Data Preprocessing – Hyper parathyroid

Musab Khan

I. 2024 & 2025 Dataset Differences

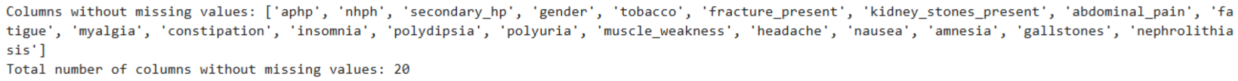
1. **Columns Removed in the New Dataset:**
   - *tersiyerhp*
   - *Medeni Durum (marital status)*
   - *24 saat idrar sodyum (urinary_sodium_24h)*
   - *Kemik spesifik ALP (bone_specific_alp)*
   - *2nd and 3rd AMELİYAT PATOLOJİ RAPORU (2nd and 3rd surgery_pathology_report)*
   - *hdl*
2. **Renaming 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.

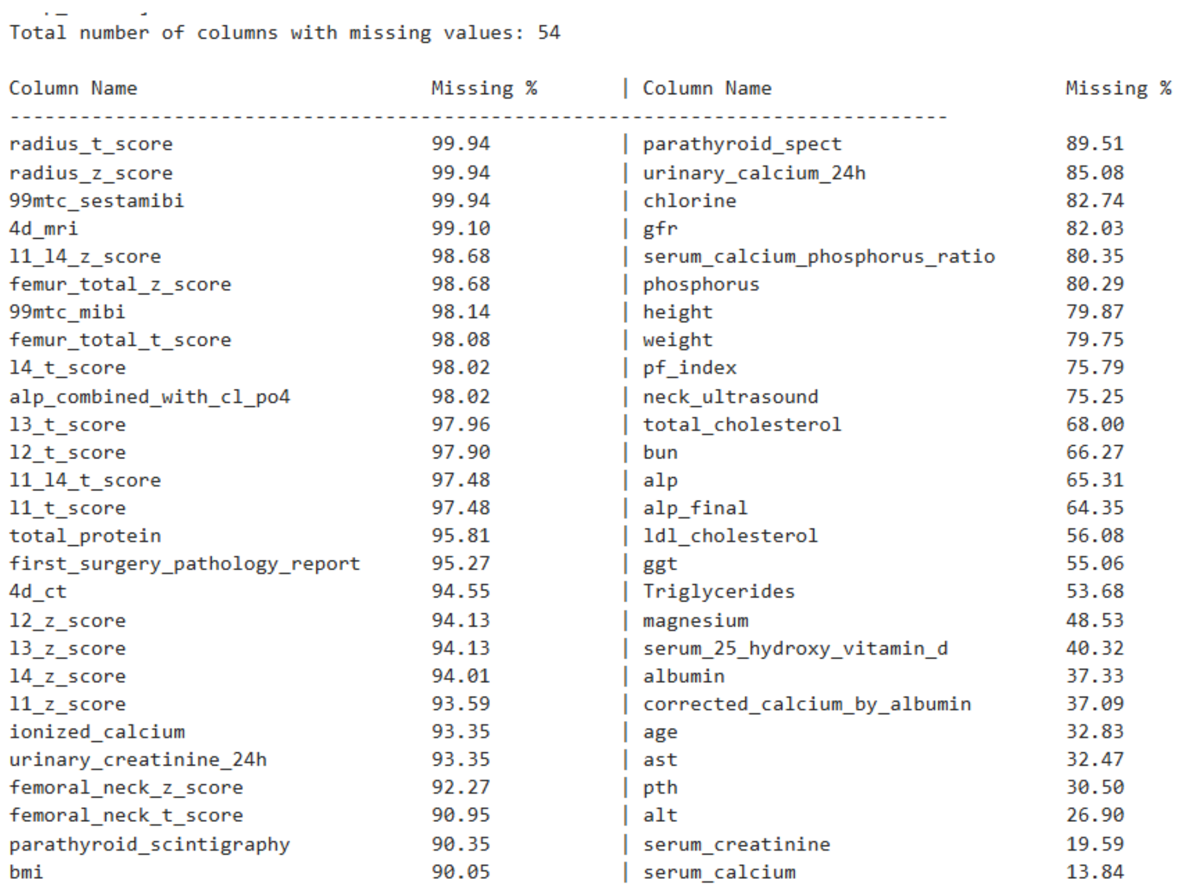
2. Handled Categorical Missing Values:



```
Columns without missing values: ['aphp', 'nhph', 'secondary_hp', 'gender', 'tobacco', 'fracture_present', 'kidney_stones_present', 'abdominal_pain', 'fa
tigue', 'myalgia', 'constipation', 'insomnia', 'polydipsia', 'polyuria', 'muscle_weakness', 'headache', 'nausea', 'amnesia', 'gallstones', 'nephrolithia
sis']
Total number of columns without missing values: 20
```

3. **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.

