Tugas Proyek Praktik: Preprocessing Dataset Rumah Sakit Indonesia Pemerintah vs Swasta per Provinsi

DISUSUN OLEH:

NAMA: FAUZAN MUHAMMAD ZAHRAN

NIM :105841116723

KELAS: 5E

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS MUHAMMADIYAH MAKASSAR

2025

Tujuan

Tujuan dari analisis dataset ini adalah untuk menganalisis distribusi jumlah rumah sakit di seluruh provinsi di Indonesia berdasarkan dua kategori utama, yaitu rumah sakit milik pemerintah dan rumah sakit swasta. Selain itu, analisis ini bertujuan untuk mengevaluasi perbandingan kapasitas dan sebaran layanan kesehatan antarwilayah guna mengidentifikasi provinsi dengan jumlah fasilitas kesehatan tertinggi dan terendah. Analisis juga dilakukan untuk menemukan ketimpangan antara sektor pemerintah dan swasta dalam penyediaan layanan kesehatan di setiap provinsi. Dataset ini dipersiapkan agar bersih, lengkap, dan terstandardisasi sehingga dapat digunakan untuk analisis lanjutan seperti pemodelan *machine learning* atau pengambilan keputusan berbasis data di sektor kesehatan. Hasil dari proses ini diharapkan dapat memberikan gambaran yang lebih jelas mengenai pemerataan fasilitas dan layanan kesehatan di Indonesia.

Detail Skenario

Proses analisis dimulai dengan pengumpulan data dari sumber resmi seperti Kementerian Kesehatan (Kemenkes) atau instansi pemerintah lainnya yang menyediakan informasi mengenai rumah sakit di Indonesia. Dataset yang digunakan memuat kolom penting seperti Provinsi, Jenis Rumah Sakit, Kepemilikan, Total Tempat Tidur, Total Layanan, dan Total Tenaga Kerja. Tahap selanjutnya adalah inspeksi awal menggunakan Python dan pustaka pandas untuk memahami struktur serta isi dataset. Pada tahap ini dilakukan pemeriksaan nilai hilang (*missing values*) dan penyesuaian format data agar konsisten antar kolom. Kolom numerik seperti *total_tempat_tidur* dan *total_tenaga_kerja* diimputasi menggunakan nilai median karena lebih tahan terhadap *outlier*, sedangkan kolom kategorikal seperti *jenis* dan *kepemilikan* diimputasi menggunakan nilai modus (mode) untuk menjaga distribusi data tetap alami. Setelah proses pembersihan, dilakukan pengelompokan (agregasi) berdasarkan provinsi dan kepemilikan rumah sakit, serta visualisasi data untuk menggambarkan sebaran dan proporsi rumah sakit pemerintah dan swasta di seluruh Indonesia. Proses ini menghasilkan dataset yang telah terolah dengan baik dan siap digunakan untuk analisis kebijakan, evaluasi pelayanan kesehatan, serta pengembangan model analisis lanjutan di sektor kesehatan.

Langkah-Langkah Tugas

1. Pengumpulan dan Inspeksi Data

Identifikasi:

Dataset **"Hospital_Indonesia_datasets.csv"** dimuat ke dalam *environment* analisis menggunakan pustaka **pandas** pada Python. Dataset ini berisi informasi tentang **daftar rumah sakit di Indonesia** beserta detail administratif, tipe, kepemilikan, dan kapasitas layanan.

Dataset ini digunakan untuk menganalisis **sebaran rumah sakit berdasarkan provinsi**, serta **perbandingan antara rumah sakit pemerintah dan swasta** di setiap wilayah. Data ini diperoleh dari **sumber resmi Kementerian Kesehatan Republik Indonesia (Kemenkes)** melalui portal terbuka data fasilitas pelayanan kesehatan.

