DATA PREPARATION DARI SUMBER OPEN SOURCE

disusun untuk memenuhi tugas

mata kuliah Pembelajaran mesin

Oleh:

Kelompok 10

Anggota :

Muhammad Habil Aswad	(2208107010013)
Rafli Afriza Nugraha	(2208107010028)
Muhammad Khalid Al Ghifari	(2208107010044)
Muhammad Ridho	(2208107010064)
Muhammad Ilzam	(2208107010087)

JURUSAN INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SYIAH KUALA

DARUSSALAM, BANDA ACEH

2025

BAB I

PENDAHULUAN

1.1 Latar Belakang

Dalam era digital saat ini, data menjadi salah satu aset paling berharga dalam berbagai bidang, termasuk bisnis, kesehatan, keuangan, dan ilmu pengetahuan. Namun, sebelum data dapat digunakan untuk analisis atau pelatihan model machine learning, diperlukan proses persiapan data yang tepat. Data yang tidak bersih atau tidak terstruktur dapat menyebabkan hasil analisis yang tidak akurat dan berdampak negatif pada keputusan yang diambil. Oleh karena itu, pemahaman mendalam tentang teknik data preparation sangat penting untuk memastikan kualitas data yang optimal sebelum digunakan dalam proses lebih lanjut.

Tugas ini bertujuan untuk memberikan pengalaman langsung dalam mengolah data dari sumber open source, seperti Kaggle dan Hugging Face. Dengan melakukan serangkaian tahap data preparation, diharapkan dapat memahami pentingnya proses ini serta menerapkan teknik yang sesuai untuk meningkatkan kualitas data yang digunakan dalam analisis atau model machine learning.

1.2 Tujuan

Adapun tujuan dari tugas ini adalah sebagai berikut:

1. Memahami proses pemilihan dan pemuatan dataset dari sumber open source.
2. Menganalisis struktur dan karakteristik dataset yang digunakan.
3. Melakukan eksplorasi data awal untuk menemukan pola dan insight penting.
4. Menerapkan teknik preprocessing untuk meningkatkan kualitas dataset, termasuk penanganan missing values, encoding, normalisasi, dan feature selection.
5. Mendokumentasikan proses data preparation secara sistematis dalam bentuk laporan.

BAB II

PEMBAHASAN

2.1 Deskripsi Dataset

Nama Dataset : **Earthquakes in Indonesia**

Sumber : **Kaggle**

Dataset "Earthquakes in Indonesia" adalah kumpulan data yang berisi catatan kejadian gempa bumi di Indonesia. Data ini mencakup informasi seperti 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.

2.2 Sampel Data

- **Jumlah Sampel** : 92.887
- **Jumlah Fitur** : 13
- **Label** : Dataset ini tidak memiliki label eksplisit untuk tugas klasifikasi. Tetapi, kolom **mag** atau **remark** bisa digunakan sebagai target dalam analisis tertentu, seperti klasifikasi gempa berdasarkan wilayah atau magnitudo.
- **Format Data** : CSV (Comma-Separated Values)

2.3 Data Loading

Dalam penelitian ini, data yang digunakan berasal dari file CSV bernama "**katalog_gempa.csv**". File ini berisi informasi mengenai kejadian gempa bumi di Indonesia yang akan dianalisis lebih lanjut. Untuk memuat dataset ini ke dalam lingkungan pemrograman Python, digunakan library Pandas, yang memungkinkan pembacaan dan manipulasi data dalam format tabular dengan efisien.

Berikut adalah langkah-langkah pemuatan data:

1. **Mengimpor Library** Untuk mempermudah proses pemuatan dan analisis data, kita menggunakan library Pandas.

```
import pandas as pd
```

2. **Membaca File CSV** Dataset dimuat ke dalam DataFrame Pandas menggunakan fungsi `pd.read_csv()`. Dengan ini, data dapat diakses dan dimanipulasi dengan lebih mudah.

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

Dengan langkah-langkah di atas, dataset telah berhasil dimuat dan siap untuk dianalisis lebih lanjut.

