**LAPORAN TUGAS**

**PENGOLAHAN DATA TB (TUBERKULOSIS)**

**KECERDASAN KOMPUTASIONAL**

**Dosen pengampu**

**Eka Mala Sari Rochman, S.Kom., M.Kom.**

****

**Disusun oleh Kelompok 5**

* **Alaikassalam : 230411100133**
* **Ivanka Ramadhan : 230411100202**
* **Bilqizah Rahma Ilayya Syahdewi : 230411100030**
* **, Adhitiya Aprillio Pratama : 230411100201**
* **Sirojul Izzeh : 230411100131**

**PRODI TEKNIK INFORMATIKA**

**JURUSAN TEKNIK INFORMATIKA  
FAKULTAS TEKNIK**

**UNIVERSITAS TRUNOJOYO MADURA**

**2025**

# Daftar Isi

[Daftar Isi i](#_Toc198145371)

[1. Deskripsi tugas 1](#_Toc198145372)

[2. Penyelesain tugas 2](#_Toc198145373)

[2.1. Penggabungan seluruh file data exel 2](#_Toc198145374)

[2.2. Seleksi atribut penting yang akan digunakan untuk analisis 3](#_Toc198145375)

[2.1. Jumlah seluruh data (record) 4](#_Toc198145376)

[2.2. Jumlah missing value setiap atribut 5](#_Toc198145377)

[2.3. Jumlah data yang memiliki NIK duplikat 5](#_Toc198145378)

# Deskripsi tugas

Laporan ini disusun untuk mendokumentasikan tahapan pengolahan data TB (Tuberkulosis) berdasarkan data Excel dari RSUD. Proses yang dilakukan meliputi :

* Penggabungan seluruh file data Excel.
* Seleksi atribut penting yang akan digunakan untuk analisis.

1. Tanggal Didaftar

2. NIK (Nomor Identitas Kependudukan)

3. Umur (Tahun)

4. Jenis Kelamin

5. Lokasi Anatomi Penyakit

6. Hasil Pemeriksaan Foto Toraks

7. Status HIV

8. Riwayat Diabetes Melitus

9. Hasil Pemeriksaan Diagnosis TBC

10. Jenis Pemeriksaan Diagnosis

11. Tipe Diagnosis

* Jumlah seluruh data (record)
* Jumlah missing value setiap atribut
* Jumlah data yang memiliki NIK duplikat

File yang perlu dikumpulkan :

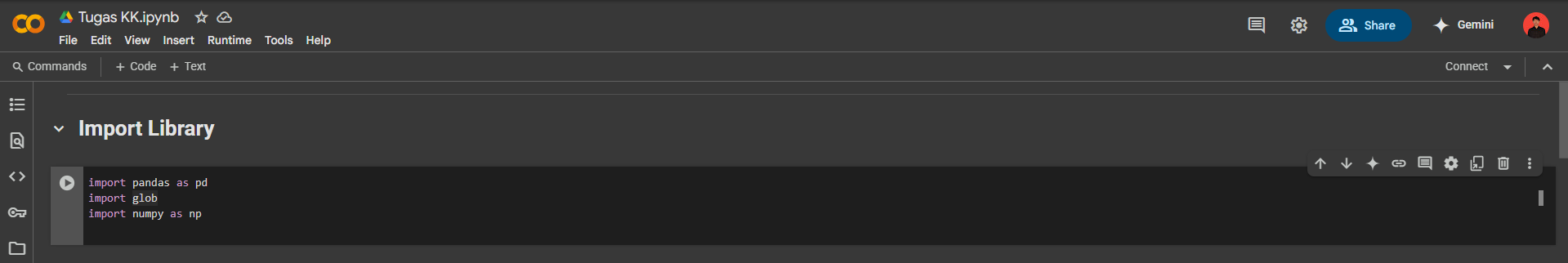
* File Excel/csv data yg sudah di drop atribut yg tidak dibutuhkan.
* File Excel/csv data yg sdh di preprocessing (clean, transformasi).
* Laporan yg berisi tugas 1-2

