**Data Scientist Case Study Documentation**

**Name**: Hazar Kağan

**Surname:** Elbeyli

**Email:**[**elbeylihazar@gmail.com**](mailto:elbeylihazar@gmail.com)

**Date: 21.09.2024**

1. **Overview of the Dataset**

The dataset used in this project consists of patient information, including drug side effects and other personal health details. The dataset has 19 columns, such as:

* **Kullanici\_id**: Unique identifier for each patient.
* **Cinsiyet**: Gender of the patient.
* **Dogum\_Tarihi**: Date of birth of the patient.
* **Ilac\_Adi**: Name of the drug used.
* **Ilac\_Baslangic\_Tarihi**: The start date of drug usage.
* **Ilac\_Bitis\_Tarihi**: The end date of drug usage.
* **Yan\_Etki**: Side effects reported by the patient.
* **Kilo**: Patient's weight.
* **Boy**: Patient's height.

The data has both numerical and categorical variables, and there are missing values present in certain fields.

**2. Handling Missing Data**

Missing data was identified during the exploratory data analysis (EDA). Here’s how missing values were handled:

* **Date Fields**: Missing or invalid dates in columns such as Dogum\_Tarihi, Ilac\_Baslangic\_Tarihi, and Ilac\_Bitis\_Tarihi were converted to NaT (Not a Time) using pd.to\_datetime with the errors='coerce' parameter. This ensures that the date fields are consistent and errors are handled gracefully.
* **Other Fields**: Missing values in other fields such as Kilo, Boy, and categorical fields were handled using methods such as imputation (mean for numerical data or mode for categorical data) or exclusion from analysis, depending on the context.

**3. Data Types and Transformations**

Several transformations were performed on the dataset:

* **Date Conversion**: Fields like Dogum\_Tarihi, Ilac\_Baslangic\_Tarihi, and Ilac\_Bitis\_Tarihi were converted to the datetime format to allow for age calculation and timeline analysis.
* **Age Calculation**: A new column Yas was derived by calculating the age of each patient at the time of starting the drug. The formula used was:

df['Yas'] = (df['Ilac\_Baslangic\_Tarihi'] - df['Dogum\_Tarihi']).dt.days / 365.25

**4. Exploratory Data Analysis (EDA)**

The exploratory analysis helped in understanding the distribution and relationships within the data. Here are some of the findings:

* **Numeric Data**: Descriptive statistics were generated for numeric columns like Kilo and Boy. The describe()function was used to summarize the mean, median, and range of these variables.
* **Categorical Data**: Categorical columns such as Cinsiyet, Uyruk, and Yan\_Etki were analyzed using countplot to observe their distributions.
* **Visualization**: Several key visualizations were created:
  + **Histograms**: For numeric data to observe the distribution of values such as weight (Kilo).
  + **Correlation Heatmap**: A heatmap was generated to show correlations between numeric variables.
  + **Scatter Plot**: A scatter plot was used to visualize the relationship between Yas (Age) and Kilo (Weight), but initially failed due to missing values in the Yas column, which was later addressed by computing age as described above.

**5. Data Preprocessing Steps**

To ensure the dataset was ready for analysis, several preprocessing steps were applied:

* **Data Cleaning**: Invalid or missing values were handled through conversion, imputation, or removal. For example, missing values in the Kilo column were filled with the column's mean value, while missing categorical values were imputed using the most frequent value in their respective columns.
* **Categorical Encoding**: Categorical variables such as Cinsiyet (Gender) and Uyruk (Nationality) were encoded using one-hot encoding, allowing them to be used in machine learning models.
* **Scaling**: The numeric data was scaled to ensure consistent units and ranges for features like Kilo and Boy. This was done using the StandardScaler from Scikit-learn to standardize the data by removing the mean and scaling it to unit variance.