Name: Kader Kılıç
Email: kaderkilicd@gmail.com

Findings Document – Data Science Intern Case Study

Dataset Overview

- Dataset size: **2235 observations, 13 features**
- Target variable: **TedaviSuresi (treatment duration in sessions)**
- Features include demographic information, medical history, diagnoses, treatments, and application details.

Exploratory Data Analysis (EDA)

- **Age (Yas):** Patients range from 6 to 87 years old. Majority are between 30–55.
- **Gender (Cinsiyet):** ~65% women, ~35% men.
- **Nationality (Uyruk):** 97% Turkey, very few from other countries.
- **Blood Type (KanGrubu):** Balanced distribution across major groups, Rh+ dominant.
- **Chronic Diseases (KronikHastalik):** 611 patients have no chronic disease; remaining have 220+ unique combinations (e.g., diabetes, asthma, muscular dystrophy).
- **Allergies (Alerji):** 944 patients report none; common allergies include pollen, dust, and drug reactions.
- **Departments (Bolum):** Mostly "Physical Medicine & Rehabilitation, Respiratory Center" (2056 patients).
- **Diagnoses (Tanilar):** High cardinality (340+ unique), dominated by back/neck pain, intervertebral disc disorders, and shoulder injuries.
- **Treatments (TedaviAdi):** 239 unique treatments; most frequent are related to back pain and disc problems.
- **Application Sites (UygulamaYerleri):** 30+ unique, commonly spine, neck, knees, shoulders.

Data Preprocessing

- **Missing Data:**
  - Few missing values handled with imputation.
- **Encoding:**

- Gender → One-Hot (Cinsiyet_Kadin, Cinsiyet_Erkek)
- Blood type → One-Hot
- Nationality → Binary (Turkey vs Other)
- Diagnoses, Treatments, Application sites → Multi-label encoding (due to multiple categories per patient).

Conclusion

The dataset is now **clean, preprocessed, and model-ready**.
It can be used for predictive modeling to estimate treatment duration based on patient demographics, chronic conditions, and diagnoses.