# Penyelesain tugas

Pada tahap ini akan melakukan preprocessing dengan menggunakan Bahasa python. Dimana akan membutuhkan library untuk digunakan mengolah data tabular (seperti tabel), numpy digunakan untuk operasi numerik, dan glob digunakan untuk mencari file dengan pola tertentu.

|  |
| --- |
| import pandas as pd  import glob  import numpy as np |

Output :



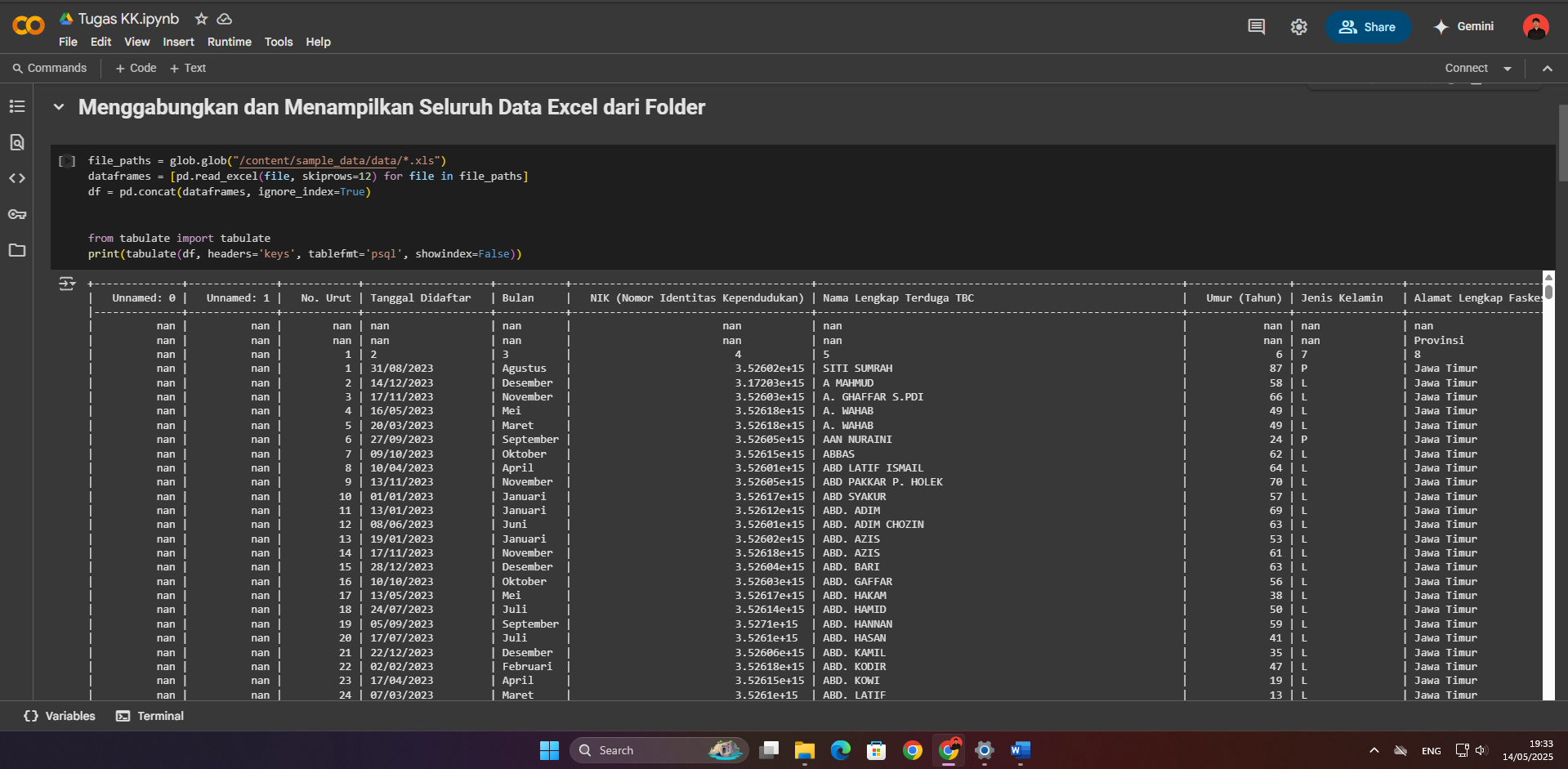
Tahapan selanjutnya itu penggabungan seluruh file data exel