4. **Columns With Missing Values:**



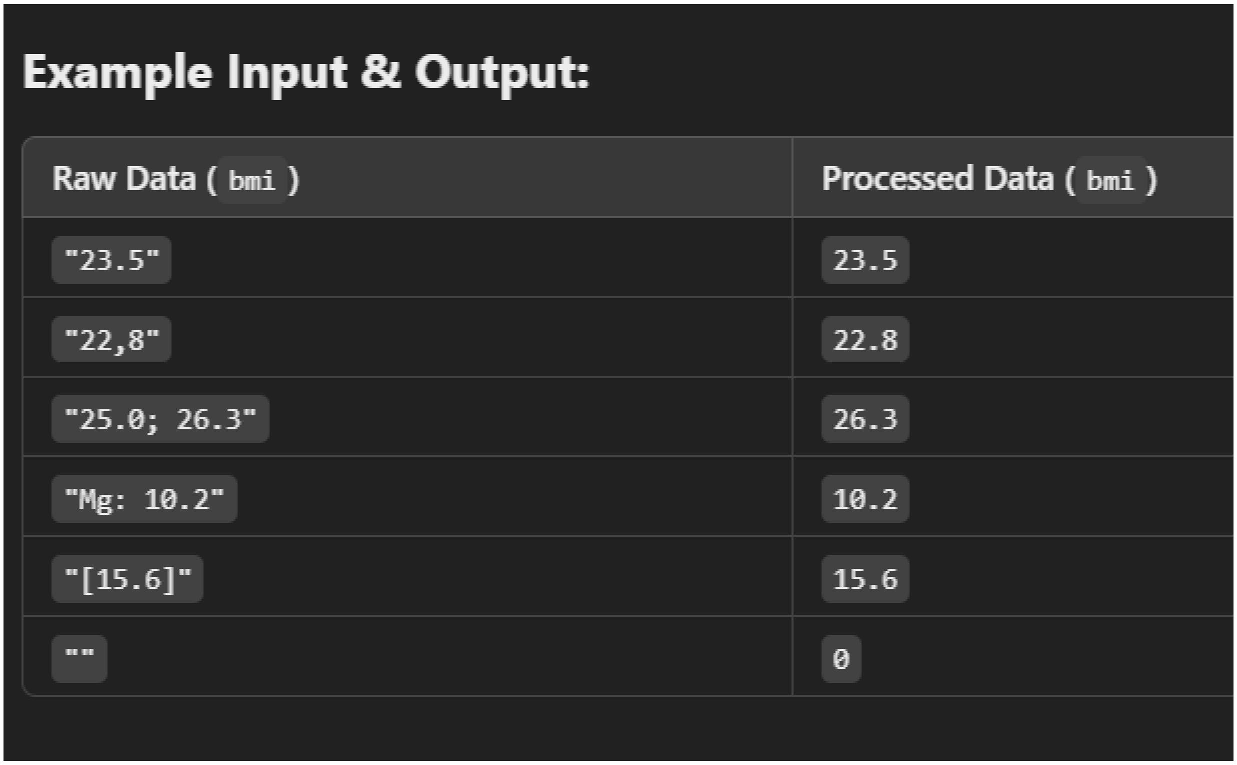
```
Total number of columns with missing values: 54

Column Name                         Missing %       | Column Name                         Missing %
--------------------------------------------------------------------------------
radius_t_score                      99.94           | parathyroid_spect                   89.51
radius_z_score                      99.94           | urinary_calcium_24h                 85.08
99mtc_sestamibi                     99.94           | chlorine                            82.74
4d_mri                              99.10           | gfr                                 82.03
l1_l4_z_score                       98.68           | serum_calcium_phosphorus_ratio      80.35
femur_total_z_score                 98.68           | phosphorus                          80.29
99mtc_mibi                          98.14           | height                              79.87
femur_total_t_score                 98.08           | weight                              79.75
l4_t_score                          98.02           | pf_index                            75.79
alp_combined_with_cl_po4            98.02           | neck_ultrasound                     75.25
l3_t_score                          97.96           | total_cholesterol                   68.00
l2_t_score                          97.90           | bun                                 66.27
l1_l4_t_score                       97.48           | alp                                 65.31
l1_t_score                          97.48           | alp_final                           64.35
total_protein                       95.81           | ldl_cholesterol                     56.08
first_surgery_pathology_report      95.27           | ggt                                 55.06
4d_ct                               94.55           | Triglycerides                       53.68
l2_z_score                          94.13           | magnesium                           48.53
l3_z_score                          94.13           | serum_25_hydroxy_vitamin_d          40.32
l4_z_score                          94.01           | albumin                             37.33
l1_z_score                          93.59           | corrected_calcium_by_albumin        37.09
ionized_calcium                     93.35           | age                                 32.83
urinary_creatinine_24h              93.35           | ast                                 32.47
femoral_neck_z_score                92.27           | pth                                 30.50
femoral_neck_t_score                90.95           | alt                                 26.90
parathyroid_scintigraphy            90.35           | serum_creatinine                    19.59
bmi                                 90.05           | serum_calcium                       13.84
```

5. **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.
6. **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'