Project Report with EDA

Group Name: DG$_{H}ealthcare$

October 6, 2024

Team Member's Details

- **Group Name**: DG Healthcare
- **Name**: Nishi Gandhi and Sachin Subramanian
- **Email**: gandhi.nis@northeastern.edu, subramanian@ucdavis.edu
- **Country**: USA
- **College/Company**: Northeastern University
- **Specialization**: Data Science

Problem Description

The project focuses on analyzing healthcare data to predict patient outcomes, specifically whether a patient will continue treatment or not. Using machine learning models, we aim to uncover the most significant factors influencing treatment persistence and improve the decision-making process in healthcare services.

Github Repo Link

The project repository can be accessed here: https://github.com/Nishi-Gandhi/DG$_{H}ealthcareassignments$

EDA Performed on the Data

Exploratory Data Analysis (EDA) has been performed to understand the structure of the dataset. Key tasks include:

- **Data Cleaning**: Missing values were handled using imputation methods.
- **Outlier Detection**: Outliers were identified and capped using IQR-based methods.
- **Distribution Analysis**: Histograms and boxplots were generated to understand the distribution and skewness of numeric features.
- **Correlation Analysis**: A correlation matrix was produced to examine relationships between features.
- **Feature Engineering**: New variables were created based on domain knowledge to improve model performance.

Final Recommendation

Based on the EDA and insights from the data, it is recommended to implement a **Random Forest** model to predict patient treatment persistence. This model is effective in handling a large number of features and can capture non-linear relationships in the data.