## Penggabungan seluruh file data exel

Data yang akan diolah disini ada lima data real format .exel. Pada tahap ini seluruh data akan di gabung menjadi satu dengan menggunakan program python dibawah ini :

|  |
| --- |
| file\_paths = glob.glob("/content/sample\_data/data/\*.xls")  dataframes = [pd.read\_excel(file, skiprows=12) for file in file\_paths]  df = pd.concat(dataframes, ignore\_index=True)  from tabulate import tabulate  print(tabulate(df, headers='keys', tablefmt='psql', showindex=False)) |

Output :



## Seleksi atribut penting yang akan digunakan untuk analisis

Setelah semua data .exel digabung menjadi satu, Langkah selanjutnya itu seleksi berdasarkan atribut. Atribut yang diperlukan disini  "Tanggal Didaftar", "NIK (Nomor Identitas Kependudukan)", "Umur (Tahun)", "Jenis Kelamin", "Lokasi Anatomi Penyakit", "Hasil Pemeriksaan Foto Toraks", "Status HIV", "Riwayat Diabetes Melitus", "Hasil Pemeriksaan Diagnosis TBC", "Jenis Pemeriksaan Diagnosis", "Tipe Diagnosis". Untuk program python berikut :

|  |
| --- |
| kolom\_yang\_dipertahankan = [      "Tanggal Didaftar",      "NIK (Nomor Identitas Kependudukan)",      "Umur (Tahun)",      "Jenis Kelamin",      "Lokasi Anatomi Penyakit",      "Hasil Pemeriksaan Foto Toraks",      "Status HIV",      "Riwayat Diabetes Melitus",      "Hasil Pemeriksaan Diagnosis TBC",      "Jenis Pemeriksaan Diagnosis",      "Tipe Diagnosis"  ]  df\_clean = df[kolom\_yang\_dipertahankan].copy()  df\_clean = df\_clean.dropna(subset=["NIK (Nomor Identitas Kependudukan)"])  df\_clean = df\_clean.drop\_duplicates(subset=["NIK (Nomor Identitas Kependudukan)"])  df\_clean["Jenis Kelamin"] = df\_clean["Jenis Kelamin"].str.upper().replace({      "L": "LAKI-LAKI",      "P": "PEREMPUAN",      "LK": "LAKI-LAKI",      "PR": "PEREMPUAN"  })  df\_clean["Tanggal Didaftar"] = pd.to\_datetime(df\_clean["Tanggal Didaftar"], errors="coerce")  df\_clean["Status HIV"] = df\_clean["Status HIV"].fillna("Tidak Diketahui")  df\_clean.to\_excel("data\_tbc\_clean.xlsx", index=False)  print(f"Data awal: {df.shape[1]} kolom | Data bersih: {df\_clean.shape[1]} kolom")  print("Kolom yang tersisa:")  print(df\_clean.columns.tolist())  from tabulate import tabulate  print(tabulate(df\_clean, headers='keys', tablefmt='grid')) |

