Pengenalan Datasets

Di subbab ini kita akan mengenal beberapa dataset terbuka yang tersedia dan dapat digunakan secara bebas. Berikut adalah beberapa sumber yang dapat Anda gunakan untuk menemukan dataset:

- **Scikit-Learn**: Scikit-Learn menyediakan sejumlah dataset bawaan yang sangat berguna untuk pemula. Anda dapat mengimpor dataset ini langsung dalam kode Python Anda. Contoh dataset yang dapat Anda gunakan adalah Iris, Wine, dan Breast Cancer.
- **Kaggle**: Kaggle adalah platform online yang menyediakan akses ke berbagai dataset dan kompetisi dalam ilmu data dan kecerdasan buatan. Anda dapat mencari dataset di Kaggle dan mengunduhnya untuk digunakan dalam proyek Anda. Beberapa dataset populer di Kaggle termasuk Titanic, MNIST, dan CIFAR-10.
- **UCI Machine Learning Repository**: Universitas California, Irvine (UCI) menyediakan repository dataset machine learning yang beragam. Mereka memiliki berbagai dataset klasifikasi yang dapat diunduh secara gratis.

Berikut ini adalah beberapa contoh datasets dan cara mengambilnya langsung dari modul Scikit-Learn. Panduan lebih lengkapnya dapat dilihat di halaman Toy Dataset Scikit-Learn

```
1 import numpy as np
2 import pandas as pd
3 from sklearn.datasets import load_iris, load_wine, load_breast_cancer
```

```
1 data_iris = load_iris(as_frame=True)['frame']
2 data_iris
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	2
146	6.3	2.5	5.0	1.9	2
147	6.5	3.0	5.2	2.0	2
148	6.2	3.4	5.4	2.3	2
149	5.9	3.0	5.1	1.8	2

150 rows × 5 columns

```
1 data_wine = load_wine(as_frame=True)['frame']
2 data_wine
```

alcohol	malic acid	ash	alcalinity of ash	magnesium	total phenols	flavanoids	nonflavanoid phenols	proanthocyanins	color inter

```
1 data_breast_cancer = load_breast_cancer(as_frame=True)['frame']
2 data_breast_cancer
```

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	0.09744	...	26.50	98.87	567.7
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	...	26.40	166.10	2027.0
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	...	38.25	155.00	1731.0
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	0.05648	...	34.12	126.70	1124.0
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	0.07016	...	39.42	184.60	1821.0
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	0.05884	...	30.37	59.16	268.6

569 rows × 31 columns

Jika ingin mengambil data custom, misalnya dari file berbentuk `*.csv`, dapat menggunakan fungsi `read_csv` dari Pandas. Lebih lengkapnya dapat dilihat di halaman resmi pandas.read_csv. Berikut ini contohnya:

```
1 data_csv = pd.read_csv("../Datasets/Sample Power Plant Datasets.csv")
2 data_csv
```

	Timestamp	Gross Load	Coal Flow
0	2023-09-01 00:00:00	0.718712	0.620595
1	2023-09-01 00:01:00	0.720401	0.619100
2	2023-09-01 00:03:00	0.715198	0.616807
3	2023-09-01 00:04:00	0.709859	0.611955
4	2023-09-01 00:05:00	0.707900	0.607111
...	...	...	...
41601	2023-09-30 23:56:00	0.701683	0.629420
41602	2023-09-30 23:57:00	0.702291	0.623076
41603	2023-09-30 23:58:00	0.706278	0.631010
41604	2023-09-30 23:59:00	0.712731	0.646545
41605	2023-10-01 00:00:00	0.717867	0.662710

41606 rows × 3 columns

Jika ingin mengambil data dari database, dapat menggunakan library bantuan yaitu `mysql-connector-python` dan fungsi `read_sql` dari Pandas. Lebih lengkapnya dapat dilihat di halaman resmi pandas.read_sql. Berikut ini contohnya:

