DATA PREPARATION

Muhammad Habil Aswad
2208107010013

Rafli Afriza Nugraha
2208107010023

Muhammad Khalid
Al Ghifari
2208107010044

Muhammad Ridho
2208107010064

Muhammad Ilzam
2208107010087

KELOMPOK 10

Dataset

Dataset "**Earthquakes in Indonesia**" adalah kumpulan data yang berisi catatan kejadian gempa bumi di Indonesia pada 2008-2024.

Berjumlah **92.887** data. Mencakup **13 fitur** antara lain tanggal dan waktu kejadian, lokasi geografis (latitude dan longitude), kedalaman, serta magnitudo gempa. Selain itu, terdapat juga beberapa parameter mekanisme sumber gempa (strike, dip, dan rake) yang menggambarkan bagaimana patahan bumi bergerak saat gempa terjadi.

kolom mag (magnitudo) atau remark bisa digunakan sebagai target untuk klasifikasi

Link Dataset : https://www.kaggle.com/datasets/kekavigi/earthquakes-in-indonesia?select=katalog_gempa.csv

Data Loading

Dataset yang digunakan berasal dari file "**katalog_gempa.csv**", berisi data kejadian gempa bumi di Indonesia. Data ini dimuat ke Python menggunakan **library Pandas** untuk analisis dan manipulasi secara efisien.

1. Mengimpor Library

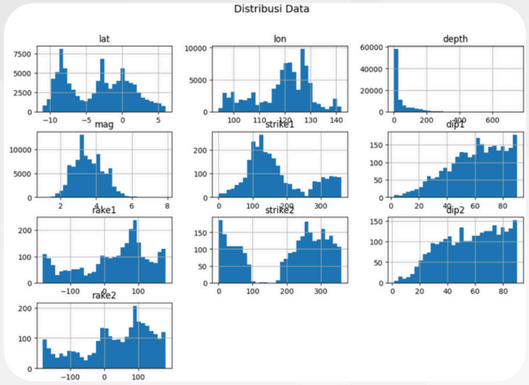
```
import pandas as pd
```

2. Membaca Dataset

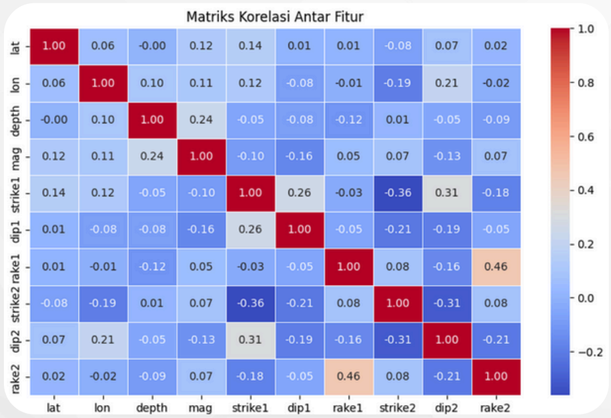
Dataset dimuat ke DataFrame Pandas menggunakan fungsi **pd.read_csv()**. Sehingga data dapat diakses dan dimanipulasi dengan lebih mudah.

```
# Memuat dataset dari file CSV
df = pd.read_csv("katalog_gempa.csv")
```

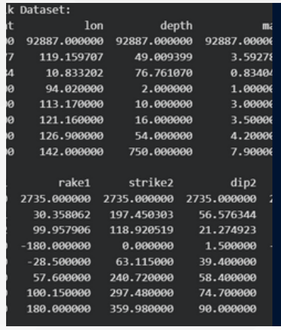
Data Understanding



- **Magnitudo (mag)**: Mayoritas gempa berkisar antara 3–6 (distribusi mendekati normal).
- **Latitude (lat) & Longitude (lon)**: Sebaran gempa sesuai dengan wilayah Indonesia.
- **Kedalaman (depth)**: Lebih banyak gempa terjadi di kedalaman dangkal (<100 km).



Kebanyakan fitur tidak berkorelasi satu dengan lainnya, hanya fitur **rake1** dan **rake 2** yang memiliki korelasi cukup tinggi yaitu 0.46.

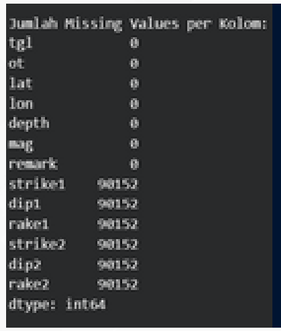


```
k Dataset:
t            lon         depth          ma
0   92887.000000  92887.000000  92887.00000
7     119.159707     49.009399      3.59278
4      10.833202     76.761070      0.83404
0      94.020000      2.000000      1.00000
0     113.170000     10.000000      3.00000
0     121.160000     16.000000      3.50000
0     126.900000     54.000000      4.20000
0     142.000000    750.000000      7.90000

          rake1       strike2          dip2
2735.000000  2735.000000  2735.000000  2
     30.358062   197.450303    56.576344
     99.957906   118.920519    21.274923
   -180.000000     0.000000     1.500000
    -28.500000    63.115000    39.400000
     57.600000   240.720000    58.400000
    100.150000   297.480000    74.700000
    180.000000   359.980000    90.000000
```

Statistik Dasar Dataset :

- Magnitudo (mag): Berkisar dari nilai terendah hingga tertinggi, dengan rata-rata sekitar nilai mean.
- Kedalaman (depth): Bervariasi dari sangat dangkal hingga sangat dalam.
- Latitude (lat) & Longitude (lon): Menunjukkan sebaran gempa di berbagai wilayah Indonesia.



```
Jumlah Missing Values per Kolom:
tgl             0
ot              0
lat             0
lon             0
depth           0
mag             0
remark          0
strike1     90152
dip1        90152
rake1       90152
strike2     90152
dip2        90152
rake2       90152
dtype: int64
```

Missing Values :

- Kolom mekanisme sumber gempa (strike1, dip1, dll.) memiliki banyak missing values.
- Kolom utama seperti lat, lon, depth, dan mag umumnya lengkap.