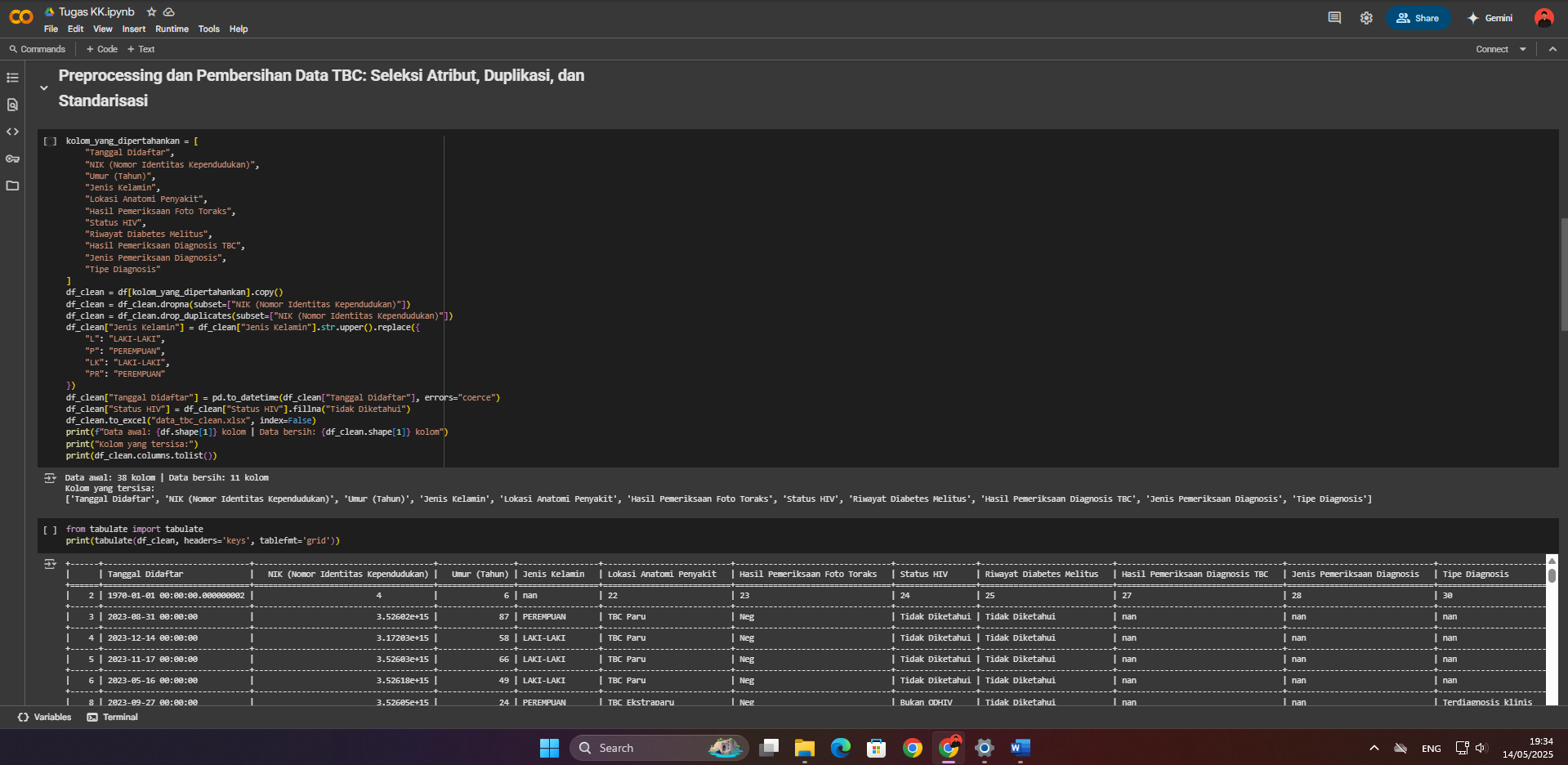
Untuk menyimpan file yang sudah diseleksi berdasarkan atribut terntentu menggunaka code program berikut :