Deskripsi Dataset:

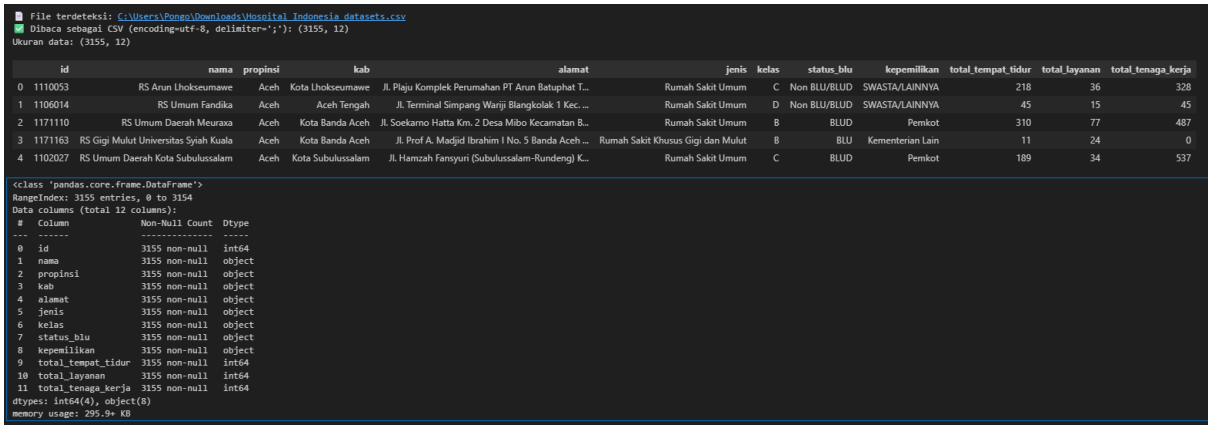
Dataset *"Hospital_Indonesia_datasets.csv"* berisi data rekapitulasi rumah sakit di seluruh Indonesia, yang mencakup kategori kepemilikan dan jenis fasilitas rumah sakit. Dataset memiliki beberapa informasi penting, yaitu:

- **Format data:** Comma Separated Values (.csv)
- **Jumlah baris dan kolom:** bervariasi (tergantung pembaruan data, umumnya ± 3.000 baris × 11 kolom)
- **Tahun data:** 2023–2024
- **Sumber data:** https://www.kaggle.com/datasets/muhammadhabibna/hospital-data-in-indonesia
- **Variabel utama:**
    - id – kode unik rumah sakit
    - nama – nama rumah sakit
    - propinsi – provinsi tempat rumah sakit berada
    - kab – kabupaten atau kota
    - alamat – alamat lengkap rumah sakit
    - jenis – jenis rumah sakit (umum, khusus, RSUD, dll.)
    - kelas – kelas rumah sakit (A, B, C, D)
    - status_blu – status Badan Layanan Umum (opsional)
    - kepemilikan – kepemilikan rumah sakit (pemerintah atau swasta)
    - total_tempat_tidur – jumlah total tempat tidur tersedia
    - total_tenaga_kerja – jumlah tenaga kerja di rumah sakit

Inspeksi Awal:

Hasil inspeksi awal terhadap dataset menunjukkan bahwa data memiliki struktur yang baik dengan beberapa kolom berformat teks dan numerik. Kolom identitas seperti *nama*, *jenis*, *kelas*, dan *kepemilikan* bertipe objek (string), sedangkan kolom numerik seperti *total_tempat_tidur* dan *total_tenaga_kerja* bertipe *float64*.

Pemeriksaan awal (df.info(), df.isnull().sum()) menunjukkan adanya beberapa nilai kosong pada kolom tertentu, namun secara umum data lengkap dan konsisten. Dataset ini siap digunakan untuk tahap pembersihan dan analisis lanjutan setelah dilakukan proses imputasi dan standarisasi nilai.



File terdeteksi: C:\Users\Pongo\Downloads\Hospital_Indonesia_datasets.csv
Dibaca sebagai CSV (encoding=utf-8, delimiter=';'): (3155, 12)
Ukuran data: (3155, 12)

	id	nama	propinsi	kab	alamat	jenis	kelas	status_blu	kepemilikan	total_tempat_tidur	total_layanan	total_tenaga_kerja
0	1110053	RS Arun Lhokseumawe	Aceh	Kota Lhokseumawe	Jl. Plaju Komplek Perumahan PT Arun Batuphat T...	Rumah Sakit Umum	C	Non BLU/BLUD	SWASTA/LAINNYA	218	36	328
1	1106014	RS Umum Fandika	Aceh	Aceh Tengah	Jl. Terminal Simpang Warjii Blangkolak 1 Kec. ...	Rumah Sakit Umum	D	Non BLU/BLUD	SWASTA/LAINNYA	45	15	45
2	1171110	RS Umum Daerah Meuraxa	Aceh	Kota Banda Aceh	Jl. Soekarno Hatta Km. 2 Desa Mibo Kecamatan B...	Rumah Sakit Umum	B	BLUD	Pemkot	310	77	487
3	1171163	RS Gigi Mulut Universitas Syiah Kuala	Aceh	Kota Banda Aceh	Jl. Prof A. Madjid Ibrahim I No. 5 Banda Aceh ...	Rumah Sakit Khusus Gigi dan Mulut	B	BLU	Kementerian Lain	11	24	0
4	1102027	RS Umum Daerah Kota Subulussalam	Aceh	Kota Subulussalam	Jl. Hamzah Fansyuri (Subulussalam-Rundeng) K...	Rumah Sakit Umum	C	BLUD	Pemkot	189	34	537

