

Atribut yang digabungkan :

1. Tanggal Didaftar
2. NIK
3. Umur
4. Jenis Kelamin
5. Lokasi Anatomi
6. Hasil Pemeriksaan Foto Toraks
7. Status HIV
8. Riwayat Diabetes Melitus
9. Hasil Pemeriksaan Diagnosis TBC
10. Jenis Pemeriksaan Diagnosis
11. Tipe Diagnosis

▼ PENGGABUNGAN DATA DAN DROP ATRIBUT YANG TIDAK DI BUTUHKAN

```
1 import pandas as pd
2 import numpy as np
3 from google.colab import drive
4
5 drive.mount('/content/drive')
6
7 file_paths = [
8     "/content/drive/MyDrive/Data/dataset_rs_udsa/RS_Umum_Daerah_Syarifah_Ambainii_Rato_Ebu_Report_TB_06_sulisrs_Januari-Desember",
9     "/content/drive/MyDrive/Data/dataset_rs_udsa/RS_Umum_Daerah_Syarifah_Ambainii_Rato_Ebu_Report_TB_06_sulisrs_Januari-Desember",
10    "/content/drive/MyDrive/Data/dataset_rs_udsa/RS_Umum_Daerah_Syarifah_Ambainii_Rato_Ebu_Report_TB_06_sulisrs_Januari-Desember",
11    "/content/drive/MyDrive/Data/dataset_rs_udsa/RS_Umum_Daerah_Syarifah_Ambainii_Rato_Ebu_Report_TB_06_sulisrs_Januari-Desember",
12    "/content/drive/MyDrive/Data/dataset_rs_udsa/RS_Umum_Daerah_Syarifah_Ambainii_Rato_Ebu_Report_TB_06_sulisrs_Januari-Desember"
13 ]
14
15 gabungan_df_data = []
16
17 for path in file_paths:
18     data_frame = pd.read_excel(
19         path,
20         skiprows=[x for x in range(0, 15)],
21         usecols="D,F,H,I,X:AA,AC,AD,AF",
22         names= ['Tanggal_Daftar', 'NIK', 'Umur', 'Jenis_Kelamin', 'Lokasi_Anatomi', 'Hasil_Pemeriksaan_Foto_Toraks', 'Status_HIV', ''],
23         dtype={
24             'Tanggal_Daftar': str,
25             'NIK': str,
26             'Umur': np.int32,
27             'Jenis_Kelamin': str,
28             'Lokasi_Anatomi': str,
29             'Hasil_Pemeriksaan': str,
30             'Hasil_Pemeriksaan_Foto_Toraks': str,
31             'Status_HIV': str,
32             'Riwayat_Diabetes_Melitus': str,
33             'Hasil_Pemeriksaan_Diagnosis_TBC': str,
34             'Tipe_Diagnosis': str
35         })
36     gabungan_df_data.append(data_frame)
37
38 gabungan_df_data = pd.concat(gabungan_df_data)
39
40 total_nik = gabungan_df_data['NIK'].value_counts()
41
42 gabungan_df_data
43
```

Mounted at /content/drive

	Tanggal_Daftar	NIK	Umur	Jenis_Kelamin	Lokasi_Anatomii	Hasil_Pemeriksaan	Foto_Toraks	Status_HIV	Riwayat
0	25/01/2020	3526022503520003	67	L	TBC Paru			Pos	Tidak Diketahui
1	28/05/2020	3526041402650003	55	L	TBC Paru			Pos	Tidak Diketahui
2	10/09/2020	3527091208620001	58	L	TBC Paru			Neg	Tidak Diketahui
3	24/06/2020	3526070709940004	25	L	TBC Paru			Neg	Tidak Diketahui
4	10/06/2020	3526061912930001	26	L	TBC Paru			Neg	Tidak Diketahui
...
1789	10/06/2024	3526036612040003	19	P	TBC Ekstraparu			Neg	Tidak Diketahui
1790	12/12/2024	3578164608730006	51	P	TBC Paru			TDL	Tidak Diketahui
1791	06/01/2024	3526064412170003	6	P	TBC Paru			Pos	Tidak Diketahui
1792	13/08/2024	3526141912970001	26	L	TBC Paru			Pos	Bukan ODHV
1793	05/04/2024	3528016608180001	5	P	TBC Paru			Neg	Tidak Diketahui

5378 rows × 11 columns

▼ DATA SETELAH DI GABUNGKAN DAN DI DROP ATRIBUT YANG TIDAK DI PERLUKAN

Tambahkanblockquote

```
1 from google.colab import sheets
2 sheet = sheets.InteractiveSheet(df=gabungan_df_data)
```

https://docs.google.com/spreadsheets/d/1N1t91gWaDPbXTug63QuMxoF123tD18sMrgwNwl1pQm0/edit#gid=0

