Semih Aşdan – semih.asdan@gmail.com

Data Preprocessing and Feature Engineering Steps

This section details which operations were applied to each original column in the raw dataset and why.

1. HastaNo (Anonymized Patient ID)

- **Operation Performed:** Deleted from the dataset before modeling.
- **Reason:** This column is a unique identifier for each patient and does not hold a meaningful relationship with the target variable, `TedaviSuresi` (Treatment Duration). It was removed to prevent the model from learning incorrect patterns by memorizing these IDs (overfitting) and to get rid of an unnecessary feature.

2. Yas (Age)

- **Operation Performed:**
  - **Numerical Value Kept:** The original `Yas` column was kept and scaled.
  - **`Yas_Karesi` Column Created:** A new feature was derived by taking the square of the `Yas` value.
  - **Categorization Performed:** Age was divided into logical groups such as 'Baby', 'Child', 'Adolescent', 'Young_Adult', 'Middle_Aged', 'Elderly', and these groups were converted into `YasGrubu_*` columns using One-Hot Encoding.
- **Reason:**
  - The original `Yas` column was kept for the model to learn if there is a linear relationship with treatment duration.
  - `Yas_Karesi` allows the model to capture a non-linear relationship between age and treatment duration (e.g., recovery time increasing exponentially with age).
  - Categorization enables the model to more easily learn distinct and significant behavioral differences between groups like "Adolescent" and "Elderly," rather than minor differences between "22 years old" and "24 years old."

3. Cinsiyet (Gender)

- **Operation Performed:** Missing values were first filled based on intra-patient consistency, then with the most frequent value (mode). It was then converted into a single binary column (`Cinsiyet_Erkek`) with 'male': 1 and 'female': 0.
- **Reason:** Instead of creating two separate One-Hot Encoding columns for a two-category feature (which carries the risk of the "dummy variable trap"), representing the same information more efficiently with a single column is the best practice.

4. KanGrubu (Blood Type)

- **Operation Performed:** Missing values were first filled based on intra-patient consistency. Remaining missing values and rare blood types with very low frequency (like 0 Rh-, AB Rh-) were grouped under a single 'Unknown' category. Finally, these simplified categories were converted into `KanGrubu_*` columns using One-Hot Encoding.
- **Reason:** This grouping was done to prevent the model from learning noise from rare categories with only a few examples and to focus its attention on statistically more significant, common blood types.

5. Uyruk (Nationality)

- **Operation Performed:** Completely deleted from the dataset.
- **Reason:** In Exploratory Data Analysis (EDA), it was observed that almost all values in this column were 'Türkiye'. For a feature to be meaningful for a model, it must show diversity (variance). Since this column had an almost constant value, it had no predictive power.

6. KronikHastalik (Chronic Conditions)

- **Operation Performed:**
  - Text cleaning was performed, and intra-patient inconsistencies were resolved. Remaining missing values were filled with "None".
  - `KronikHastalik_Sayisi`: The total number of chronic diseases for the patient was calculated.
  - `Genel_Saglik_Skoru`: A weighted health score was created by assigning a risk score to each disease based on its severity.
  - `KronikHastalik_AdvFreq`: A frequency score was calculated to show how common the patient's primary (first) chronic disease is.
  - The original text column was deleted.
- **Reason:** The goal was to transform a single text column into multiple rich numerical features that the model can understand. Different pieces of information, such as the number of diseases, the overall health burden, and the rarity of a disease, provide separate signals for the model to make more accurate predictions.

7. Bolum (Department/Clinic)

- **Operation Performed:** Missing values were filled with the most frequent value. For the top two categories that constituted over 90% of the data (Physical Medicine and Orthopedics), two separate binary columns named `Bolum_Is_FizikselTip` and `Bolum_Is_Ortopedi` were created. If both are 0, it is understood that the patient is in one of the other rare departments.
- **Reason:** Instead of expanding more than 10 categories with One-Hot Encoding, a simpler and more effective feature set was created by highlighting the dominant and important categories in the dataset and grouping the rest into a single "other" category.

