Studi Kasus: Titanic

1. Informasi Dataset

Tahapan-tahapan dalam *preprocessing* dapat berbeda-beda, tergantung kepada permasalahan yang ada. Dataset yang digunakan adalah *dataset* Titanic. Dataset ini berisi sejumlah daftar penumpang kapal Titanic yang menjadi korban dan selamat pada kecelakaan kapal tersebut. Dataset ini digunakan untuk memprediksi apakah seseorang selamat atau tidak berdasarkan data-data lainnya. Adapun informasi fitur pada data ini dapat dilihat pada Tabel di bawah ini.

Variable	**Definisi**	**Penjelasan Isi**
survival	Kolom label yang menyatakan penumpang selamat atau tidak	0 = Tidak selamat, 1 = Selamat
pclass	Kelas tiket penumpang	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Jenis kelamin penumpang	
Age	Umur penumpang dalam tahun	
sibsp	jumlah saudara kandung / pasangan di kapal Titanic	
parch	jumlah orang tua / anak di kapal Titanic	
ticket	Nomer tiket	

fare	Tarif Penumpang	
cabin	Nomor Kabin	
embarked	Pelabuhan keberangkatan	C = Cherbourg, Q = Queenstown, S = Southampton

2. Memuat dan menganalisa tipe data

Dataset ini terdiri atas dua dokumen yaitu train.csv dan test.csv yang dapat diunduh pada https://www.kaggle.com/competitions/titanic/overview. Train.csv berisi data yang akan digunakan untuk pelatihan dan test.csv adalah data yang akan digunakan untuk evaluasi. Untuk memuat data ke dalam notebook kita dapat menggunakan pandas dengan fungsi read_csv dengan parameter nama file. Pada kasus ini langkah tersebut dilakukan pada baris 11 dan 12. Setelah berhasil memuat data, kita tampilkan 5 data pertama dengan menggunakan fungsi head() (Baris 13)

```
1  # data analysis and wrangling
2  import pandas as pd
3  import numpy as np
4  import random as rnd
5
6  # visualization
7  import seaborn as sns
8  import matplotlib.pyplot as plt
9  %matplotlib inline
10
11 train_df = pd.read_csv('train.csv')
12 test_df = pd.read_csv('test.csv')
13 train_df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Berdasarkan 5 data pertama dari fungsi head, kita telah dapat menentukan tipe-tipe data. Adapun tipe data yang ada pada dataset ini adalah

- - Kategori : *Survived, Sex*, dan *Embarked*.
- - Ordinal : Pclass

- - Continous: *Age, Fare*.
- - Discrete: SibSp, Parch.

Sedangkan untuk tipe data Ticketmerupakan gabungan antara numerik dan alpha-numerik dan data Cabin adalah alpha-numerik dan dari analisa awal terdapat missing value (bernilai NaN). Dalam komputasi, NaN adalah singkatan dari *Not a Number*. NaN adalah nilai tipe data numerik yang mewakili nilai yang tidak ditentukan atau tidak terwakili.