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3155 entries, 0 to 3154
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   id                  3155 non-null   int64
 1   nama                3155 non-null   object
 2   propinsi            3155 non-null   object
 3   kab                 3155 non-null   object
 4   alamat              3155 non-null   object
 5   jenis               3155 non-null   object
 6   kelas               3155 non-null   object
 7   status_blu          3155 non-null   object
 8   kepemilikan         3155 non-null   object
 9   total_tempat_tidur  3155 non-null   int64
 10  total_layanan       3155 non-null   int64
 11  total_tenaga_kerja  3155 non-null   int64
dtypes: int64(4), object(8)
memory usage: 295.9+ KB
```

Hasil pengumpulan dan inspeksi data menunjukkan bahwa file **Hospital_Indonesia_datasets .csv** berhasil ditemukan dan dibaca dengan baik .

dari direktori C:\Users\Pongo\Downloads\Hospital Indonesia datasets.csv menggunakan pustaka **pandas** pada Python. File ini dibaca menggunakan **encoding UTF-8** dan **delimiter titik koma (;)**, dengan ukuran data sebesar **3.155 baris** dan **12 kolom**. Dataset tersebut berisi informasi lengkap mengenai daftar rumah sakit di seluruh Indonesia, mencakup atribut seperti *id* (kode rumah sakit), *nama*, *propinsi*, *kabupaten/kota*, *alamat*, *jenis rumah sakit*, *kelas*, *status_blu*, *kepemilikan*, *total_tempat_tidur*, *total_layanan*, dan *total_tenaga_kerja*. Berdasarkan hasil inspeksi awal terhadap lima baris pertama, terlihat bahwa data mencakup berbagai jenis rumah sakit dari Provinsi Aceh, seperti RS Arun Lhokseumawe, RS Umum Fandika, dan RS Umum Daerah Meuraxa. Hasil analisis struktur dataset melalui perintah df.info() menunjukkan bahwa seluruh kolom memiliki **3155 nilai non-null**, artinya **tidak terdapat nilai hilang (missing values)**. Dari segi tipe data, terdapat **4 kolom bertipe numerik (int64)** dan **8 kolom bertipe kategorikal (object)**, dengan total penggunaan memori sebesar sekitar **296 KB**

2. Penanganan Nilai Hilang (Missing Values)

Proses penanganan nilai hilang dilakukan berdasarkan tipe data setiap kolom. Untuk kolom **numerik** seperti *total_tempat_tidur* dan *total_tenaga_kerja*, digunakan metode **imputasi median** karena median lebih robust terhadap *outlier* dan tidak mengubah distribusi secara ekstrem.

```
Numerik: ['id', 'total_tempat_tidur', 'total_layanan', 'total_tenaga_kerja']
Kategorikal: ['nama', 'propinsi', 'kab', 'alamat', 'jenis', 'kelas', 'status_blu', 'kepemilikan']
Cek missing setelah imputasi:

id                  0
nama                0
propinsi            0
kab                 0
alamat              0
jenis               0
kelas               0
status_blu          0
kepemilikan         0
total_tempat_tidur  0
total_layanan       0
total_tenaga_kerja  0
dtype: int64
```

Hasil pemeriksaan pada gambar tersebut menunjukkan bahwa proses penanganan nilai hilang (*missing values*) pada dataset **Hospital_Indonesia_datasets.csv** telah berhasil dilakukan dengan baik. Data dibagi menjadi dua kelompok utama, yaitu fitur numerik dan fitur kategorikal. Fitur numerik meliputi kolom id, total_tempat_tidur, total_layanan, dan total_tenaga_kerja, yang berisi nilai angka terkait kapasitas dan sumber daya rumah sakit. Sementara itu, fitur kategorikal mencakup kolom nama, propinsi, kab, alamat, jenis, kelas, status_blu, dan kepemilikan, yang berisi informasi teks tentang identitas dan karakteristik setiap rumah sakit. Setelah dilakukan proses imputasi, di mana kolom numerik diisi dengan **nilai median** dan kolom kategorikal diisi dengan **nilai modus (mode)**, dilakukan pemeriksaan ulang menggunakan fungsi df.isnull().sum(). Hasilnya menunjukkan bahwa seluruh kolom memiliki nilai nol (0) pada jumlah data yang hilang, yang berarti **tidak ada lagi nilai kosong atau data yang hilang** dalam dataset.

3. Penanganan Data Kategorikal dan High Cardinality

