EDA and Preprocessing Report

Esma Nur Arslan
esmanurarslan2016@gmail.com

Physical Medicine & Rehabilitation Dataset: EDA and Preprocessing Report

1. Introduction

The objective of this project was to perform a comprehensive exploratory data analysis (EDA) and prepare a dataset from a physical medicine and rehabilitation clinic for a potential predictive modeling task. The primary goal was to make the data model-ready for predicting the target variable, **TedaviSuresi** (Treatment Duration in Sessions). This report details the steps taken, from initial data inspection and cleaning to advanced feature engineering and encoding.

2. Exploratory Data Analysis (EDA)

The analysis began with an initial inspection of the dataset, which contained 2235 rows and 13 columns.

2.1. Initial Findings & Data Quality Issues

- **Duplicate Data**: A significant data quality issue was the presence of **928 completely duplicate rows**. These were removed to prevent bias in analysis and modeling.
- **Inconsistent Formatting**: Numerical columns like TedaviSuresi and UygulamaSuresi contained text suffixes ("Seans", "Dakika"), requiring cleaning before they could be used as numbers.
- **Messy Text Data**: Several key categorical columns (KronikHastalik, Alerji, Tanilar, TedaviAdi, UygulamaYerleri) contained comma-separated lists with numerous inconsistencies, including typos, abbreviations, and varied terminology.
- **Missing Values**: Missing data was identified in several columns, including Cinsiyet, KanGrubu, Alerji, and KronikHastalik.
- **Correlations**: An initial correlation heatmap of the raw numerical features showed very weak linear relationships, indicating that robust feature engineering would be crucial.

3. Data Preprocessing & Feature Engineering

Based on the EDA findings, a multi-stage preprocessing pipeline was implemented to clean and enrich the dataset.

3.1. General Cleaning

- **Column Renaming**: All column names were converted to snake_case for consistency.
- **Type Conversion**: After removing text suffixes, tedavi_seans_sayisi (target) and seans_suresi_dk were converted to integer types.
- **Lowercasing**: All textual data was converted to lowercase.

3.2. Feature-Specific Cleaning and Engineering

- **Alerji & KronikHastalik**:
  - Missing values were logically filled with **'Yok'** (None).

    - The comma-separated strings were split into lists of individual conditions.
    - Typos were corrected and related conditions were grouped under a standard name.
- **Tanilar (Diagnoses)**:
    - This column underwent extensive cleaning to standardize text.
    - A new, highly relevant feature, **SGK_Tani_Grubu**, was engineered. This process was rigorously guided by the official **ICD-10 Physical Therapy and Rehabilitation Diagnosis List (Annex EK-4/D)**. Diagnoses were systematically mapped and classified into their official treatment groups (A, B, C, D). This methodology ensures that the grouping is not arbitrary but is based on established clinical and administrative standards, directly reflecting the expected complexity and reimbursement levels for different conditions.
- **TedaviAdi (Treatment Name)**:
    - This was the most complex feature. A robust function was developed to:
        1. Correct common abbreviations ("öçb" -> "ön çapraz bağ") and typos.
        2. Standardize treatment terminology ("ftr" -> "rehabilitasyonu").
        3. Use regular expressions to parse and extract directionality (e.g., (sağ), (sol), (bilateral)) and treatment phases (e.g., (evre 1)).
    - A simplified version (tedavi_cleaned_bert) was created for the NLP model.
- **UygulamaYerleri (Application Sites)**:
    - The list of body parts was cleaned and standardized.
    - Missing values were imputed by a rule-based system that searched for body-part keywords within the corresponding tedavi_cleaned entry.

4. Preparation for Modeling

To make the data suitable for machine learning algorithms, it was split and encoded.

4.1. Train-Test Split

The dataset was split into **80% training** and **20% testing** sets. This split was performed **before** applying any encoding methods to prevent data leakage.

4.2. Advanced Encoding Strategy

A variety of encoding techniques were employed:

- **One-Hot Encoding**: Applied to simple categorical features like cinsiyet.
- **Multi-Label Binarization**: Used for list-based features like kronik_hastalik_cleaned.
- **BERT Embeddings + HDBSCAN Clustering**: To capture the rich semantic meaning of the TedaviAdi feature, an advanced NLP technique was used.
    1. **Embeddings**: A Turkish BioBERT model converted each cleaned treatment name into a high-dimensional numerical vector.
    2. **Clustering**: The HDBSCAN algorithm was trained on these embeddings to group similar treatments into clusters, converting the complex text feature into a single, meaningful categorical feature (tedavi_hdbscan_label).

5. Conclusion

Through a systematic process of EDA, cleaning, and advanced feature engineering—including the application of official ICD-10 standards for diagnosis classification—the raw dataset has been transformed into a fully numerical, high-quality format. The data is now ready for the development of robust predictive models.