```
1 con = f"""mysql+mysqlconnector://rfamro@mysql-rfam-public.ebi.ac.uk:4497/Rfam"""
2 data_db = pd.read_sql("SELECT rfam_acc, description, author, created FROM family LIMIT 10", con=con)
3 data_db
```

	rfam_acc	description	author	created
0	RF00001	5S ribosomal RNA	Griffiths-Jones SR, Mifsud W, Gardner PP	2013-10-03 20:41:44
1	RF00002	5.8S ribosomal RNA	Griffiths-Jones SR, Mifsud W	2013-10-03 20:47:00
2	RF00003	U1 spliceosomal RNA	Griffiths-Jones SR, Mifsud W, Moxon SJ, Ontive...	2013-10-03 20:57:11
3	RF00004	U2 spliceosomal RNA	Griffiths-Jones SR, Mifsud W, Gardner PP	2013-10-03 20:58:30
4	RF00005	tRNA	Eddy SR, Griffiths-Jones SR, Mifsud W	2013-10-03 21:00:26

Visualisasi dataset

Visualisasi data membantu dalam mengidentifikasi pola, tren, anomali, dan wawasan yang mungkin sulit ditemukan dalam data mentah. Ada beberapa jenis library yang dapat digunakan untuk visualisasi data.

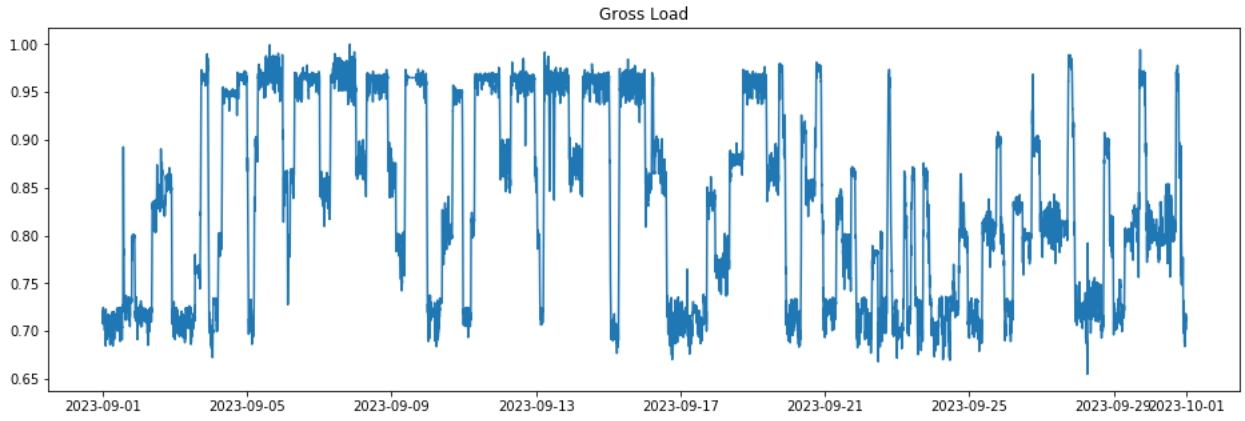
1. Matplotlib

```
import matplotlib.pyplot as plt

x = pd.to_datetime(data_csv['Timestamp'])
y = data_csv['Gross Load']

plt.figure(figsize=(16, 5))
plt.plot(x, y)
plt.title('Gross Load')
plt.show()
```

```
C:\Users\ichsa\AppData\Roaming\Python\Python37\site-packages\matplotlib\cbook\__init__.py:1377: FutureWarning: Support for multi-dimensior
  x[:, None]
C:\Users\ichsa\AppData\Roaming\Python\Python37\site-packages\matplotlib\axes\_base.py:237: FutureWarning: Support for multi-dimensional ir
  x = x[:, np.newaxis]
C:\Users\ichsa\AppData\Roaming\Python\Python37\site-packages\matplotlib\axes\_base.py:239: FutureWarning: Support for multi-dimensional ir
  y = y[:, np.newaxis]
```



2. Plotly

```
import plotly.express as px

fig = px.scatter(data_breast_cancer, x='mean radius', y='mean texture', color='target')
fig.update_layout(height=600)
fig.show()
```

3. Seaborn

```
import seaborn as sns

sns.catplot(
    data=data_iris[['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)']].melt(),
    x="value", y="variable", kind="violin"
```

6)

<seaborn.axisgrid.FacetGrid at 0x20b38ee6988>