Data Cleansing & Transformation - Week 9

Team Member's Details

Group Name: The Closer

Member 1 Name: Rongala Sreedhar

Member 1 Email: rongalasreedhar@gmail.com

Member 1 Country: Italy

Member 1 College/Company: University of Naples Federico II

Member 1 Specialization: Data Science

Problem Description

The objective of this task is to perform Data Cleansing and Transformation on the Healthcare dataset. The data contains patient records which may have inconsistencies, missing values, or outliers. Our goal is to apply multiple techniques to clean this data and also demonstrate NLP featurization.

GitHub Repo Link

https://github.com/sreedharsiddhu/Data-Glacier

Data Cleansing and Transformation

We applied the following cleansing techniques (demonstrated in the attached Notebook):

1. Handling Missing Values:
   - Approach A (Member 1): Simple Imputation. We used the Median for skewed numeric columns and Mode for categorical columns. This is a robust baseline.
   - Approach B (Member 2): Model-Based Imputation (KNN). We used K-Nearest Neighbors to estimate missing values based on similar records, preserving local structure.

2. Handling Outliers:
   - Approach A: Removal using IQR (Interquartile Range). Any data point falling below Q1-1.5*IQR or above Q3+1.5*IQR was removed.
   - Approach B: Log Transformation. Instead of removing data, we applied a log transformation to compress the scale of outliers.

3. NLP Featurization (Column: Ntm_Speciality):

- Cleaning: Lowercasing and removing punctuation using Regex.

- Featurization: Applied CountVectorizer (Bag of Words) and TF-IDF to convert text data into numerical features.

Peer Review Comments

Reviewer: Peer Member (Simulated)

- "The use of Median imputation is appropriate for 'Count_Of_Risks' due to its discrete nature."
- "IQR outlier removal significantly reduced the dataset size; Log transformation (Approach 2) might be safer for this small dataset."
- "Regex cleaning successfully standardized the 'Speciality' column."