Data Preprocessing – Hyper parathyroid

Musab Khan

I. Differences Between the New and Old Datasets

1. **Missing Columns in the New Dataset:**
    - *tersiyerhp*
    - *Medeni Durum (marital status)*
    - *24 saat idrar sodyum (urinary_sodium_24h)*
    - *Kemik spesifik ALP (bone_specific_alp)*
    - *2nd and 3rd AMELİYAT PATOLOJİ RAPORU (2nd and 3rd surgery_pathology_report)*
    - *hdl*
2. **Renaming of Columns:**
    - The column formerly named **"nhph(normokals)"** is now renamed to **"NORMOKALSEM?K"**.
3. **New Column Inclusions:**
    - The new dataset includes **"SEMTOMATIKPHP"**, a column that was not part of the previous mapping.
    - Two versions of the ALP measurement exist: **"alp"** and **"ALP"**.

II. Data Preprocessing Steps

1. **Column Removal:**
    - **After ALP:** All columns following the ALP column were removed.
    - **Sparse Columns:** The following columns were removed due to sparse data:
        - Unnamed: 89, Unnamed: 92, Unnamed: 79, Unnamed: 84, Unnamed: 85, Unnamed: 95, Unnamed: 95.1, Unnamed: 96, Unnamed: 97, Unnamed: 98, Unnamed: 99, Unnamed: 100.

- **Additional Dropped Columns:** The following columns were dropped:
    - 'hospital_name', 'living_city', 'mobile_phone_number', 'patient_name_1st_visit', 'date_of_1st_visit', 'patient_name_2nd_visit', 'date_of_2nd_visit', 'patient_name_3rd_visit', 'date_of_3rd_visit', 'medication_name', 'comorbidity', and 'semtomatikphp'.
- The **"semtomatikphp"** target column was removed entirely because it contained no data.

2. **Handling Dual ALP Columns:**
    - For the two ALP columns (**"alp"** and **"ALP"**), is following strategy correct:
        - **If one column is empty** and the other contains a value, use the non-empty value.
        - **If both columns have values,** compute the mean of the two values and use that as the final ALP value.
3. **Categorical Missing Values:**
    - Should empty cells be filled with mean?
4. **Conversion of Categorical Variables:**
    - **aphp Column:** Map 'Asemptomatik' and 'EVET' to 1; missing values are set to 0.
    - **nhph Column:** Map 'nhph(normokals)' and 'Normokalsemik' to 1; map 'HAYIR' and missing values to 0.
    - **secondary_hp Column:** Map various affirmative responses (e.g., 'EVET' and entries resembling 'Sekonder hp') to 1; map 'HAYIR' and missing values to 0.
5. **Gender Encoding:**

- 'Men' → 0
- 'Women' → 1
- Missing or invalid values → -1

III. Data Format Corrections

- **Empty/Whitespace: " ", " ", " "**
- **Comma as Decimal: "0,6", "0,69"**
- **Square Brackets: "[1.0726]", "[103.9000]"**
- **Multiple Values (semicolon): "8.5; 11.2", "47; 22", "30.6; 115.3"**
- **Multiple Dots: "8.511.2"**
- **Thousand Separators: "1,063,125"**

Example Input & Output:

Raw Data (`bmi`)	Processed Data (`bmi`)
`"23.5"`	`23.5`
`"22,8"`	`22.8`
`"25.0; 26.3"`	`26.3`
`"Mg: 10.2"`	`10.2`
`"[15.6]"`	`15.6`
`""`	`0`

---

1. **Undecided Handling for Specific Diagnostic/Imaging Columns:**
    - The following columns require further decisions on handling:
        - 'parathyroid_scintigraphy'
        - 'parathyroid_spect'
        - 'neck_ultrasound'
        - '4d_ct'
        - '4d_mri'

- '99mtc_mibi'
- '99mtc_sestamibi'
- 'first_surgery_pathology_report'