

LAPORAN PRAKTIKUM MINGGU KE-1
PEMROSESAN DATA



Oleh :

Bayu Kurniawan

(3322600019)

Sains Data Terapan

Politeknik Elektronika Negeri Surabaya




2023


A. Import dataset dari Kaggle(google colab)

1. Pilih dataset yang ingin di download

Twitter Financial News

Data Card Code (0) Discussion (0) 25 New Notebook


train_data.csv (2.4 MB)   

Detail Compact Column 2 of 2 columns 

A text

Tweets relating to Finance

2. Buat API token baru dari account pada halaman profil Kaggle.








Bayuk6667
Add occupation
Add organization
Add location
Joined 16 minutes ago · last seen in the past day


Competitions Novice


Home Competitions Datasets Code Discussion Followers Notifications Account Edit Public Profile


3. Buka icon files pada bagian kiri google colab kemudian upload file “kaggle.json”

Files 

 ..

 sample_data

 kaggle.json

4. Jalankan kode-kode berikut.

```
[2] !pip install kaggle

[3] !mkdir ~/.kaggle

[6] !cp kaggle.json ~/.kaggle/

[7] !chmod 600 ~/.kaggle/kaggle.json
```

5. Download dataset Kaggle yang telah dipilih sebelumnya dengan kode: `! kaggle competitions download <nama dataset>` Nama dataset diambil dari nama user/nama dataset yang dapat disalin dari link dataset tersebut.

```
! kaggle datasets download sulphatet/twitter-financial-news

Downloading twitter-financial-news.zip to /content
 0% 0.00/1.08M [00:00<?, ?B/s]
100% 1.08M/1.08M [00:00<00:00, 94.4MB/s]

! unzip twitter-financial-news.zip

Archive:  twitter-financial-news.zip
 inflating: train_data.csv
 inflating: valid_data.csv
```

Setelah itu, bila dataset yang didownload berformat zip , kita bisa unzip file tersebut dengan kode `! unzip <nama file zip>`

B. Tutorial Python(Jupyter)

1. Baca data csv dengan library pandas

```
[21] import pandas as pd
      df1 = pd.read_csv("/content/data.csv")
      print(df1.to_string())
```

	Duration	Pulse	Maxpulse	Calories
0	60	110	130	409.1
1	60	117	145	479.0
2	60	103	135	340.0
3	45	109	175	282.4
4	45	117	148	406.0
5	60	102	127	300.0
6	60	110	136	374.0
7	45	104	134	253.3
8	30	109	133	195.1
9	60	98	124	269.0
10	60	103	147	329.3

```
[22] import pandas as pd
df2 = pd.read_json("/content/data.js")
print(df2.to_string())
```

	Duration	Pulse	Maxpulse	Calories
0	60	110	130	409.1
1	60	117	145	479.0
2	60	103	135	340.0
3	45	109	175	282.4
4	45	117	148	406.0
5	60	102	127	300.5
6	60	110	136	374.0
7	45	104	134	253.3
8	30	109	133	195.1

2. Bersihkan null value atau NaN dengan method dropna()

```
df1.dropna(inplace = True)
print(df1.to_string())
```

	Duration	Pulse	Maxpulse	Calories
0	60	110	130	409.1
1	60	117	145	479.0
2	60	103	135	340.0
3	45	109	175	282.4
4	45	117	148	406.0
5	60	102	127	300.0
6	60	110	136	374.0
7	45	104	134	253.3
8	30	109	133	195.1

```
[25] df2.dropna(inplace = True)
print(df2.to_string())
```

	Duration	Pulse	Maxpulse	Calories
0	60	110	130	409.1
1	60	117	145	479.0
2	60	103	135	340.0
3	45	109	175	282.4
4	45	117	148	406.0
5	60	102	127	300.5
6	60	110	136	374.0
7	45	104	134	253.3
8	30	109	133	195.1
9	60	98	124	269.0
10	60	103	147	329.3

3. Perbaiki data yang salah seperti data 450 pada duration yang salah karena dataset tersebut adalah dataset sesi workout dan seseorang biasanya menghabiskan waktu maksimal 120 menit untuk workout sehingga data duration yang lebih dari 120 bisa diganti dengan 120

```
[26] for x in df1.index:
      if df1.loc[x, "Duration"] > 120:
          df1.loc[x, "Duration"] = 120
      print(df1.to_string())
```

	Duration	Pulse	Maxpulse	Calories
0	60	110	130	409.1
1	60	117	145	479.0
2	60	103	135	340.0
3	45	109	175	282.4
4	45	117	148	406.0
5	60	102	127	300.0
6	60	110	136	374.0
7	45	104	134	253.3
8	30	109	133	195.1

```
for x in df2.index:
    if df2.loc[x, "Duration"] > 120:
        df2.loc[x, "Duration"] = 120
    print (df2.to_string())
```

	Duration	Pulse	Maxpulse	Calories
0	60	110	130	409.1
1	60	117	145	479.0
2	60	103	135	340.0
3	45	109	175	282.4
4	45	117	148	406.0
5	60	102	127	300.5
6	60	110	136	374.0
7	45	104	134	253.3
8	30	109	133	195.1
9	60	98	124	269.0
10	60	103	147	329.3

4. Baca apakah ada baris yang terulang/duplikat, jika ada bisa dibersihkan

```
print(df1.duplicated())
```

```
0      False
1      False
2      False
3      False
4      False
...
164     False
165     False
166     False
167     False
168     False
Length: 164, dtype: bool
```

```
print(df2.duplicated())
```

```
0      False
1      False
2      False
3      False
4      False
...
164     False
165     False
166     False
167     False
168     False
Length: 164, dtype: bool
```