```
Dimensi setelah OHE: (3151, 75)
```

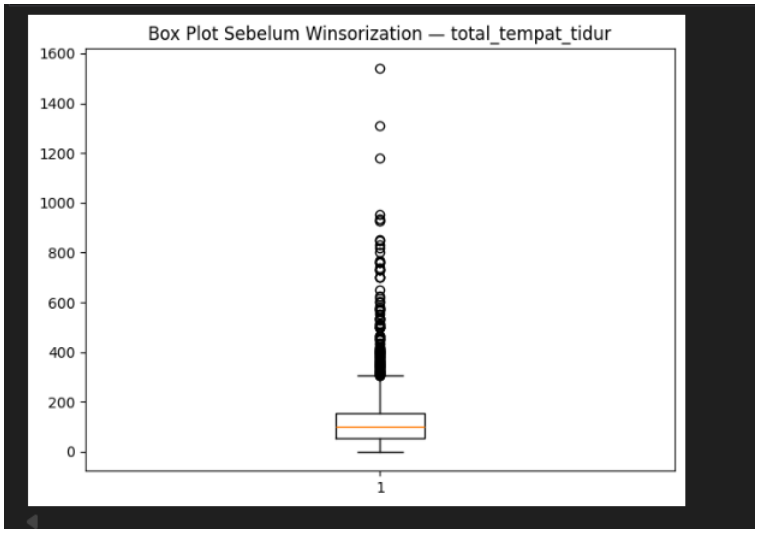
	id	total_tempat_tidur	total_layanan	total_tenaga_kerja	nama_Lain-Lain	propinsi_Aceh	propinsi_Bali	propinsi_Banten	propinsi_Dki Jakarta	propinsi_Jambi	...	kepemilikan_Organisasi Katholik	kepemilikan_Organisasi Sosial
0	1110053	218	36	328	1	1	0	0	0	0	...	0	0
1	1106014	45	15	45	1	1	0	0	0	0	...	0	0
2	1171110	310	77	487	1	1	0	0	0	0	...	0	0
3	1171163	11	24	0	1	1	0	0	0	0	...	0	0
4	1102027	189	34	537	1	1	0	0	0	0	...	0	0

menunjukkan keluaran dari proses penanganan data kategorikal dan high cardinality menggunakan metode One-Hot Encoding (OHE). Setelah dilakukan pengelompokan

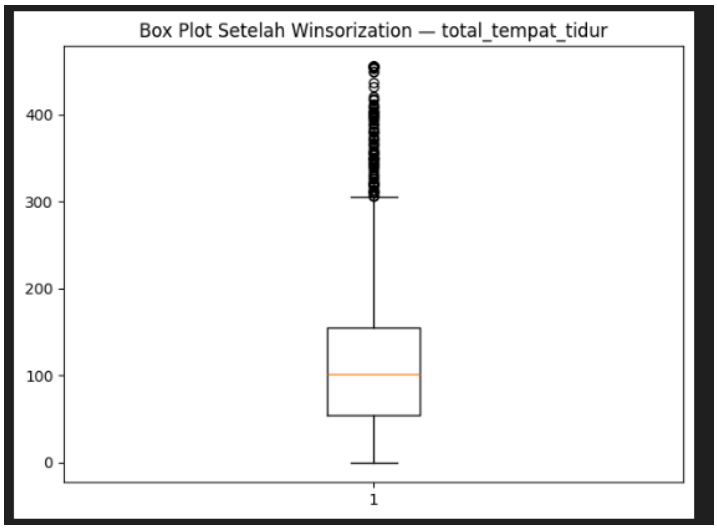
kategori dengan frekuensi rendah menjadi “Lain-Lain” dan proses konversi variabel kategorikal menjadi bentuk numerik biner, dataset kini memiliki 3.151 baris dan 75 kolom. Kolom-kolom teks seperti propinsi, kepemilikan, dan jenis telah diubah menjadi serangkaian kolom baru dengan nilai 0 dan 1, di mana angka 1 menunjukkan keberadaan kategori tertentu pada baris tersebut. Sebagai contoh, kolom propinsi_Aceh bernilai 1 menandakan rumah sakit tersebut berada di Provinsi Aceh, sedangkan kolom kepemilikan_Organisasi Sosial bernilai 0 berarti rumah sakit tersebut tidak termasuk dalam kategori kepemilikan tersebut

4. Penanganan Outlier

- Box Plot Sebelum Winsorization



- Box Plot Setelah Winsorization



menunjukkan hasil visualisasi **box plot** yang digunakan untuk mendeteksi dan menangani **outlier** pada variabel numerik total_tempat_tidur, yaitu jumlah tempat tidur di setiap rumah

sakit. Pada grafik pertama, **Box Plot Sebelum Winsorization**, terlihat bahwa terdapat banyak titik data yang berada jauh di atas *whisker* (garis batas atas), menunjukkan adanya nilai ekstrem yang sangat tinggi dibandingkan dengan distribusi umum data. Nilai-nilai ekstrem tersebut dapat mengganggu analisis karena berpotensi mendistorsi rata-rata dan variabilitas data.

Setelah dilakukan proses **Winsorization**, seperti yang ditunjukkan pada grafik kedua (**Box Plot Setelah Winsorization**), nilai-nilai ekstrem di bagian atas telah dibatasi hingga berada dalam kisaran yang wajar berdasarkan metode **Interquartile Range (IQR)**, yaitu nilai di luar batas atas (Q3 + 3×IQR) dan batas bawah (Q1 − 3×IQR) dipotong ke batas tersebut tanpa dihapus. Hasilnya, distribusi data menjadi lebih seimbang dan tidak terlalu dipengaruhi oleh nilai ekstrem, meskipun masih terdapat sedikit titik di luar batas yang dianggap wajar secara statistik