2.4 Data Understanding

Untuk memahami dataset secara keseluruhan, kita bisa melihat kita dapat melihat Statistik Dasar dari Dataset seperti berikut:

1. Ringkasan Statistik Dasar

```
Ringkasan Statistik Dataset:
                lat           lon         depth           mag      strike1  \
count  92887.000000  92887.000000  92887.000000  92887.000000  2735.000000
mean      -3.404577    119.159707     49.009399      3.592788   170.142852
std        4.354584     10.833202     76.761070      0.834042    88.359267
min      -11.000000     94.020000      2.000000      1.000000     0.000000
25%       -7.885000    113.170000     10.000000      3.000000   107.550000
50%       -2.910000    121.160000     16.000000      3.500000   144.600000
75%        0.140000    126.900000     54.000000      4.200000   217.500000
max        6.000000    142.000000    750.000000      7.900000   359.200000

              dip1        rake1      strike2         dip2        rake2
count  2735.000000  2735.000000  2735.000000  2735.000000  2735.000000
mean     60.202121    30.358062   197.450303    56.576344    35.250018
std      19.699252    99.957906   118.920519    21.274923    98.235894
min       2.300000  -180.000000     0.000000     1.500000  -180.000000
25%      46.950000   -28.500000    63.115000    39.400000   -19.900000
50%      62.300000    57.600000   240.720000    58.400000    56.500000
75%      76.400000   100.150000   297.480000    74.700000   112.600000
max      90.000000   180.000000   359.980000    90.000000   180.000000
```

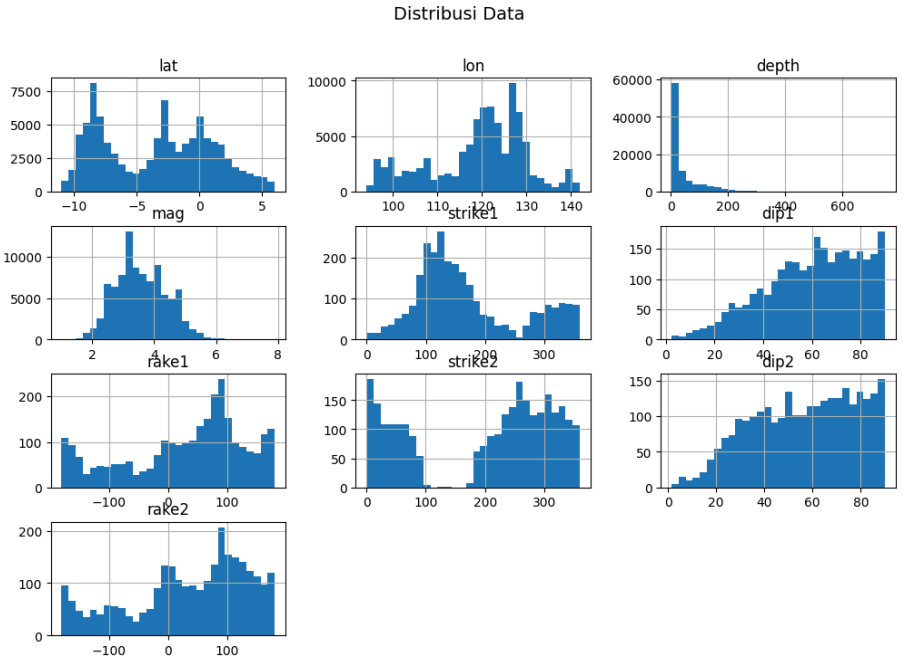
Berdasarkan ringkasan statistik, magnitudo gempa (**mag**) berkisar antara terendah dan tertinggi, dengan rata-rata sekitar nilai mean. Kedalaman gempa (**depth**) sangat bervariasi, dengan beberapa gempa sangat dangkal dan beberapa sangat dalam. Sebaran latitude (**lat**) dan longitude (**lon**) menunjukkan cakupan gempa di berbagai wilayah di Indonesia.

