🏛 General-Purpose Datasets

- Kaggle Datasets: A vast repository of datasets across domains like finance, healthcare, NLP, and computer vision. Many datasets are accompanied by kernels and community discussions.
- UCI Machine Learning Repository: A classic collection featuring over 600 datasets, including well-known ones like Iris, Wine, and Adult Income.
- OpenML: Offers a wide range of datasets with rich metadata, facilitating automated processing and benchmarking.

🎁 1. Kaggle Datasets (requires API key)

Before running this code:

- Create a Kaggle account.
- Get your `kaggle.json` from: https://www.kaggle.com/account
- Place it in `~/.kaggle/` (or configure as shown)

```
python
CopyEdit
import os
from kaggle.api.kaggle_api_extended import KaggleApi

# Configure Kaggle API
os.environ['KAGGLE_USERNAME'] = 'your_username'
os.environ['KAGGLE_KEY'] = 'your_key'

api = KaggleApi()
api.authenticate()

# Example: Download Titanic dataset
api.dataset_download_files('heptapod/titanic', path='datasets/titanic',
unzip=True)

import pandas as pd

df = pd.read_csv('datasets/titanic/train.csv')
print(df.head())
```

📂 2. UCI Machine Learning Repository

Example: Load **Wine** dataset using `sklearn`

```
python
CopyEdit
from sklearn.datasets import load_wine
import pandas as pd
```

```
wine = load_wine()
df = pd.DataFrame(data=wine.data, columns=wine.feature_names)
df['target'] = wine.target

print(df.head())
```

Or manually load CSV from UCI:

```
python
CopyEdit
url = "https://archive.ics.uci.edu/ml/machine-learning-
databases/iris/iris.data"

df = pd.read_csv(url, header=None)
df.columns = ["sepal_length", "sepal_width", "petal_length", "petal_width",
"class"]

print(df.head())
```

🔄 3. OpenML Datasets

Using the `openml` Python package:

```
python
CopyEdit
import openml

# List all datasets (optional)
# datasets = openml.datasets.list_datasets(output_format='dataframe')
# print(datasets.head())

# Load a specific dataset by ID
# Example: ID 61 = Iris dataset
dataset = openml.datasets.get_dataset(61)
X, y, _, _ = dataset.get_data(target=dataset.default_target_attribute)

print(X.head())
print("Target:", y[:5])
```

To install `openml`:

```
bash
CopyEdit
pip install openml
```