5. Penskalaan Fitur

	id	total_tempat_tidur	total_layanan	total_tenaga_kerja	nama_Lain-Lain	propinsi_Aceh	propinsi_Bali	propinsi_Banten	propinsi_Dki Jakarta	propinsi_Jambi	...	kepemilikan_Organisasi Katholik	kepemilikan_Organisasi Sosial	kepemilikan_Pemkab
1295	0.266339	0.028509	0.082707	0.073567	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
2737	0.649321	0.368421	0.338346	0.141146	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0
764	0.253425	0.173246	0.413534	0.277160	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0

3 rows × 75 columns

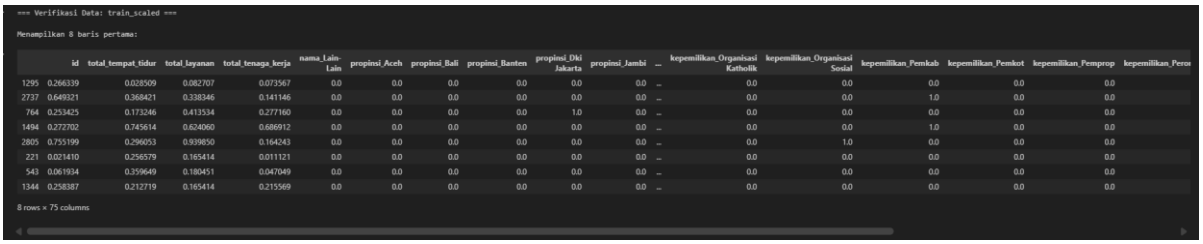
	id	total_tempat_tidur	total_layanan	total_tenaga_kerja	nama_Lain-Lain	propinsi_Aceh	propinsi_Bali	propinsi_Banten	propinsi_Dki Jakarta	propinsi_Jambi	...	kepemilikan_Organisasi Katholik	kepemilikan_Organisasi Sosial	kepemilikan_Pemkab
3141	1.000000	0.285088	0.345865	0.216424	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0
1839	0.295484	0.144737	0.210526	0.094953	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
2285	0.306260	0.223684	0.165414	0.085543	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0

3 rows × 75 columns

hasil akhir dari proses penskalaan fitur (feature scaling) menggunakan metode normalisasi Min-Max (MinMaxScaler). Pada tahap ini, seluruh fitur numerik dalam dataset — seperti id, total_tempat_tidur, total_layanan, dan total_tenaga_kerja — telah diubah ke dalam rentang nilai antara 0 hingga 1. Tujuan dari penskalaan ini adalah untuk menyeimbangkan skala antarvariabel sehingga tidak ada fitur yang mendominasi proses analisis statistik atau model *machine learning*.

Terlihat bahwa setiap baris data kini memiliki nilai desimal terstandarisasi, misalnya total_tempat_tidur bernilai 0.26 atau 0.64, yang menunjukkan posisi relatif nilai tersebut terhadap nilai minimum dan maksimum dari kolom aslinya. Selain itu, kolom hasil *One-Hot Encoding* seperti propinsi_Aceh, propinsi_Dki Jakarta, atau kepemilikan_Pemkab tetap bernilai 0 atau 1 karena merupakan variabel kategorikal yang tidak memerlukan penskalaan.

6. Hasil Akhir dan Dokumentasi



