Talent Academy Data Science Case Study & Preprocessing Summary

Ecem Ceylan

ecemceyllan12@gmail.com

1. General Overview

This study was conducted on a physical medicine and rehabilitation dataset. The dataset contains 2235 rows and 13 columns. The target variable is Treatment Duration (TedaviSuresi). The purpose of this project is to analyze missing values in the dataset and prepare the data for modeling through preprocessing.

2. Data Understanding

Columns in the dataset:

- **HastaNo**: Patient ID
- **Yas**: Patient age
- **Cinsiyet, KanGrubu, Uyruk, Bolum, TedaviAdi, UygulamaYerleri, KronikHastalik, Alerji, Tanilar**: Categorical variables
- **TedaviSuresi**: Target variable (minutes)
- **UygulamaSuresi**: Application duration (minutes)

3. Data Preprocessing

- Numerical Columns: Extracted numeric values from columns with strings (e.g., "20 Dakika" → 20).
  Filled missing values using mean imputation with SimpleImputer.
- Categorical Columns:
  Filled missing values with mode (most frequent value).
  Applied OneHotEncoder to convert categorical variables into numerical format suitable for modeling.
- ColumnTransformer was used to apply OneHotEncoding while leaving numerical columns unchanged.

4. Next Steps

Considering the large number of features after one-hot encoding and the continuous nature of the target variable (TedaviSuresi), the **Random Forest Regressor** is an appropriate choice. This model effectively manages high-dimensional data, captures non-linear relationships, and remains robust to outliers.