Pusula Talent Academy Case Study

Name and Surname: Murathan Elcuman
Email Address: mrthelcuman@gamil.com

1. Introduction and Goal

This document summarizes the data analysis and preprocessing performed in the Python Jupyter Notebook. The objective of the notebook is to analyze a patient dataset to uncover insights and prepare it for a machine learning model. The analysis focuses on understanding the relationships between patient demographics, medical history, and treatment plans, with the ultimate goal of transforming the raw data into a clean, structured feature set ready for model training. The target variable for a potential predictive model is Treatment Duration.

2. Data Loading and Sanity Check

The analysis begins by loading the dataset file into a Pandas DataFrame. An initial sanity check is performed to understand the dataset's structure and quality.

Shape and Info

- The dataset contains 2235 rows and 13 columns. **df.info()** reveals that most columns are of **object** data type and, with **Hastano** and **Yas** being numerical (**int64**)

Missing Values

- Several columns have a significant number of missing values. The highest percentages are:
  - **Alerji**: 42.2%
  - **KanGrubu**: 30.2%
  - **KronikHastalik**: 27.3%
  - **UygulamaYerleri**: 9.9%

Duplicates

- The notebook identifies 928 duplicate rows in the dataset.

Data Type Issues

- The TedaviSuresi and UygulamaSuresi columns, which represent durations, are stored as strings. These require cleaning to be used as numerical features.

3. Exploratory Data Analysis (EDA)

EDA was conducted to explore the data's underlying distributions and relationships.

Numerical Data Analysis

- Histograms and boxplots for Yas show a wide distribution, ranging from 2 to 92, with a mean age of approximately 47. Some outliers are present in the upper age range.
- A correlation heatmap **df.corr()** reveals very weak linear correlations between the numerical features Yas, HastaNo, TedaviSuresi, UygulamaSuresi.