8. Alerji (Allergies)

- **Operation Performed:** Missing values were filled with intra-patient consistency, and the rest with an 'Allergy_None' category. Then, One-Hot Encoding (`get_dummies`) was applied to create a separate column for each potential allergen. Additionally, a new feature named `Coklu_Ilac_Alerjisi_Sayisi` was derived by summing up only the allergy columns related to medications.
- **Reason:** Since a patient can have multiple allergies at the same time (a "multi-label" situation), representing the presence or absence of each allergy in a separate column is the most accurate method. The number of drug allergies represents a specific condition that could affect the treatment process.

9. Tanilar (Diagnoses)

- **Operation Performed:**
  - Text cleaning was performed, and missing values were filled relationally.
  - `Tani_Sayisi`: The total number of diagnoses for the patient was calculated.
  - `Tani_Ciddiyeti_Yuksek`: A binary column indicating the severity of the diagnosis was created by scanning for keywords like "fracture," "op," "rupture."
  - `Tanilar_AdvFreq`: A frequency score was calculated to show how common the patient's primary (first) diagnosis is.
  - The original text column was deleted.
- **Reason:** As with the `KronikHastalik` column, the goal was to derive multiple meaningful and numerical features from a single complex text field, such as the number, severity, and prevalence of diagnoses, which the model can process more easily.

10. TedaviAdi (Treatment Name)

- **Operation Performed:** Text cleaning and relational filling were performed. Rare treatments with very low frequency were grouped as "Other_Treatment." Finally, advanced frequency encoding (`TedaviAdi_AdvFreq`) was applied to this simplified category list. The original text column was deleted.
- **Reason:** One-Hot Encoding was not suitable due to the very high number of unique categories (180+). Frequency encoding was the most effective way to solve this high cardinality problem with a single meaningful column.

11. TedaviSuresi (Treatment Duration in Sessions)

- **Operation Performed:** The data distribution was analyzed, and it was converted from a regression target to a classification target. The values were divided into 4 categories: `0_Very_Short`, `1_Medium`, `2_Standard_15`, `3_Long`, and a numerical target column named `y_target_sinif` was created.
- **Reason:** The concentration of a large portion of the data at a specific value (15 sessions) is not suitable for regression models. Transforming the problem into a classification problem, which better reflects the real-world process of "assigning a treatment package," is a more accurate and meaningful modeling approach.

12. UygulamaYerleri (Application Sites)

- **Operation Performed:** First, directional terms like "right" and "left" were removed to merge categories (right shoulder and left shoulder -> shoulder). Missing data was filled from `TedaviAdi`. Frequency encoding (`UygulamaYerleri_AdvFreq`) was applied to this cleaned column, and more general categorical features like `Tedavi_Odak_Bolgesi` (Treatment Focus Area: Upper/Lower Extremity, Trunk) were derived.
- **Reason:** The goal was to transform raw application site information into both a numerical score indicating its prevalence and higher-level body region categories that the model can generalize from more easily.

13. UygulamaSuresi (Application Duration)

- **Operation Performed:** Kept as a numerical feature and scaled with `StandardScaler`.
- **Reason:** The duration of treatment sessions can be an important indicator for the total treatment package duration (`TedaviSuresi`), so it was used in the model as a valuable numerical feature.

Applied Feature Engineering Strategies

The following strategies were applied to derive new, meaningful, and numerical features from the original columns in the raw dataset, enabling the model to understand the data more easily and make more accurate predictions:

1. KronikHastalik_Sayisi (ChronicDisease_Count)

- **How it was created:** The text in the `KronikHastalik` column was split by commas, and the total number of chronic diseases for each patient was calculated.
- **Why it was created:** This feature quantifies a patient's overall health burden in a simple and powerful way. It allows the model to understand the difference between a patient with a single chronic disease and a complex case with multiple diseases. A higher number of chronic diseases could potentially mean a longer and more complex treatment process.