2. Jumlah Missing Values di setiap kolom

```
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
```

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

3. Visualisasi Distribusi data

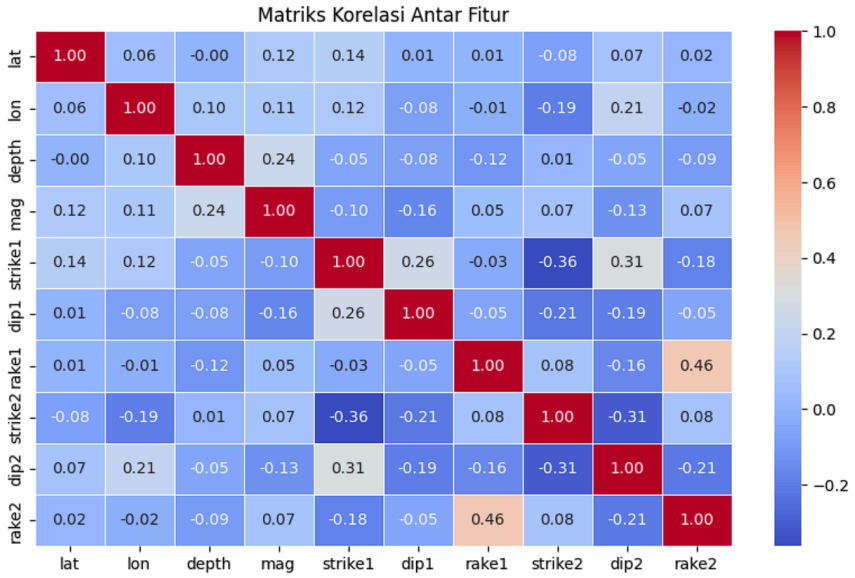


- Magnitudo (**mag**): Distribusi mendekati normal, dengan mayoritas gempa berkisar antara 3 hingga 6.
- Latitude (**lat**) dan **Longitude** (**lon**): Menunjukkan sebaran lokasi gempa di wilayah Indonesia.

- Kedalaman (**depth**): Distribusi miring ke kanan, menunjukkan lebih banyak gempa terjadi di kedalaman dangkal.
- **Strike**, **Dip**, dan **Rake**: Memiliki pola distribusi yang beragam, beberapa menunjukkan variasi yang tinggi

.

4. Korelasi antar fitur



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

5. Insight dari Eksplorasi Data

- Magnitudo: Mayoritas gempa berkisar antara 3–6, sedangkan gempa besar (>7) jarang terjadi.
- Sebaran Lokasi: Gempa tersebar di sepanjang Cincin Api Pasifik, terutama di Sumatra, Jawa, Sulawesi, dan Papua.
- Kedalaman Gempa: Mayoritas kurang dari 100 km, menunjukkan aktivitas tektonik dangkal.
- Mekanisme Patahan: Beragam jenis patahan sesuai dengan pergerakan lempeng di Indonesia.
- Missing Values: Tidak ada nilai yang hilang, dataset siap untuk analisis lebih lanjut.

2.5 Data Preparation

Pada tahap ini, dilakukan preprocessing untuk menyesuaikan dataset agar dapat digunakan secara optimal dalam analisis lebih lanjut. Jika tujuan utama dari dataset ini adalah untuk membuat model machine learning, seperti klasifikasi atau klastering, maka penanganan terhadap data yang kosong atau tidak relevan sangatlah penting. Oleh karena itu, kami melakukan beberapa langkah preprocessing sebagai berikut:

1. **Memisahkan tahun, bulan dan tanggal.**

```
# 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)

df.tail()
```

	ot	lat	lon	depth	mag	remark	strike1	dip1	rake1	strike2	dip2	rake2	tahun	bulan	tanggal
92882	02:25:09.288	3.24	127.18	10	4.0	Talaud Islands - Indonesia	NaN	NaN	NaN	NaN	NaN	NaN	2023	1	26
92883	02:15:03.893	2.70	127.10	10	3.9	Northern Molucca Sea	NaN	NaN	NaN	NaN	NaN	NaN	2023	1	26
92884	01:57:08.885	-7.83	121.07	10	3.8	Flores Sea	NaN	NaN	NaN	NaN	NaN	NaN	2023	1	26
92885	01:46:21.009	3.00	127.16	10	4.1	Northern Molucca Sea	NaN	NaN	NaN	NaN	NaN	NaN	2023	1	26
92886	00:00:35.181	-8.87	118.95	10	2.4	Sumbawa Region - Indonesia	NaN	NaN	NaN	NaN	NaN	NaN	2023	1	26