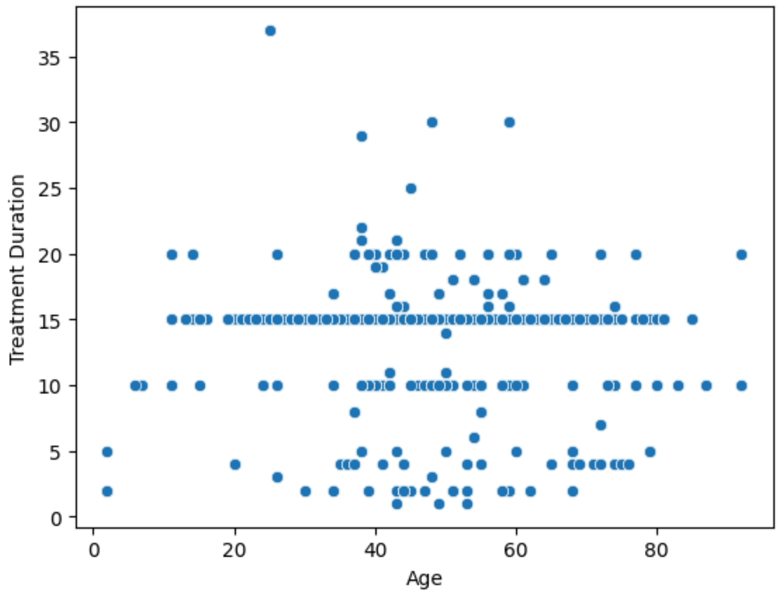

Figure 1 Age - Treatment Duration Graph

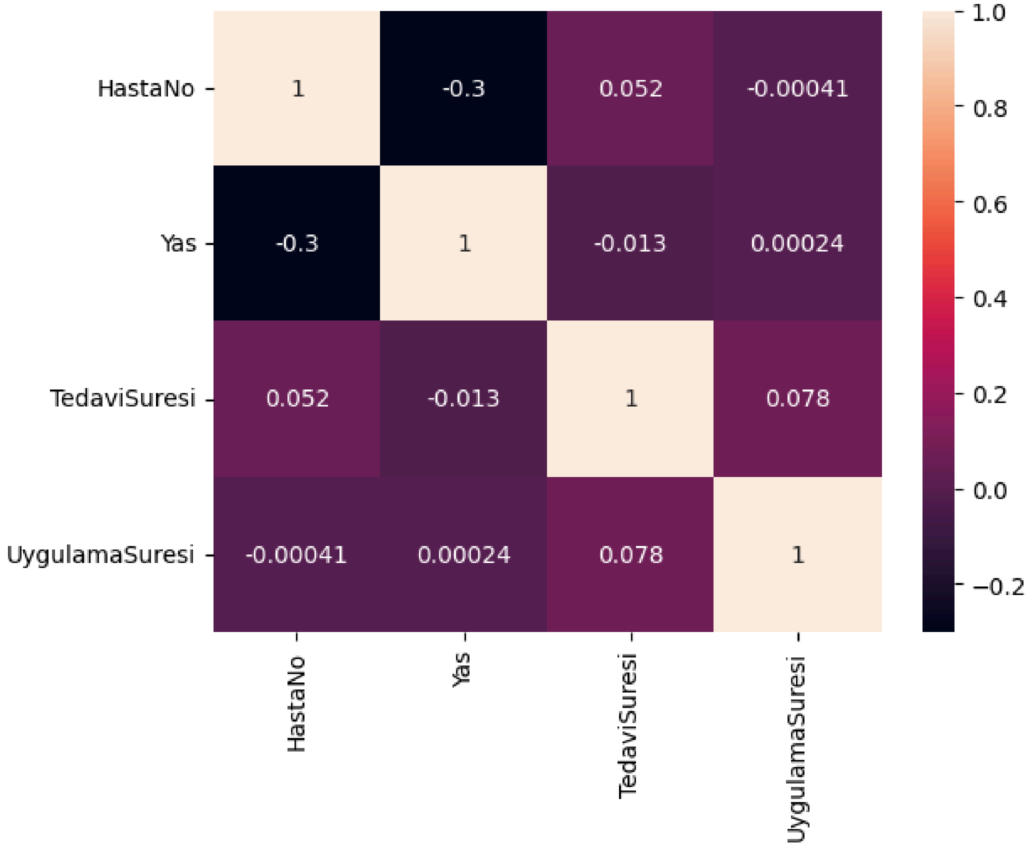

Figure 2 Heatmap

Categorical Data Analysis

- **Distributions: value_counts()** shows that the patient population is predominantly female and from Turkey.

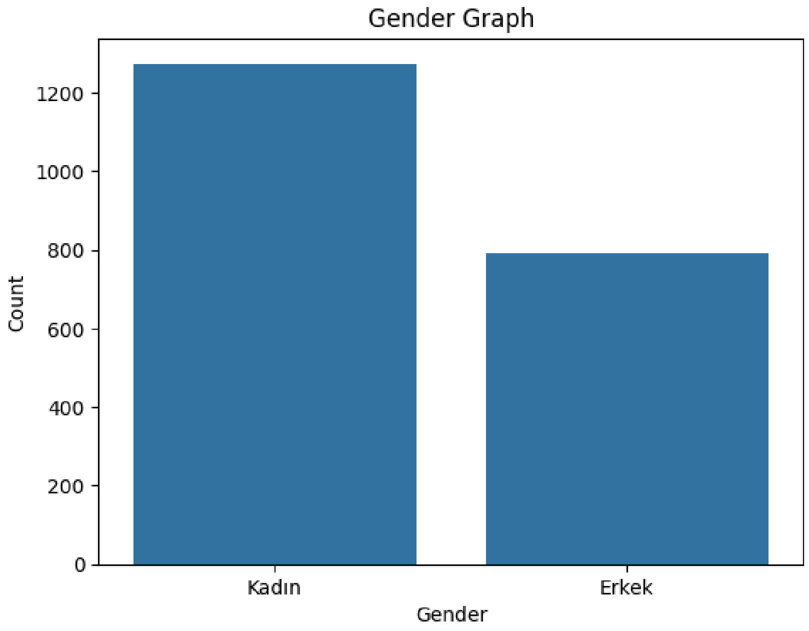

Figure 3 Gender Ratio

- **Relationships:** Violin plots and boxplots were used to examine relationships between categorical features and numerical targets. For instance, the relationship between Gender and Treatment Duration was explored, as was the relationship between Chronic Disease and Age.

