**Kübra Nur Akçakaya**

[**kubrakcakaya@outlook.com**](mailto:kubrakcakaya@outlook.com)

**REPORT**

**1-Introduction**

In this study, the Physical Medicine & Rehabilitation dataset was examined. The dataset contains 2235 observations and 13 features. The project aims to explore the dataset (EDA), handle missing values, convert categorical variables to numerical format, standardize numerical values, and prepare the dataset for modelling.

**2-Exploraty Data Analysis (EDA)**

There are a total of 2706 missing values ​​in the data set. The columns containing missing values ​​are as follows:

-Cinsiyet

-KanGrubu

-KronikHastalik

-Bolum

-Alerji

-Tanilar

-UygulamaYerleri

The missing values ​​shown above are visualized with appropriate graphs to understand the distribution.

**3- Handling Missing Values**

SimpleImputer from the scikit-learn library was used to fill in missing values.

The columns "Cinsiyet", "KanGrubu", "Bolum" and "UygulamaYerleri" were filled with the Mod (most frequent) .

The columns "KronikHastalik", "Alerji" and"Tanilar" were filled with a fixed value ("unknown").

This process resulted in no missing values ​​in the dataset.

**4- Encoding Categorical Variables**

Numveric inputs are needed to maket he dataset Model-ready. Therefore, categorical variables were converted to numeric format. Label Encoding and One-Hot encoding were used in this step.kullanılmıştır.

Label Encoding was used for the "Cinsiyet" column.

One-Hot Encoding was applied to the "KanGrubu", "Uyruk", "Bolum", "UygulamaYerleri" columns.

**5- Standardizing Numerical Values**

StandardScaler was used. Numerical columns were adjusted to a specific scale.

**6-Summary**

The dataset was explored using EDA.

Missing values ​​were processed.

Categorical variables were converted to numeric format.

Numerical values ​​were standardized.

The dataset was prepared for modeling.