	A1	Tanggal_Daftar							
	A	B	C	D	E	F	G	H	I
16	20/04/2020	6371021009870	32	L	TBC Ekstraparu	Neg	Tidak Diketahui	Tidak	
17	08/01/2020	3526020107760	43	L	TBC Paru	Pos	Tidak Diketahui	Tidak	Rif Sen
18	09/11/2020	3526052507970	23	L	TBC Paru	Neg	Tidak Diketahui	Tidak Diketahui	Neg
19	21/07/2020	3526130604870	33	L	TBC Paru	Pos	Tidak Diketahui	Tidak Diketahui	Rif Sen
20	19/08/2020	3526040101010	19	L	TBC Paru	Pos	Tidak Diketahui	Tidak Diketahui	
21	24/07/2020	3526011109010	18	L	TBC Paru	Neg	Tidak Diketahui	Tidak Diketahui	Neg
22	08/01/2020	3526131912170	2	L	TBC Paru	Pos	Tidak Diketahui	Tidak	
23	06/09/2020	3526042002680	52	L	TBC Paru	Pos	Tidak Diketahui	Tidak Diketahui	Rif Sen
24	10/08/2020	3526071709650	54	L	TBC Paru	Pos	Tidak Diketahui	Tidak Diketahui	Rif Sen
25	23/09/2020	3526141011810	38	L	TBC Paru	Pos	Bukan ODHV	Tidak	Rif Sen
26	16/06/2020	3526033112560	63	L	TBC Paru	Pos	Tidak Diketahui	Tidak	
27	16/11/2020	3526112705450	75	L	TBC Paru	Pos	Tidak Diketahui	Tidak Diketahui	Rif Sen
28	14/09/2020	3526021212670	52	L	TBC Paru	Pos	Tidak Diketahui	Tidak Diketahui	
29	14/09/2020	3526021212670	52	L	TBC Paru	Pos	Tidak Diketahui	Tidak Diketahui	
30	14/09/2020	3526021212670	52	L	TBC Paru	Pos	Tidak Diketahui	Tidak Diketahui	
31	14/11/2020	3526112001870	33	L	TBC Paru	Neg	Tidak Diketahui	Tidak Diketahui	Neg
32	09/06/2020	3526090101540	66	L	TBC Paru	Pos	Tidak Diketahui	Tidak	Neg
33	10/09/2020	3526092105860	34	L	TBC Paru	Pos	Bukan ODHV	Tidak	
34	18/02/2020	3526090107580	61	L	TBC Paru	Pos	Tidak Diketahui	Tidak	Neg
35	06/09/2020	3526010412630	54	L	TBC Paru	Neg	Tidak Diketahui	Tidak Diketahui	Neg

✓ Menghitung Jumlah Data, MISSING Value, dan Jumlah NIK yang sama

```

1 print("Jumlah Data: \n", gabungan_df_data.shape[0])
2 print()
3 print("Jumlah Missing Value Setiap attribut: \n")
4 print(gabungan_df_data.isnull().sum())
5 print("Total Jumlah Missing value: ", gabungan_df_data.isnull().sum().sum())
6 print()
7 print("Jumlah NIK yang sama: ", total_nik[total_nik > 1].sum())

```

⤵ Jumlah Data:
5378

Jumlah Missing Value Setiap attribut:

Tanggal_Daftar	0
NIK	0
Umur	0
Jenis_Kelamin	1
Lokasi_Anatomi	21
Hasil_Pemeriksaan_Foto_Toraks	438
Status_HIV	0
Riwayat_Diabetes_Melitus	0
Hasil_Pemeriksaan_Diagnosis_TBC	2247
Jenis_Pemeriksaan_Diagnosis	2247
Tipe_Diagnosis	3398
dtype: int64	
Total Jumlah Missing value:	8352

Jumlah NIK yang sama: 616

✓ MENGHAPUS BARIS YANG TERDAPAT MISSING VALUE

- DATA KOSONG
- ERROR

```

1 deleted_row_missing_value = gabungan_df_data.dropna()
2 deleted_row_missing_value = deleted_row_missing_value[deleted_row_missing_value['Hasil_Pemeriksaan_Diagnosis_TBC'] != 'ERROR']
3 deleted_row_missing_value

