Data Science Intern Case Study – EDA & Preprocessing Report

Pınar Gökhan – pinar.gkhn1@gmail.com

1. Exploratory Data Analysis (EDA) Findings

- **Dataset shape**: 2235 observations, 13 columns.
- **Column check**: All expected fields (HastaNo, Yas, Cinsiyet, KanGrubu, Uyruk, KronikHastalik, Bolum, Alerji, Tanilar, TedaviAdi, TedaviSuresi, UygulamaYerleri, UygulamaSuresi) are present. No extra columns.
- **Missing values**: Some categorical variables (KanGrubu, Alerji, KronikHastalik) have significant missingness. Numerical fields like *Yas* and *HastaNo* are fully complete.
- **Duplicates**: A few potential duplicates were detected (same patient ID, treatment, and application site).
- **Target variable (TedaviSuresi)**: Stored as strings (e.g. "15 Seans"), requires numeric extraction. Distribution is right-skewed with common values around 5, 10, and 15 sessions.
- **Numerical correlations**: Weak correlations between age and treatment duration. Heatmap revealed limited relationships between numeric columns (*Yas*, derived durations).
- **Categorical insights**:
  - Gender mostly balanced but with inconsistent labels ("Kadın / KADIN / Kadin").
  - Blood group contains "0 Rh+" notation that should be normalized to "O Rh+".
  - *Bolum* contains multi-department entries; primary department extraction is needed.
- **Multi-valued fields**: *KronikHastalik, Alerji, Tanilar, UygulamaYerleri* include multiple comma/semicolon-separated values that need splitting.

2. Data Preprocessing Steps

1. **Numeric extraction**
   - Converted TedaviSuresi → TedaviSuresi_num (number of sessions).
   - Converted UygulamaSuresi → UygulamaSuresi_num (minutes).
2. **Text normalization**
   - Trimmed spaces, unified casing, normalized Turkish characters.
   - Gender mapped to standardized labels ("Kadin", "Erkek").
   - Blood group "0 Rh+" fixed to "O Rh+".
3. **Feature engineering**
   - Extracted primary department (Bolum_Primary).
   - Split multi-valued fields into lists and created count features (*_count).
   - Built multi-hot encoded columns for top 15 frequent conditions, allergies, diagnoses, and application sites.
4. **Handling missing values**
   - Numerical features imputed with median.
   - Categorical features imputed with most frequent.
   - List-type columns converted to empty lists where missing.

5. **Outlier treatment**
    - Applied IQR capping on “TedaviSuresi_num, UygulamaSuresi_num”, and “Yas” to reduce extreme values.
6. **Deduplication**
    - Identified potential duplicate rows (same HastaNo, TedaviAdi, UygulamaYerleri). Left removal as optional.
7. **Final datasets**
    - Saved “model_ready.csv” after cleaning.
    - Constructed processed feature matrix (model_matrix.csv and .parquet) using ColumnTransformer with:
        - StandardScaler for numeric features.
        - OneHotEncoder for categorical features.
    - Verified all features are numeric and target has no missing values.