LAPORAN PRAKTIKUM MINGGU KE-1 PEMROSESAN DATA



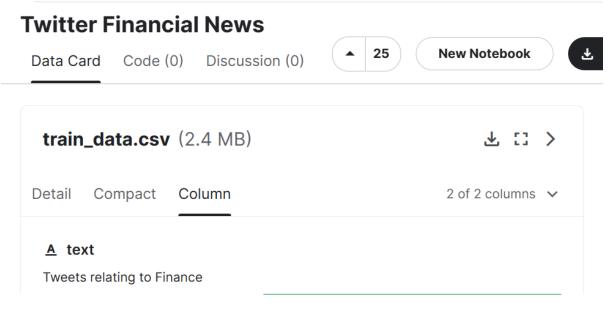
Oleh:

Bayu Kurniawan (3322600019)

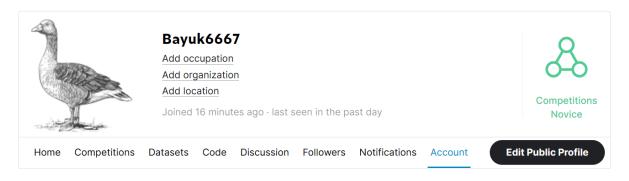
Sains Data Terapan Politeknik Elektronika Negeri Surabaya 2023

A. Import dataset dari Kaggle(google colab)

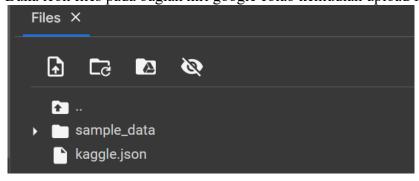
1. Pilih dataset yang ingin di download



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



3. Buka icon files pada bagian kiri google colab kemudian upload file "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
</pre>
```

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

```
[22] import pandas as pd
     df2 = pd.read_json("/content/data.js")
     print(df2.to_string())
                    Pulse Maxpulse Calories
          Duration
     0
                60
                       110
                                 130
                                          409.1
     1
                60
                       117
                                 145
                                          479.0
                                 135
                                          340.0
     2
                60
                       103
     3
                45
                       109
                                 175
                                          282.4
     4
                45
                       117
                                 148
                                          406.0
     5
                60
                       102
                                 127
                                          300.5
     6
                60
                       110
                                 136
                                          374.0
                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
                      117
                                 145
                                         479.0
    1
                60
    2
                                         340.0
                60
                      103
                                 135
    3
                45
                                 175
                                         282.4
                      109
    4
                45
                      117
                                 148
                                         406.0
    5
                60
                      102
                                127
                                         300.0
    6
                60
                      110
                                136
                                         374.0
                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
                                          409.1
                                  130
     1
                 60
                       117
                                  145
                                          479.0
                 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
                 30
     8
                       109
                                  133
                                          195.1
     9
                 60
                        98
                                  124
                                          269.0
     10
                       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
                       110
                                  130
                                          409.1
                 60
                                  145
                                          479.0
                       117
     2
                 60
                       103
                                  135
                                          340.0
     3
                 45
                       109
                                  175
                                          282.4
                 45
                       117
                                  148
                                          406.0
     5
                 60
                       102
                                          300.0
                                  127
     6
                 60
                       110
                                  136
                                          374.0
                 45
                                  134
                       104
                                          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())
C→
          Duration
                    Pulse
                            Maxpulse
                                       Calories
                60
                       110
                                  130
                                          409.1
    0
    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
                                          300.5
                60
                       102
                                  127
    6
                60
                       110
                                  136
                                          374.0
                45
                       104
                                          253.3
                                  134
    8
                30
                       109
                                  133
                                          195.1
    9
                        98
                                          269.0
                60
                                  124
    10
                60
                       103
                                  147
                                           329.3
```

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

```
print(df1.duplicated())
[→ 0
           False
           False
    2
           False
           False
    4
           False
    164
           False
    165
           False
    166
           False
    167
           False
           False
    168
    Length: 164, dtype: bool
```

```
print(df2.duplicated())
[→ 0
           False
           False
    1
           False
           False
           False
    164
          False
    165
          False
    166
           False
    167
           False
    168
           False
    Length: 164, dtype: bool
```