```

⤵

	Tanggal_Daftar	NIK	Umur	Jenis_Kelamin	Lokasi_Anatomi	Hasil_Pemeriksaan_Foto_Toraks	Status_HIV	Riwayat
0	25/01/2020	3526022503520003	67		L	TBC Paru	Pos	Tidak Diketahui
1	28/05/2020	3526041402650003	55		L	TBC Paru	Pos	Tidak Diketahui
7	05/02/2020	3526030507860004	33		L	TBC Ekstraparuh	Neg	Tidak Diketahui
8	14/05/2020	0000000000000000	26		L	TBC Paru	Pos	Tidak Diketahui
11	24/10/2020	3526011304610003	59		L	TBC Paru	Pos	Tidak Diketahui
...
1778	02/12/2024	3526164708080002	16		P	TBC Ekstraparuh	Neg	Bukan ODHIV
1779	14/08/2024	3526123112650062	58		L	TBC Paru	Pos	Tidak Diketahui
1784	23/12/2024	3526061308860002	38		L	TBC Paru	Pos	Bukan ODHIV
1788	10/04/2024	3526034104900008	34		P	TBC Paru	Pos	Bukan ODHIV
1792	13/08/2024	3526141912970001	26		L	TBC Paru	Pos	Bukan ODHIV

1046 rows × 11 columns

Langkah berikutnya: [Lihat plot yang direkomendasikan](#) [New interactive sheet](#)

Klik dua kali (atau tekan Enter) untuk mengedit

▼ DATA SETELAH DI HAPUS BARIS YANG TERDAPAT MISSING VALUE

```
1 sheet = sheets.InteractiveSheet(df=deleted_row_missing_value)

🔗 https://docs.google.com/spreadsheets/d/13Qj16Yn0Y9Kvxn7F66hY\_OEcUvA25Nwlcd-rB1RTe3w/edit#gid=0
```

File Edit Tampilan Sisipkan Format Data Alat Ekstensi Bantuan

A1	A	B	C	D	E	F	G	H
82	18/11/2020	3526090107530	67	L	TBC Paru	Pos	Tidak Diketahui	Rif S
83	22/12/2020	3526010611960	24	L	TBC Paru	Pos	Bukan ODHIV	Rif S
84	03/10/2020	3526032310030	16	L	TBC Paru	Pos	Tidak Diketahui	Rif S
85	28/05/2020	3526010309680	51	L	TBC Paru	Pos	Bukan ODHIV	Tidak
86	24/02/2020	3526042405610	58	L	TBC Paru	Pos	Tidak Diketahui	Neg
87	18/08/2020	3526093112610	58	L	TBC Paru	Pos	Tidak Diketahui	Neg
88	25/02/2020	3526032803630	56	L	TBC Paru	Pos	Tidak Diketahui	Rif S
89	23/03/2020	3526040711530	66	L	TBC Paru	Pos	Tidak Diketahui	Tidak
90	04/02/2020	3526130107780	41	L	TBC Ekstraparuu	Neg	Tidak Diketahui	Rif S
91	17/04/2020	3526092511950	24	L	TBC Paru	Pos	Tidak Diketahui	Tidak Diketahui
92	04/10/2020	3526033011540	65	L	TBC Paru	Pos	Tidak Diketahui	Tidak Diketahui
93	21/07/2020	3526032507700	49	L	TBC Paru	Pos	Tidak Diketahui	Neg
94	09/03/2020	3526064107650	54	P	TBC Paru	Pos	Tidak Diketahui	Neg
95	30/03/2020	3526061512890	29	L	TBC Paru	Pos	Tidak Diketahui	Neg
96	17/02/2020	3526166503620	57	P	TBC Paru	Pos	Tidak Diketahui	Rif S
97	09/03/2020	3526037112470	73	P	TBC Paru	Pos	Tidak Diketahui	Neg
98	26/11/2020	3526024507750	45	P	TBC Paru	Pos	Tidak Diketahui	Tidak Diketahui
99	09/03/2020	3526020510690	50	L	TBC Paru	Pos	Tidak Diketahui	Tidak
100	28/07/2020	3526055308920	26	P	TBC Paru	Pos	Tidak Diketahui	Rif S
101	23/01/2020	3526160112520	67	I	TBC Paru	Pos	Tidak Diketahui	Tidak

+ ⏪ Sheet1 ⏴ ⏵

▼ TRANSFORMASI DATA DARI KATEGORIKAL KE NUMMERIK

- Jenis Kelamin
L = 0
P = 1
- Lokasi Anatomi
TBC E = 0
TBC P = 1
- Hasil Pemeriksaan Foto Toraks
Neg = 0
Pos = 1
TDL = 2
- Status HIV
Bukan ODHIV = 0
ODHIV = 1
Tidak Diketahui = 2
- Riwayat Diabetes Melitus
Tidak = 0
Tidak Diketahui = 1
Ya = 2
- Hasil Pemeriksaan Diagnosis TBC
Neg = 0
Rif Indet = 1
Rif Res = 2
Rif Sen = 3