Insight dari Eksplorasi Data :

- Mayoritas gempa memiliki magnitudo 3–6, sedangkan gempa besar (>7) jarang.
- Sebaran lokasi gempa mengikuti Cincin Api Pasifik, terutama di Sumatra, Jawa, Sulawesi, dan Papua.
- Aktivitas tektonik dangkal mendominasi (kedalaman <100 km).
- Mekanisme patahan beragam, sesuai dengan pergerakan lempeng di Indonesia.
- Dataset siap untuk analisis karena tidak ada missing values pada kolom utama.

Data Preparation

Dilakukan **preprocessing** untuk menyesuaikan dataset agar optimal untuk analisis. Jika digunakan untuk **machine learning**, penanganan data kosong atau tidak relevan sangat penting. Beberapa langkah preprocessing yang dilakukan antara lain sebagai berikut:

1 Memisahkan Tahun, Bulan, dan Tanggal

- Kolom **tgl** dipecah menjadi **tahun, bulan, dan tanggal** dengan **.str.split("/")**
- Ubah ke integer (**.astype(int)**) agar mudah dianalisis.
- Kolom **tgl** asli dihapus karena sudah dipisahkan.

```
# Memisahkan tahun, bulan, dan tanggal
df[['tahun', 'bulan', 'tanggal']]
= df['tgl'].str.split('/', expand=True)
df[['tahun', 'bulan', 'tanggal']]
= df[['tahun', 'bulan', 'tanggal']].astype(int)

df.drop(columns=['tgl'], inplace=True)
```

2 Konversi Kolom Kategorikal ke Numerik

Kolom **remark** yang sebelumnya bertipe **Object** diubah menjadi **numerik** menggunakan **Label Encoding** agar bisa digunakan dalam model machine learning.

```
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()
df['remark'] = le.fit_transform(df['remark'])
```

```
df.head()
✓ 0.0s
```

	ot	lat	lon	depth	mag	remark	strike1	dip1	rake1	strike2	dip2	ra
0	21:02:43.058	-9.18	119.06	10	4.9	43	NaN	NaN	NaN	NaN	NaN	N
1	20:58:50.248	-6.55	129.64	10	4.6	4	NaN	NaN	NaN	NaN	NaN	N
2	17:43:12.941	-7.01	106.63	121	3.7	15	NaN	NaN	NaN	NaN	NaN	N
3	16:24:14.755	-3.30	127.85	10	3.2	32	NaN	NaN	NaN	NaN	NaN	N
4	16:20:37.327	-6.41	129.54	70	4.3	4	NaN	NaN	NaN	NaN	NaN	N

```
print(f"Jumlah Data: {len(df)}")
print(f"Jumlah Baris dengan Missing Value: {df.isnull().any(axis=1).sum()}")
✓ 0.0s

Jumlah Data: 92887
Jumlah Baris dengan Missing Value: 90152
```

3 Pemisahan Dataframe untuk Fleksibilitas Analisis

Dataset dibagi menjadi **df** (lengkap) untuk eksplorasi, **df1** (umum) untuk machine learning yang lebih bersih, dan **df2** (khusus) dengan fitur seismik lengkap. Pemisahan ini menjaga fleksibilitas analisis tanpa kehilangan data penting.

```
df1 = df.drop(columns=['dip1', 'strike1', 'rake1', 'dip2',

# Jika untuk membuat model clustering, maka fitur waktu tida
df1.drop(columns=['ot'], inplace=True)

df1.head()
✓ 0.0s
```

	lat	lon	depth	mag	remark	tahun	bulan	tanggal
0	-9.18	119.06	10	4.9	43	2008	11	1
1	-6.55	129.64	10	4.6	4	2008	11	1
2	-7.01	106.63	121	3.7	15	2008	11	1
3	-3.30	127.85	10	3.2	32	2008	11	1

```
# Jika untuk membuat model clustering, maka fitur waktu tidak dibutuhkan
df2.drop(columns=['ot'], inplace=True)

df2.head()
✓ 0.0s
```

	lat	lon	depth	mag	remark	strike1	dip1	rake1	strike2	dip2	rake2	tahun	bulan	tanggal
8739	-9.28	122.46	139	5.7	31	343.7	43.0	-168.1	244.97	81.9	-47.6	2011	6	27
8801	4.37	97.52	10	3.4	28	197.1	40.0	-55.6	335.29	58.0	-115.4	2011	7	4
8804	1.45	96.95	13	4.8	29	333.8	75.6	-128.5	226.40	40.7	-22.4	2011	7	5
8910	-1.27	137.83	10	5.1	21	100.5	71.9	95.8	262.47	18.9	72.9	2011	7	19
9026	-3.11	130.92	10	4.8	32	110.9	16.3	-146.7	348.74	81.1	-76.2	2011	8	6

```
print(f"Jumlah Data: {len(df2)}")
print(f"Jumlah Baris dengan Missing Value: {df2.isnull().any(axis=1).sum()}")
✓ 0.0s
```