2. Tani_Ciddiyeti_Yuksek (High_Diagnosis_Severity)

- **How it was created:** The `Tanilar` and `TedaviAdi` columns were scanned for keywords indicating severity or surgical intervention, such as "fracture," "rupture," "surgery," "implant," and "paralysis." If any of these words were found, this new column was assigned a value of 1 (severe); otherwise, it was assigned 0 (not severe).
- **Why it was created:** This helps the model learn the critical difference between a simple muscle ache and a condition that required surgical intervention. High-severity diagnoses often require longer and more intensive rehabilitation programs.

3. Ortopedi_ve_Ciddi_Tani (Orthopedics_and_Severe_Diagnosis - Interaction Feature)

- **How it was created:** The previously created `Bolum_Is_Ortopedi` and `Tani_Ciddiyeti_Yuksek` columns were combined. If a patient was both in the Orthopedics department (1) and had a severe diagnosis (1), this new column received a value of 1; in all other cases, it received 0.
- **Why it was created:** This is an interaction feature. It helps the model understand the difference between a "simple case in the Orthopedics department" and a "severe, surgical case in the Orthopedics department." The treatment durations for these two scenarios can be very different, and this feature captures this specific scenario.

4. Genel_Saglik_Skoru (Overall_Health_Score)

- **How it was created:** Each disease in the `KronikHastalik` column was assigned a risk score (from 1 to 3) based on its medical severity and progressive nature. Each patient's total score was calculated to create this new column.
- **Why it was created:** This is a more advanced metric than `KronikHastalik_Sayisi`. It teaches the model that the severity of "Asthma" (score 1) is not the same as "Duchenne Muscular Dystrophy" (score 3). This allows for a more precise measurement of the patient's health burden.

5. Coklu_Ilac_Alerjisi_Sayisi (Multiple_Drug_Allergy_Count - Polypharmacy)

- **How it was created:** From the One-Hot Encoded `Alerji_*` columns, only those representing medications (novalgin, voltaren, etc.) were selected and summed row-wise.
- **Why it was created:** The fact that a patient is allergic to multiple drugs may indicate a general sensitivity or a situation requiring more caution in treatment planning. This feature represents a more specific medical risk than environmental allergies like pollen.

6. Tani_Sayisi (Diagnosis_Count)

- **How it was created:** The text in the `Tanilar` column was split by commas, and the total number of diagnoses for each patient was calculated.
- **Why it was created:** The situation of a patient with a single, clear diagnosis is different from that of a patient with multiple comorbid (accompanying) diagnoses. An increase in the number of diagnoses is an indicator that can increase the complexity of the case and, consequently, the potential treatment duration.

7. Tedavi_Odak_Bolgesi (Treatment_Focus_Area and its One-Hot Encoded Columns)

- **How it was created:** Specific body regions in the `UygulamaYerleri` column ("right shoulder," "left knee," etc.) were grouped into three main anatomical regions: `Ust_Ekstremite` (Upper_Extremity), `Alt_Ekstremite` (Lower_Extremity), and `Govde` (Trunk). This new categorical column was then converted into `OdakBolgesi_*` columns using One-Hot Encoding.
- **Why it was created:** It helps the model generalize that 'right wrist' and 'left shoulder area' are both 'Upper_Extremity' treatments and may have similar dynamics, instead of trying to

learn the difference between them. This allows the model to capture stronger patterns with less data.

*8. Yas_Karesi (Age_Squared) and Yas_Grubu_ (Age_Group_*) Columns**

- **How it was created:** `Yas_Karesi` was created by squaring the original age value. `Yas_Grubu` was created by binning age into categories like Baby, Child, Adolescent, Adult, Elderly using `pd.cut`, and then One-Hot Encoding was applied.
- **Why it was created:** These two features were created to model the multi-faceted effect of age on treatment duration. While the original `Yas` column captures a linear relationship, `Yas_Karesi` helps the model learn non-linear relationships, such as the effect accelerating or decelerating with age, and the `Yas_Grubu_*` columns allow it to learn sharp behavioral changes between different age groups.