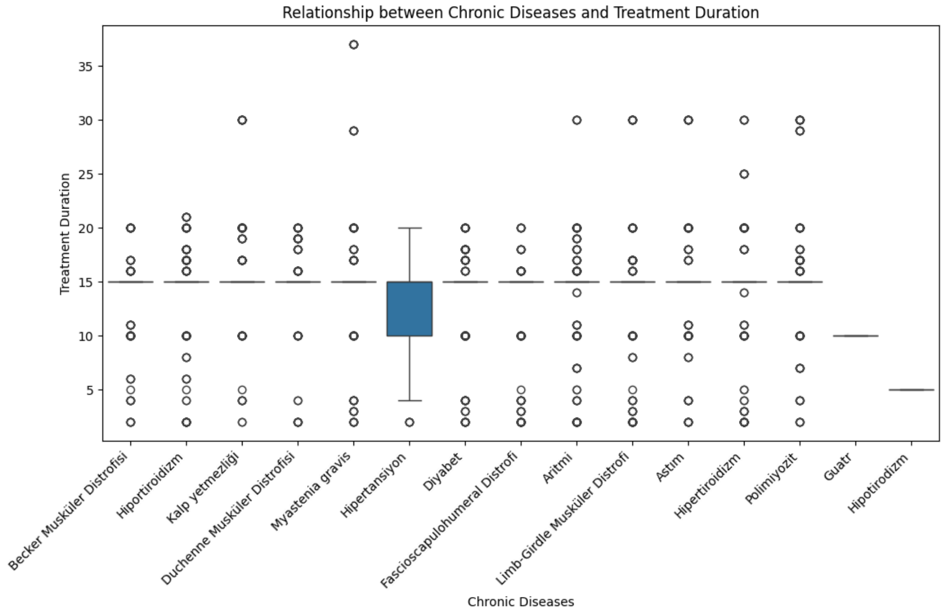

Figure 4 Treatment Duration - Chronic Disease Graph

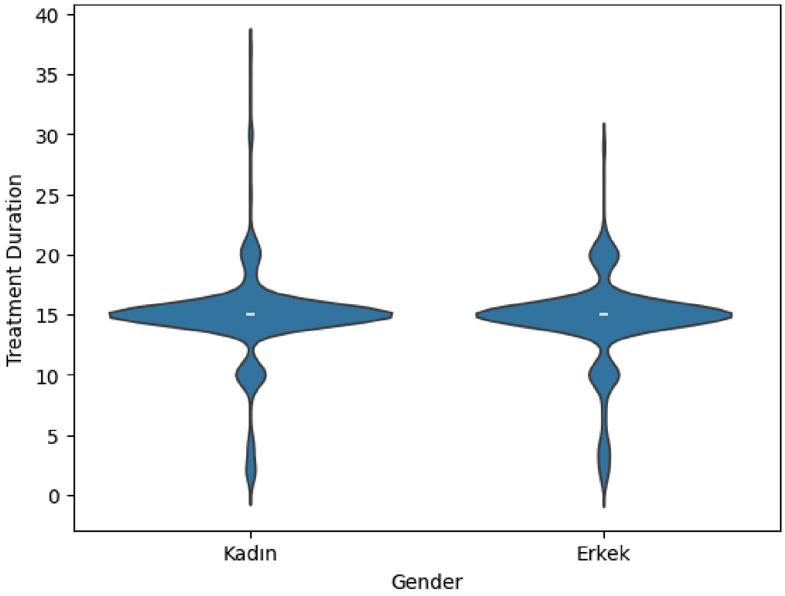

Figure 5 Treatment Duration - Gender Graph

4. Data Preprocessing and Feature Engineering

This stage transforms the raw data into a clean and structured format suitable for machine learning.

Handling Missing Values

- For single-label categorical columns, NaN values were filled with the string Unknown.
- For multi-label columns, NaN values were filled with None to signify the absence of a condition, allergy, or chronic disease.

Data Type Conversion

- The numerical values were extracted from the treatment duration and application duration columns, and the columns were converted to a float type.

Categorical Feature Encoding

- **Multi-Label Encoding**: Chronic Disease and Allergy often contain multiple comma-separated values. These were processed using **Multi-Hot Encoding**, which creates a new binary column for each unique disease or allergy.
- **Single-Label Encoding**: Other categorical columns were transformed using **One-Hot Encoding** to create binary columns for each category.

Numerical Feature Scaling

- The numerical columns Age and Application Duration were scaled using StandardScaler. This standardizes the features by removing the mean and scaling to unit variance, preventing features with larger scales from dominating the model.

5. Final Output

The notebook concludes by creating the final feature matrix X and the target vector y.

- **X**: A concatenated DataFrame containing the scaled numerical features and the encoded categorical features.
- **y**: The treatment duration column, which serves as the target variable for prediction