Prepared By : Enes Samet Aydı (esa.aydi@gmail.com)

Exploratory Data Analysis and Data Preprocessing for Machine Learning on Side Effect Dataset

1. Introduction

Project Overview

This project focuses on data analysis and data preprocessing for machine learning on medical dataset. The goal is to prepare the dataset for machine learning by conducting Exploratory Data Analysis (EDA), performing data preprocessing.

2. Dataset Description

Features

- **Kullanici_id** (ID)
- **Cinsiyet** (Gender)
- **Dogum_Tarihi** (Date of birth)
- **Uyruk** (Nationality)
- **Il** (City)
- **Ilac_Adi** (Drug name)
- **Ilac_Baslangic_Tarihi** (Drug start date)
- **Ilac_Bitis_Tarihi** (Drug end date)
- **Yan_Etki** (Side effect)
- **Yan_Etki_Bildirim_Tarihi** (Side effect notification date)
- **Alerjilerim** (Allergies)
- **Kronik Hastaliklarim** (Chronic Diseases)
- **Baba Kronik Hastaliklari** (Father's Chronic Diseases)
- **Anne Kronik Hastaliklari** (Mother's Chronic Diseases)
- **Kiz Kardes Kronik Hastaliklari** (Sister's Chronic Diseases)

- **Erkek Kardes Kronik Hastaliklari** (Brother's Chronic Diseases)
- **Kan Grubu** (Blood type)
- **Kilo** (Height)
- **Boy** (Weight)

3. Exploratory Data Analysis (EDA)

Summary of Findings

- **Missing Values:** Several columns had significant missing values**.**
- **Column Types:** The columns consist of int, float, datetime and object data types.
- **Side Effect Distribution:** The most common side effect is 'Agizda Farkli Bir Tat'
- **Age Distribution:** Most people are between the ages of 40 and 60.
- **Drug Side Effect:** Each drug has more than one side effect and same drug doesn't have same effect on every person.
- **The Relationship Between Time Stamps and Side Effects:** Time stamps have weak effects on side effects.

4. Data Pre-processing

The following steps were followed for data preprocessing:

i. Dropped unnecessary columns
ii. Filling missing values with KNNImputer
iii. Encoding for categorical columns
iv. Normalization for numerical columns