Can Berk Durak
canberk.cbd@gmail.com

Exploratory Data Analysis (EDA) for Pusula case

1. Introduction

This document summarizes the findings from the exploratory data analysis (EDA) conducted on the drug side effects dataset and outlines the steps taken for data preprocessing. The dataset contains information about drugs, side effects, patient demographics, allergies, and other relevant health-related information.

2. Data Overview

The dataset contains information about the patients to focus on side effects that they have faced.
Patient information : Gender, height, weight, city
Health related patient information: Chronic illness, blood type, allergies
Drug information: Drug name, drug start date, finish date
Side effects: Side effects, side effect informed date

3. Data Preprocessing

a. Handling Missing Data

The dataset was checked for missing values using the isnull() function, and the proportion of missing values in each column was calculated

b. Handling Invalid Categorical Values

For categorical variables, we checked for invalid or unexpected values:
Kan Grubu (Blood Type): The valid values were restricted to ['A+', 'A-', 'B+', 'B-', 'AB+', 'AB-', 'O+', 'O-']. Any deviations from this list were flagged as invalid.

Cinsiyet (Gender): The valid values were ['Kadın' (Female), 'Erkek' (Male)]. Any other entries were also flagged as invalid.

c. **Detecting Anomalies in Numerical Columns**

For numerical columns such as height, weight, and age, we checked for anomalies using the Interquartile Range (IQR) method: The IQR was used to detect outliers. Values outside 1.5 times the IQR were flagged as potential anomalies.

.

4. **EDA**

a. **Univariate Analysis**

Firstly, before analyzing the relations between the data columns let's focus on the frequencies of informations

Gender Distribution: The dataset was predominantly composed of two genders, with some missing values.

Drug Frequency: The drug ***Chlordiazepoxide-Amitriptyline*** was the most frequently prescribed drug.

Side Effects: A wide variety of side effects were reported, with some side effects being more common for specific drugs. Most common side effect is ***"Ağızda farklı bir tat***"

b. **Cross Analysis**

After implementing univariate analysis now focus on the cross analysis to understand the relations between informations to detect why those side effects cause.

Cross Analysis 1: Drugs and Side Effects

After I examined the relations between drugs and side effects, I observed that levomilnacipran is mostly causing a different taste in the mouth etc. With a good observation of heatmap we can clearly understand the drugs and side effects relation directly, but we have to consider other information about the patient also.

Cross Analysis 2: Side Effects and Chronic Illnesses

I performed a cross-tabulation between chronic illnesses and side effects, we can clearly observe that there is a great relation between alzheimer and different taste in the mouth.

Cross Analysis 3: Side Effects and Relatives' Chronic Illnesses

I examined the relationship between relatives of patients' chronic illness and side effects to understand that is there relation between those two information. Cross analysis are implemented between mother,father,male sibling, female sibling's chronic illnesses and side effects, separately.

Cross Analysis 4: Side Effects and Blood Type

I examined the relationship between blood type and side effects, after that analysis we can observe that increased blood pressure and AB Rh- has a strong relationship etc.

Cross Analysis 5: Side Effects and Allergies
I examined the relationship between allergies and side effects, after that analysis we can observe that increased blood pressure and fish allergy have a strong relationship.

Cross Analysis 6: Side Effects and Drug Usage Period
After calculating the drug usage period, I have analyzed the relationship of this period and side effects. I couldn't find any significant relationship.

Cross Analysis 7: Side Effects and Side Effect Reporting Period
After calculating the side effect reporting period, I have analyzed the relationship of this period and side effects. Skin bruising, increased appetite, diarrhea reporting periods are clearly longer than the other side effects.

5. **Key Findings**
   Drug-Specific Patterns: The drug Chlordiazepoxide-Amitriptyline showed unique side effect patterns, and further analysis could focus on understanding the risk factors for these effects.

   *Geographic Variation*s: There were notable geographic variations in drug prescriptions and side effect reporting.

   Anomalies in Data: Outliers were detected in numerical fields such as height and weight, which should be further examined to ensure data quality.

6. **Conclusion**

The EDA and preprocessing steps revealed important insights into the dataset. Further analysis should be conducted to explore the clinical significance of these patterns, especially regarding the drug-specific side effects and their relationship to allergies, blood type, and geographic location.