- Kolom **tgl** (tanggal) dipecah menjadi tiga bagian: **tahun, bulan, dan tanggal** menggunakan metode `.str.split("/", expand=True).`
- Setelah pemisahan, tipe data dikonversi ke integer menggunakan `.astype(int),` sehingga mempermudah analisis berbasis waktu.
- Kolom **tgl** yang asli kemudian dihapus (`df.drop(columns=['tgl'], inplace=True)`) karena informasinya telah dipisahkan ke dalam kolom yang lebih spesifik.

2. **Mengonversi kolom kategorikal remark menjadi numerik dengan Label Encoding.**

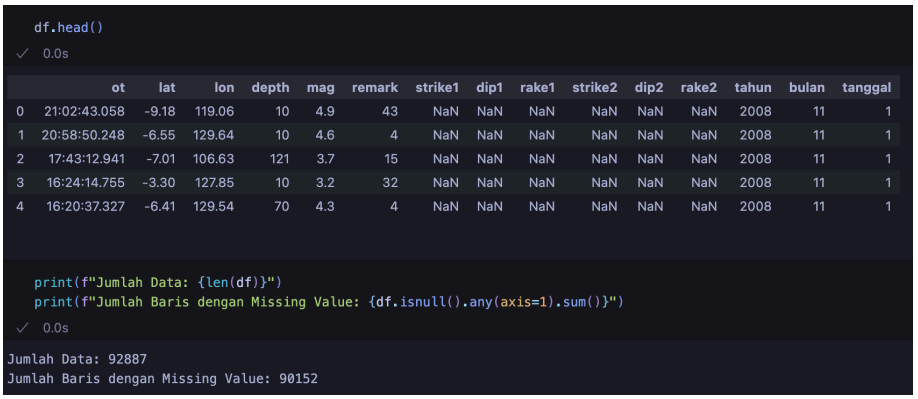
```
# Mengubah kolom kategorikal remark menjadi numerik dengan Label Encoding
from sklearn.preprocessing import LabelEncoder

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

df.tail()
```

✓ 0.0s

	ot	lat	lon	depth	mag	remark	strike1	dip1	rake1	strike2	dip2	rake2	tahun	bulan	tanggal
92882	02:25:09.288	3.24	127.18	10	4.0	46	NaN	NaN	NaN	NaN	NaN	NaN	2023	1	26
92883	02:15:03.893	2.70	127.10	10	3.9	27	NaN	NaN	NaN	NaN	NaN	NaN	2023	1	26
92884	01:57:08.885	-7.83	121.07	10	3.8	11	NaN	NaN	NaN	NaN	NaN	NaN	2023	1	26
92885	01:46:21.009	3.00	127.16	10	4.1	27	NaN	NaN	NaN	NaN	NaN	NaN	2023	1	26
92886	00:00:35.181	-8.87	118.95	10	2.4	44	NaN	NaN	NaN	NaN	NaN	NaN	2023	1	26

- Kolom **remark** yang berisi informasi kategorikal diubah menjadi nilai numerik menggunakan **Label Encoding** dari pustaka `sklearn.preprocessing`.
- **Label Encoding** mengubah kategori dalam kolom tersebut menjadi nilai numerik sehingga dapat digunakan dalam analisis lebih lanjut atau model machine learning.

3. **Membagi dataframe menjadi dataframe lengkap, dataframe umum, dan dataframe khusus.**
   - **Dataframe lengkap (df)**: Dataframe yang digunakan untuk analisis dan visualisasi, sehingga missing value dihiraukan karena dapat berguna untuk menambah informasi dari suatu data.

