you have got for analysis

The dataset consists of both numerical and categorical features. Numerical features include counts of medical encounters and exam frequencies, while categorical features include medical codes, encounter types, and patient demographics. The target variable is the **Persistency_Flag**, which indicates whether a patient continues treatment.

What are the problems in the data

Some of the key challenges identified in the dataset include:

- Missing values in several key features such as treatment frequencies and comorbidities.
- Presence of outliers in certain numerical features, particularly encounter frequencies.
- Skewed distributions in some features, which can affect model performance.

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

Several preprocessing steps are being employed to address the data issues:

- **Imputation techniques** are used to fill missing values. For numerical features, the median is used to avoid skewing the data, while for categorical features, the mode is imputed.
- **Outliers** are being handled by capping extreme values based on the interquartile range (IQR) method.
- To handle **skewed data**, a log transformation is being applied to normalize distributions.

Github Repo Link

The project code and analysis can be accessed on the following GitHub repository:
https://github.com/Sachinsub0/DG$_Healthcare_Project$