Binnur Ersöz

binnurersoz@gmail.com

Physical Medicine & Rehabilitation Dataset – EDA & Preprocessing Documentation

1. **Dataset Overview**
   - Observations: 2235
   - Features: 13
   - Target Variable: TedaviSuresi (Treatment Duration)
2. **Exploratory Data Analysis (EDA)**

   2.1 Dataset Shape
   - Rows: 2235
   - Columns: 13

   2.2 Data Types & Missing Values

#	Column	Non-Null Count	Missing %
0	Alerji	1291	42.237136
1	KanGrubu	1560	30.201342
2	KronikHastalik	1624	27.337808
3	UygulamaYerleri	2014	9.888143
4	Cinsiyet	2066	7.561521
5	Tanilar	2160	3.355705
6	Bolum	2224	0.492170
7	HastaNo	2235	0.000000
8	Yas	2235	0.000000
9	Uyruk	2235	0.000000
10	TedaviAdi	2235	0.000000
11	TedaviSuresi	2235	0.000000
12	UygulamaSuresi	2014	0.000000

   - Target (TedaviSuresi) has no missing values
   - Visualized missing values with missingno.matrix and heatmap.

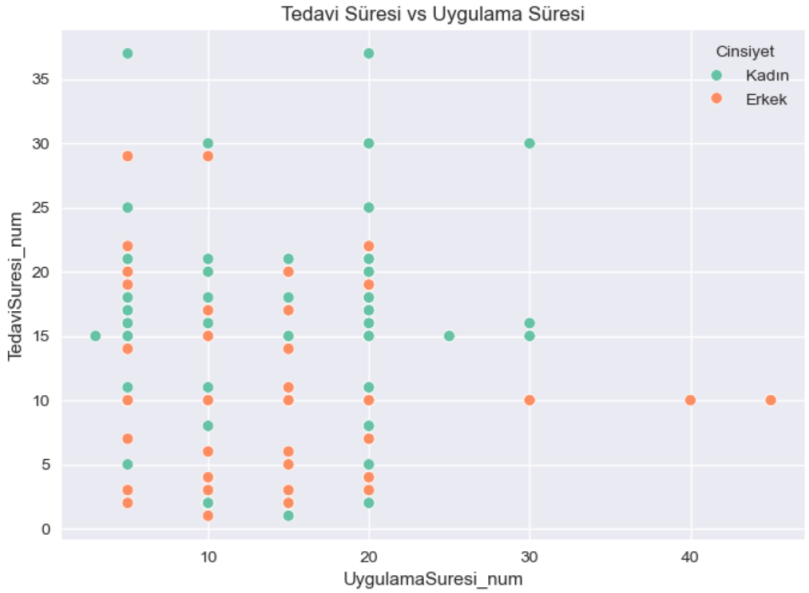
   2.3 Duplicate Records
   - Total duplicates: 928
   - Duplicates based on HastaNo + TedaviAdi + UygulamaYerleri + UygulamaSuresi: 1060
   - Removed duplicates to avoid bias in analysis.

   2.4 Summary Statistics of Target Variable

count	2235.000000
mean	14.570917
std	3.725322
min	1.000000
25%	15.000000
50%	15.000000
75%	15.000000
max	37.000000

Visualizations:

- Histogram of TedaviSuresi shows right-skewed distribution.
- Boxplots used to detect outliers.
- Scatterplot of UygulamaSuresi vs TedaviSuresi with Cinsiyet hue:



2.5 Categorical Analysis

- Gender Distribution: Kadın > Erkek
- Most Frequent Diagnoses:
  - DORSALJİ, DİĞER, LUMBOSAKRAL BÖLGE
  - Omuzun darbe sendromu
  - İntervertebral disk bozuklukları, tanımlanmamış
- Most Frequent Chronic Conditions:
  - Myastenia gravis
  - Aritmi
  - Fascioscapulohumeral Distrofi

2.6 Correlations

- Moderate positive correlation between UygulamaSuresi_num and TedaviSuresi_num.
- Age (Yas) shows low correlation with TedaviSuresi.

2.7 Department Analysis

- Avg. TedaviSuresi by Department:
  - Fiziksel Tıp Ve Rehabilitasyon,Solunum Merkezi: 15.14
  - Göğüs Hastalıkları: 13.12
  - Ortopedi Ve Travmatoloji: 4.82

3. Data Preprocessing Steps

3.1 Handling Missing Values

- Categorical columns (Cinsiyet, KanGrubu, KronikHastalik, Bolum, Alerji, Tanilar, UygulamaYerleri) missing values filled with "Missing" using SimpleImputer.

3.2 Removing Duplicates

- Removed rows duplicated in HastaNo + TedaviAdi + UygulamaYerleri + UygulamaSuresi.
- Dataset reduced from 2235 → 1175 rows.

3.3 Numeric Transformations

- Extracted numeric values from TedaviSuresi and UygulamaSuresi.
- Applied StandardScaler to Yas, TedaviSuresi_num, UygulamaSuresi_num.

3.4 Encoding Categorical Variables

- Used LabelEncoder for:
  Cinsiyet, KanGrubu, KronikHastalik, Bolum, Alerji, Tanilar, UygulamaYerleri.
- Resulting df_final includes scaled numeric + encoded categorical columns.

4. Final Dataset (df_final)

- Rows: 1175
- Columns: Scaled numeric + encoded categorical
- Ready for predictive modeling with TedaviSuresi_num as target.

df_final:

	Yas	TedaviSuresi_num	UygulamaSuresi_num	Cinsiyet	KanGrubu	KronikHastalik	Bolum	Alerji	Tanilar	UygulamaYerleri
0	0.831408	-2.150290	0.842893	2	0	52	1	24	15	0
1	-1.232020	0.206824	0.842893	1	0	82	1	1	215	3
2	-1.232020	0.206824	0.842893	1	0	82	1	1	215	6
3	-1.232020	0.206824	-1.188279	1	0	82	1	1	215	3
6	0.831408	-0.971733	2.197007	1	0	85	1	11	226	9

5. Summary

- Dataset cleaned, missing values imputed, duplicates removed.
- Target (TedaviSuresi_num) distribution analyzed.
- Numeric columns scaled, categorical columns encoded.