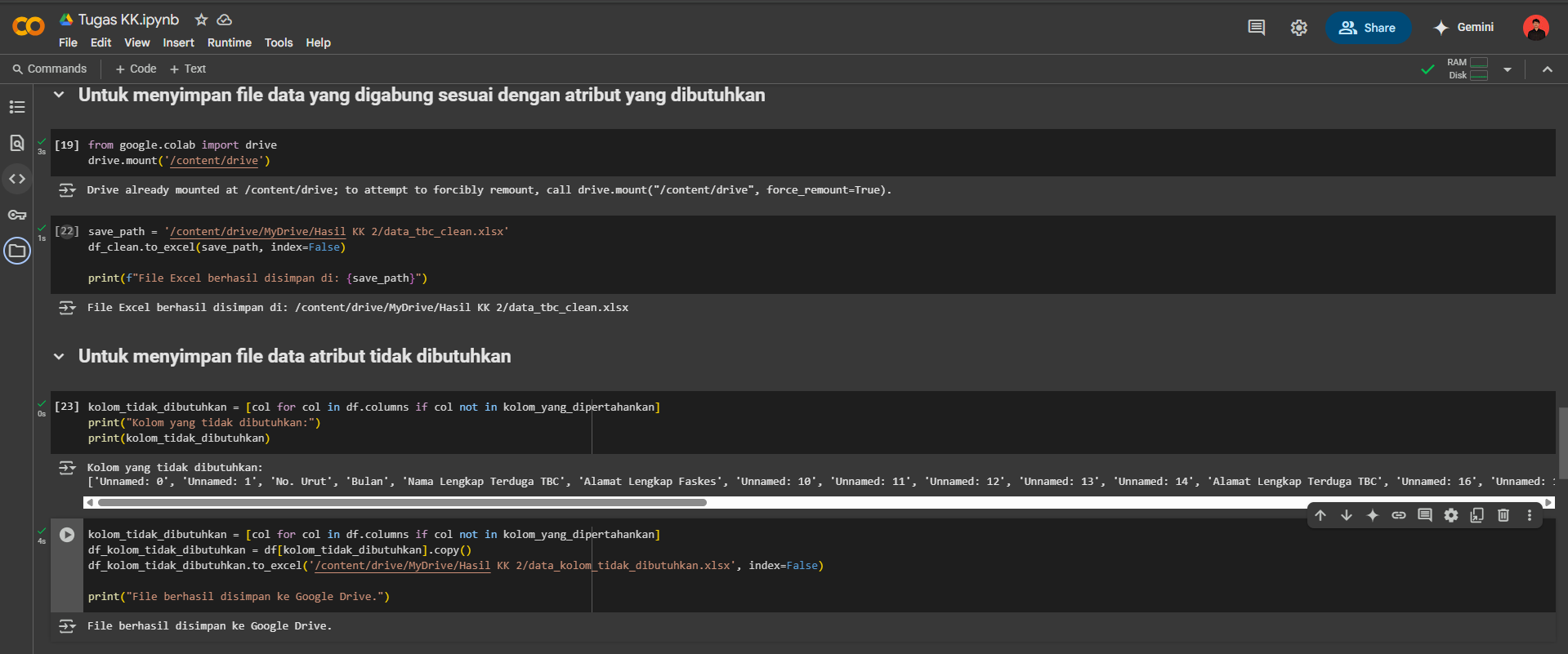
|  |
| --- |
| from google.colab import drive  drive.mount('/content/drive')  save\_path = '/content/drive/MyDrive/Hasil KK 2/data\_tbc\_clean.xlsx'  df\_clean.to\_excel(save\_path, index=False)  print(f"File Excel berhasil disimpan di: {save\_path}") |

Untuk menyimpan file data atribut yang tidak dibutuhkan setelah digabung berdasarkan atribut yang diminta maka menggunakan kode program berikut :

|  |
| --- |
| kolom\_tidak\_dibutuhkan = [col for col in df.columns if col not in kolom\_yang\_dipertahankan]  print("Kolom yang tidak dibutuhkan:")  print(kolom\_tidak\_dibutuhkan)  kolom\_tidak\_dibutuhkan = [col for col in df.columns if col not in kolom\_yang\_dipertahankan]  df\_kolom\_tidak\_dibutuhkan = df[kolom\_tidak\_dibutuhkan].copy()  df\_kolom\_tidak\_dibutuhkan.to\_excel('/content/drive/MyDrive/Hasil KK 2/data\_kolom\_tidak\_dibutuhkan.xlsx', index=False)  print("File berhasil disimpan ke Google Drive.") |

Output :