```
=== Verifikasi Data: train_scaled ===

Menampilkan 8 baris pertama:
```

	id	total_tempat_tidur	total_layanan	total_tenaga_kerja	nama_Lain-Lain	propinsi_Aceh	propinsi_Bali	propinsi_Banten	propinsi_Dki Jakarta	propinsi_Jambi	...	kepemilikan_Organisasi Katholik	kepemilikan_Organisasi Sosial	kepemilikan_Pemkab	kepemilikan_Pemkot	kepemilikan_Pemprop	kepemilikan_Perse
1295	0.266339	0.028509	0.082707	0.073567	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
2737	0.649321	0.368421	0.338346	0.141146	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	0.0	
764	0.253425	0.173246	0.413534	0.277160	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
1494	0.272702	0.745614	0.624060	0.686912	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	0.0	
2805	0.755199	0.296053	0.939850	0.164243	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0	
221	0.021410	0.256579	0.165414	0.011121	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
543	0.061934	0.359649	0.180451	0.047049	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
1344	0.258387	0.212719	0.165414	0.215569	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	

8 rows × 75 columns

```
--- Jumlah nilai hilang per kolom ---
id                                0
total_tempat_tidur                0
total_layanan                     0
total_tenaga_kerja                0
nama_Lain-Lain                    0
propinsi_Aceh                     0
propinsi_Bali                     0
propinsi_Banten                   0
propinsi_Dki Jakarta              0
propinsi_Jambi                    0
propinsi_Jawa Barat               0
propinsi_Jawa Tengah              0
propinsi_Jawa Timur               0
propinsi_Kalimantan Barat         0
propinsi_Kalimantan Selatan       0
propinsi_Kalimantan Tengah        0
propinsi_Kalimantan Timur         0
propinsi_Kepulauan Riau           0
propinsi_Lain-Lain                0
propinsi_Lampung                  0
propinsi_Nusa Tenggara Barat      0
propinsi_Nusa Tenggara Timur      0
propinsi_Riau                     0
...
- (tidak ada)
```

Verifikasi:

Berdasarkan hasil verifikasi akhir, seluruh proses pra-pemodelan terhadap dataset "Hospital_Indonesia_datasets" telah dilakukan dengan baik. Hasil pemeriksaan menunjukkan bahwa:

- Tidak terdapat lagi nilai hilang pada semua kolom setelah proses imputasi dan pembersihan.
- Kolom numerik seperti *total_tempat_tidur*, *total_layanan*, dan *total_tenaga_kerja* telah dikonversi menjadi tipe float64.
- Kolom kategorikal seperti *jenis*, *kelas*, dan *kepemilikan* telah terstandardisasi dan bebas dari nilai kosong.
- Data numerik telah melalui proses penskalaan (normalisasi) sehingga siap digunakan untuk analisis statistik atau model prediktif.

Kesimpulan

Secara keseluruhan, dataset rumah sakit di Indonesia telah melalui proses pembersihan, imputasi, encoding, penanganan outlier, dan penskalaan fitur dengan baik.
Dataset kini bersih, konsisten, dan siap digunakan untuk analisis distribusi rumah sakit berdasarkan provinsi dan kepemilikan, evaluasi ketersediaan layanan kesehatan,