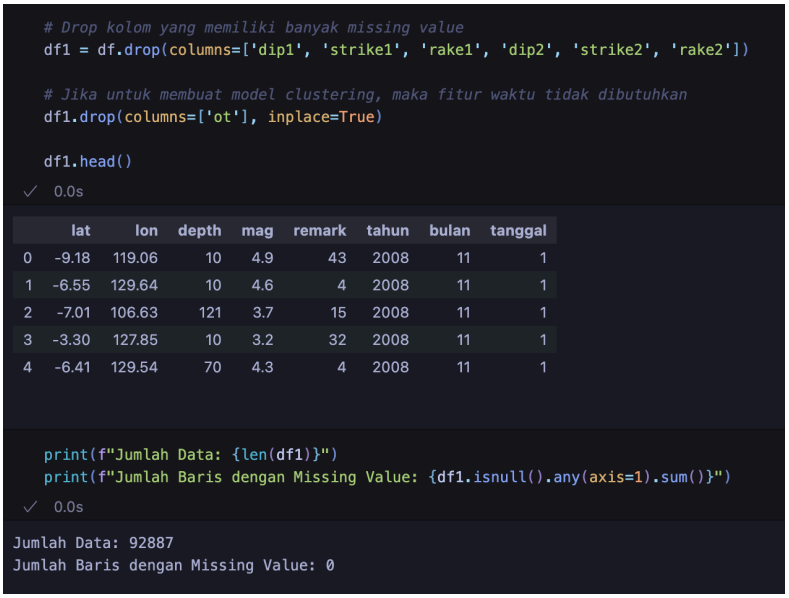


```
df.head()
✓ 0.0s
```

	ot	lat	lon	depth	mag	remark	strike1	dip1	rake1	strike2	dip2	rake2	tahun	bulan	tanggal
0	21:02:43.058	-9.18	119.06	10	4.9	43	NaN	NaN	NaN	NaN	NaN	NaN	2008	11	1
1	20:58:50.248	-6.55	129.64	10	4.6	4	NaN	NaN	NaN	NaN	NaN	NaN	2008	11	1
2	17:43:12.941	-7.01	106.63	121	3.7	15	NaN	NaN	NaN	NaN	NaN	NaN	2008	11	1
3	16:24:14.755	-3.30	127.85	10	3.2	32	NaN	NaN	NaN	NaN	NaN	NaN	2008	11	1
4	16:20:37.327	-6.41	129.54	70	4.3	4	NaN	NaN	NaN	NaN	NaN	NaN	2008	11	1

```
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
```

- **Dataframe umum (df1):** Dataframe yang digunakan untuk machine learning yang memiliki data-data umum saja (tidak memiliki kolom dip1, strike1, rake1, dip2, strike2, and rake2).



```
# Drop kolom yang memiliki banyak missing value
df1 = df.drop(columns=['dip1', 'strike1', 'rake1', 'dip2', 'strike2', 'rake2'])

# Jika untuk membuat model clustering, maka fitur waktu tidak dibutuhkan
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
4	-6.41	129.54	70	4.3	4	2008	11	1

```
print(f"Jumlah Data: {len(df1)}")
print(f"Jumlah Baris dengan Missing Value: {df1.isnull().any(axis=1).sum()}")
✓ 0.0s
Jumlah Data: 92887
Jumlah Baris dengan Missing Value: 0
```

- **Dataframe khusus (df2):** Dataframe yang digunakan untuk machine learning yang memiliki data-data khusus yang tidak terlalu banyak dimiliki oleh data umum biasanya.

```
# Drop baris dengan missing value
df2 = df.dropna().copy()

# 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

```
Jumlah Data: 2735
Jumlah Baris dengan Missing Value: 0
```

Alasan pemisahan:

Tiga dataframe dibuat untuk fleksibilitas analisis dan machine learning. **Df** dipertahankan tanpa menghapus missing value untuk eksplorasi dan visualisasi. **Df1** dibuat dengan menghapus kolom yang memiliki lebih dari 97.7% missing value agar lebih bersih dan ringkas. **Df2** hanya menyimpan data dengan fitur seismik lengkap, cocok untuk model yang membutuhkan informasi spesifik. Pemisahan ini memastikan setiap dataset dapat digunakan sesuai kebutuhan tanpa kehilangan informasi penting.