```
[2]  1  train_df.info()
     2
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

```
[3]  1  train_df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

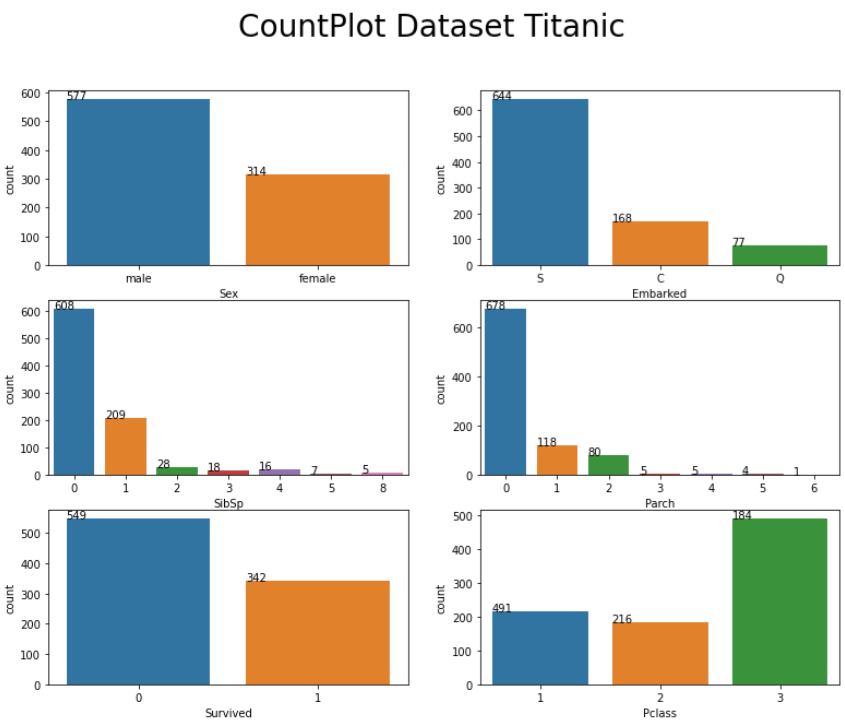
3. Pengecekan Konsistensi Data

Berdasarkan analisa awal, kita memiliki tiga fitur yang bertipe kategori yaitu *Survived, Sex,* dan *Embarked,* dua discrete dan satu ordinal. Ketiga data tersebut akan dilihat sebaran nilainya dan konsistensi nilai (untuk kategori).

```
#pengaturan layout
n_rows=3
n_cols=2
#daftar kolom
kolom=["Sex","Embarked","SibSp","Parch","Survived","Pclass"]

fig, axes = plt.subplots(nrows=n_rows, ncols=n_cols, figsize=(13, 10))
i=0
fig.suptitle('CountPlot Dataset Titanic', fontsize="28")
for row in range(n_rows):
    for col in range(n_cols):
      #menampilkan
      sns.countplot(x=kolom[i], data=train_df, ax=axes[row,col])
      #menampilkan label nilai per unique value
      for p, label in zip(axes[row,col].patches, train_df[kolom[i]].value_counts()):
        axes[row,col].annotate(label, (p.get_x(), p.get_height()))
      i=i+1


```



Pada fitur sex dan embark ini tidak ditemukan inkonsistensi data. Pada fitur Sex terdapat dua nilai unik yakni male dan female dan kolom Embarked memiliki tiga nilai unik yakni S,C dan Q. Contoh inkonsistensi data pada *dataset* adalah apabila misalnya pada sex ditemukan nilai seperti male, M, Female, FEMALE. Data yang bernilai male dan M adalah sama tetapi memiliki value yang berbeda. Solusinya adalah dengan menyatukan data-data yang sama menjadi nilai yang sama.

4. Missing Data, Feature Selection dan Scaling

Pada dataset ini, kolom nama, tiket, Informasi kabin tidak digunakan sehingga kolom-kolom ini dapat dihapus

```
#Hapus Kolom yang tidak di inginkan.
train_df=train_df.drop(["Name","Ticket","Cabin","PassengerId"],axis=1)
train_df.head()
```

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	20.0	1	0	7.0	S
1	1	1	female	36.0	1	0	71.0	C
2	1	3	female	24.0	0	0	7.0	S
3	1	1	female	33.0	1	0	52.0	S
4	0	3	male	33.0	0	0	7.0	S

Selanjutnya, perubahan data kategorikal menjadi data numerik. Kita dapat menggunakan fungsi map pada dataframe. Pada kolom *sex, female* akan diubah menjadi 0 dan male menjadi 1. Begitu juga untuk Embarked S=0, C=1 dan Q=2

Kolom *Age* masih memiliki missing value oleh karena itu kita dapat menggunakan median untuk mengisi umur yang kosong

Agar performa model bagus, kita dapat melakukan scaling terhadap dua nilai yaitu age dan fare

```
#KONVERT STRING VALUES(CATEGORICAL VALUES) MENJDI INTEGER
train_df.Sex=train_df.Sex.map({"female":0,"male":"1"})
train_df.Embarked=train_df.Embarked.map({"S":0,"C":"1","Q":2})
train_df.head()
```

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	1	20.0	1	0	7.0	0
1	1	1	0	36.0	1	0	71.0	1
2	1	3	0	24.0	0	0	7.0	0
3	1	1	0	33.0	1	0	52.0	0
4	0	3	1	33.0	0	0	7.0	0

