Sağre Buse Tonkaz /busetonkazz@gmail.com

Data Preprocessing and Analysis Report

After importing the necessary libraries into the project, the first step was to load the dataset. Once the data was successfully imported, I checked its contents and explored the dataset dimensions. I also examined the data types for each column. Ideally, all columns should have 2,357 non-null values, indicating no missing data. However, I found that only eight columns met this condition, and the remaining columns had missing values. the columns such as allergies, chronic diseases, father's chronic diseases, mother's chronic diseases, and sibling's chronic diseases had missing data but given that not every patient or their family members may have chronic conditions, I deemed it reasonable to disregard this missing data. However, five columns contained missing values that required further attention. Of these, three were categorical variables, and two were numerical variables. Using the data.describe() function, I explored the statistical properties of the numerical variables, gaining insights into measures of central tendency and distribution. This helped me better understand the structure of the dataset and provided valuable information for choosing methods in the later stages, such as handling missing values, normalization, outlier detection, and model selection. For example, I observed that all numerical variables exhibited symmetric distributions and this played a crucial role in making decisions for data processing. After thoroughly investigating the variables with missing data, I decided to create new features by converting the datetime64 types columns into numerical ones before filling in the missing values. For instance, I defined a new feature called "drug usage duration" by calculating the difference between the start and end dates of drug usage. Similarly, I calculated the patient's age at the start of drug usage as "age at drug start," and the time between the start of the drug and the report of side effects as "side effect onset duration." After creating these new features, I removed the original date columns from the dataset. I then clearly defined the categorical and numerical columns and proceeded with outlier detection using the Interquartile Range (IQR) method. In the outlier detection step, I applied the IQR method with a multiplier of 1.5 and found no outliers in the numerical variables.

One of the challenges in this dataset was that, in cases where a patient or their family members had more than one chronic disease, these diseases were listed together in the same cell, separated by commas. To address this, I used the explode function to separate each disease into its own row, allowing for a more detailed analysis. Additionally, I created visualizations to analyze the relationships between numerical variables and represent categorical variables graphically. I confirmed that all numerical variables exhibited symmetric distributions, with the mean and median values being close to each other.