- Jenis Pemeriksaan Diagnosis

TCM Xpert = 0

- Tipe Diagnosis

Terdiagnosis klinis = 0

Terkonfirmasi bakteriologis = 1

```

1 jk = deleted_row_missing_value["Jenis_Kelamin"].unique()
2 la = deleted_row_missing_value["Lokasi_Anatomii"].unique()
3 hp = deleted_row_missing_value["Hasil_Pemeriksaan_Foto_Toraks"].unique()
4 sh = deleted_row_missing_value["Status_HIV"].unique()
5 rdm = deleted_row_missing_value["Riwayat_Diabetes_Melitus"].unique()
6 hpdt = deleted_row_missing_value["Hasil_Pemeriksaan_Diagnosis_TBC"].unique()
7 jpd = deleted_row_missing_value["Jenis_Pemeriksaan_Diagnosis"].unique()
8 td = deleted_row_missing_value["Tipe_Diagnosis"].unique()
9
10 print(sorted(jk))
11 print(sorted(la))
12 print(sorted(hp))
13 print(sorted(sh))
14 print(sorted(rdm))
15 print(sorted(hpdt))
16 print(sorted(jpd))
17 print(sorted(td))

```

→ ['L', 'P']
 ['TBC Ekstraparu', 'TBC Paru']
 ['Neg', 'Pos', 'TDL']
 ['Bukan ODHIV', 'ODHIV', 'Tidak Diketahui']
 ['Tidak', 'Tidak Diketahui', 'Ya']
 ['Neg', 'Rif Indet', 'Rif Res', 'Rif Sen']
 ['TCM Xpert']
 ['Terdiagnosis klinis', 'Terkonfirmasi bakteriologis']

```

1 def transformToDict(list_cat):
2     result = {}
3     for i in range(len(list_cat)):
4         result[list_cat[i]] = i
5
6     return result
7
8 deleted_row_missing_value['Jenis_Kelamin'] = deleted_row_missing_value['Jenis_Kelamin'].map(transformToDict(sorted(jk)))
9 deleted_row_missing_value['Lokasi_Anatomii'] = deleted_row_missing_value['Lokasi_Anatomii'].map(transformToDict(sorted(la)))
10 deleted_row_missing_value['Hasil_Pemeriksaan_Foto_Toraks'] = deleted_row_missing_value['Hasil_Pemeriksaan_Foto_Toraks'].map(transformToDict(sorted(hp)))
11 deleted_row_missing_value['Status_HIV'] = deleted_row_missing_value['Status_HIV'].map(transformToDict(sorted(sh)))
12 deleted_row_missing_value['Riwayat_Diabetes_Melitus'] = deleted_row_missing_value['Riwayat_Diabetes_Melitus'].map(transformToDict(sorted(rdm)))
13 deleted_row_missing_value['Hasil_Pemeriksaan_Diagnosis_TBC'] = deleted_row_missing_value['Hasil_Pemeriksaan_Diagnosis_TBC'].map(transformToDict(sorted(hpdt)))
14 deleted_row_missing_value['Jenis_Pemeriksaan_Diagnosis'] = deleted_row_missing_value['Jenis_Pemeriksaan_Diagnosis'].map(transformToDict(sorted(jpd)))
15 deleted_row_missing_value['Tipe_Diagnosis'] = deleted_row_missing_value['Tipe_Diagnosis'].map(transformToDict(sorted(td)))
16

```

▼ DATA SETELAH TRANSFORMASI DAN CLEANING

```
1 sheet = sheets.InteractiveSheet(df=deleted_row_missing_value)
```

https://docs.google.com/spreadsheets/d/1igClzhyIEmK5tyG1DWvKgTKh4vbTP30R_ZIF1smRT4/edit#gid=0

File Edit Tampilan Sisipkan Format Data Alat Ekstensi Bantuan

Q Menu ↲ ⌛ 100% \$ % .00 123 Default... - + B I A 🔍 ⌂ ⌂

G111 fx 1

	A	B	C	D	E	F	G	H	I
43	19/02/2020	35260111044800	71			1	2	0	3
44	30/11/2020	35260852071500	5			2	2	1	3
45	28/07/2020	35260801075800	62			1	2	0	3
46	03/11/2020	35261113049200	28			1	2	1	3
47	18/02/2020	36031607075200	67			1	2	0	3
48	12/10/2020	35270941076800	60			1	0	0	3
49	18/08/2020	35260456116300	56			1	2	1	2
50	08/01/2020	35260471126000	61			1	2	0	3
51	11/08/2020	35261207090200	17			1	2	1	3