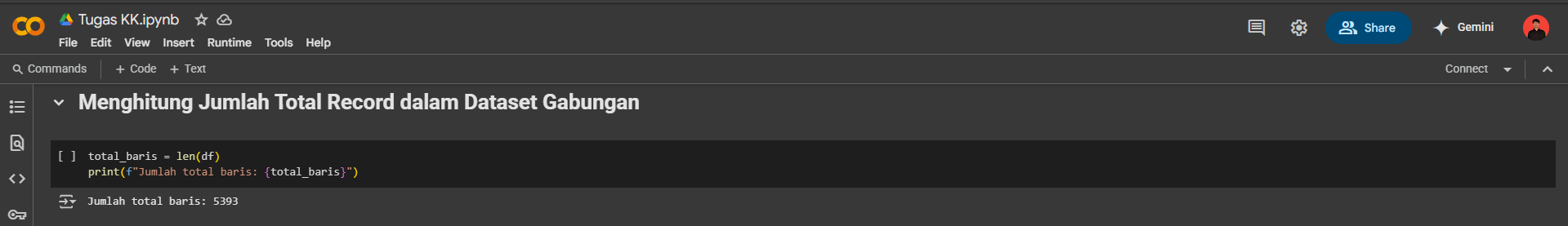


## Jumlah seluruh data (record)

Selanjutnya akan menghitung jumlah seluruh data (record) yang telah di seleksi dengan menggunakan code program berikut :

|  |
| --- |
| total\_baris = len(df)  print(f"Jumlah total baris: {total\_baris}") |

Output :

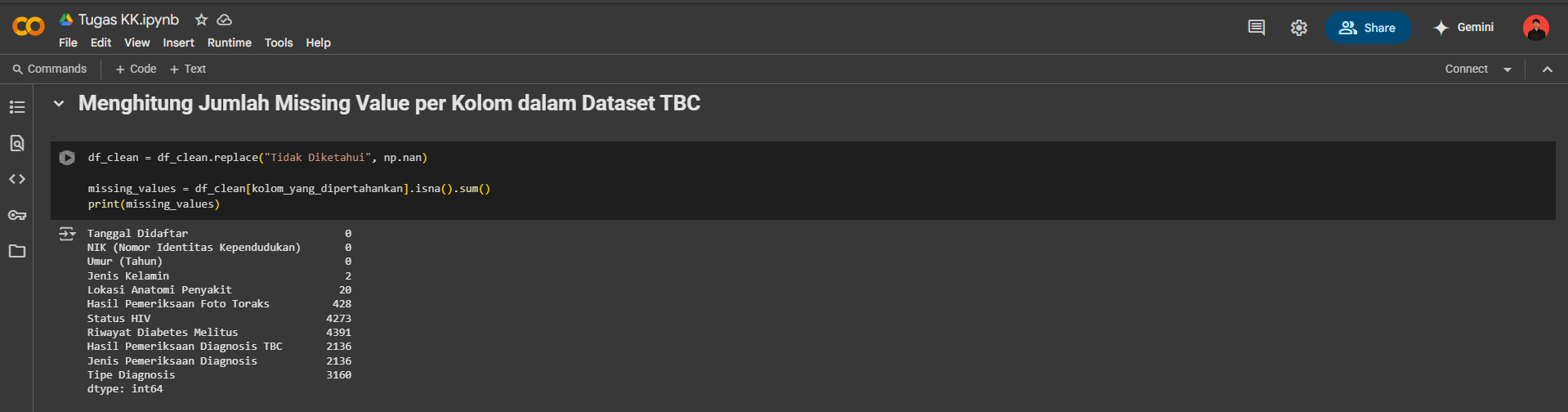


## Jumlah missing value setiap atribut

Tahapan selanjut untuk menghitung jumlah Missing Value setiap atribut menggunakan kode program berikut :

|  |
| --- |
| df\_clean = df\_clean.replace("Tidak Diketahui", np.nan)  missing\_values = df\_clean[kolom\_yang\_dipertahankan].isna().sum()  print(missing\_values) |

Output :



## Jumlah data yang memiliki NIK duplikat

Tahapan terakhir yang diminta adalah menghitung jumlah data yang memiliki NIK sama/duplikat dengan menggunakan kode program berikut :

|  |
| --- |
| from tabulate import tabulate  if 'NIK (Nomor Identitas Kependudukan)' in df.columns:      duplikasi = df.duplicated(subset=['NIK (Nomor Identitas Kependudukan)'], keep=False)      jumlah\_duplikat = duplikasi.sum()      print(f"Jumlah data duplikat berdasarkan NIK: {jumlah\_duplikat}")      data\_duplikat = df[duplikasi]        print("\nContoh 20 baris data yang terduplikat berdasarkan NIK:")      print(tabulate(data\_duplikat.head(5000), headers='keys', tablefmt='psql', showindex=False)) |

Output :

