Zahide Erva Bilen

1erva.bilen@gmail.com

# Summary of the Case Study

First, I changed the names of the columns in the data set to the appropriate format to avoid problems in coding.

## 1. Exploratory Data Analysis (EDA)

- There are 2357 observations (rows) and 19 features (columns) in the dataset. There are 12 categorical, 3 (kullanici_id, kilo, boy) numerical and 4 date-time variables.
- 11 features have missing values, especially in the “gender” variable.

Conclusions from categorical data:

- Although the difference is not much, there are more female individuals. But there is a lot of missing data.
- All persons have the same nationality in the dataset.
- Alzheimer's disease is unlikely to occur alone (This is also true for other diseases). Asthma, osteoporosis, and hypertension are particularly common in people with this disease.
- The city where users are located is varied, but Adana is the most represented city.
- There is a wide variety of drug names. Chlordiazepoxide-amitriptyline is the most used drug.
- People have a variety of allergies. Tomato is the most frequently reported allergy. Some allergies show clear gender differences, for example, people allergic to seafood and dog are more likely to be female. Some allergies seem to be common among both genders. Like caviar and kefir.
- The most common side effect after drug use is a change in taste in the mouth and high blood pressure.
- People with blood type AB Rh- are in the majority.

Conclusions from numerical data:

- The user id is numbered from 1 to 197, so we can say that one person took more than one medication.
- Age data is not symmetric. We can say that most people are between the ages of 30 and 60.
- Most of the people are between 80-100 kg. There is density in the area where the weight increases.
- We can say that the average height is between 170 cm and 190 cm.

Conclusions from date time data:

- The medicines show side effects between 25-40 days after they are started.

## 2. Data Pre-Processing

Since I calculate the age variable, I remove the date of birth variable. Since everyone has the same nationality, I am also removing this variable. Additionally, I also removed the id variable in the same way.

Handling missing values:

- Missing numerical and categorical data were filled with the KNNImputer method. Categorical data were filled after label encoding.
- I did not fill in the missing data in kronik_hastaliklarim, baba_kronik_hastaliklari, anne_kronik_hastaliklari, kiz_kardes_kronik_hastaliklari, erkek_kardes_kronik_hastaliklari because I think it has a meaning (if the person does not have a chronic disease, etc.)

Encoding categorical variables:

- I separated the data for chronic disease variables to create groups for each disease.
- Then I applied the LabelEncoder method to variables that I have not encoded before.

Standardizing numerical features:

- I applied the StandardScaler method to numerical features.