```
train_df['Age'] = train_df['Age'].fillna((train_df['Age'].median()))
```

```
from sklearn.preprocessing import StandardScaler
train_df["Age"]=round((train_df.Age-train_df.Age.mean()/train_df.Age.std()))
train_df["Fare"]=round((train_df.Fare-train_df.Fare.mean()/train_df.Fare.std()))
```

Studi Kasus: Pengumpulan data dengan metode scrapping berita detik.com

Sebagai contoh kita ingin mengambil data berita dari detik.com. Data yang diambil adalah data berita yang berisi judul, isi berita dan tanggal. Kita tidak mungkin mencatat satu per satu item yang ada di detik.com. Kita dapat membuat sebuah robot yang akan mengumpulkan data yang kita inginkan.

Adapun langkah- langkah yang dilakukan adalah :

1. Install library yang dibutuhkan

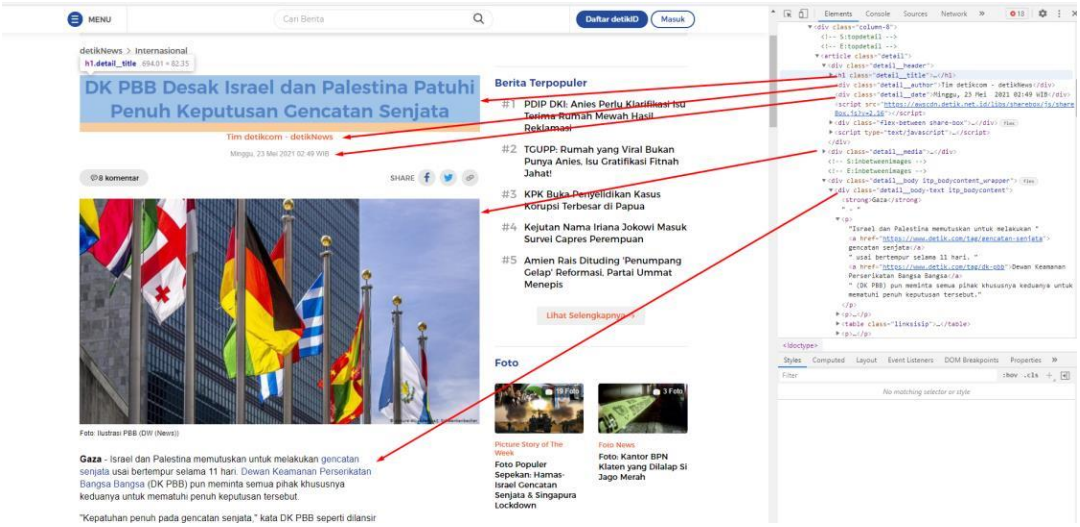
Untuk melakukan *scapping*, dibutuhkan beberapa *library* diantaranya Beautiful Soup dan Request. Untuk melakukan instalasi dapat menggunakan perintah berikut

```
[1]   1   !pip install BeautifulSoup4
      2   !pip install requests
```

Setelah library terinstall semua maka kita sudah suap melakukan scrapping.

2. Mengekstrak isi Halaman Web

Inspeksi bertujuan untuk mengetahui elemen apa yang harus dipanggil untuk melakukan ektraksi konten. Sebagai contoh satu halaman berita detik.com. Dengan bantuan tool inspect dari google chrome maka kita dapat menganlisa struktur halaman berita detik.com. Untuk mengetahui harus melihat source code atau menggunakan tool inspektor pada Gambar di bawah ini.



Hasil inspeksi menunjukkan bahwa:

- Judul berita terletak pada sebuah tag H1 dengan class detail__title
- Penulis terletak pada sebuah tag DIV dengan class detail__author
- Tanggal terletak pada sebuah tag DIV dengan class detail__data
- Isi berita terletak pada DIV dengan class detail__body-text, namun didalamnya ada iklan-iklan ke halaman lain dalam bentuk tabel, oleh karena itu ilkan ini harus dibersihkan

3. Parsing halaman webpage menggunakan Beautiful Soup

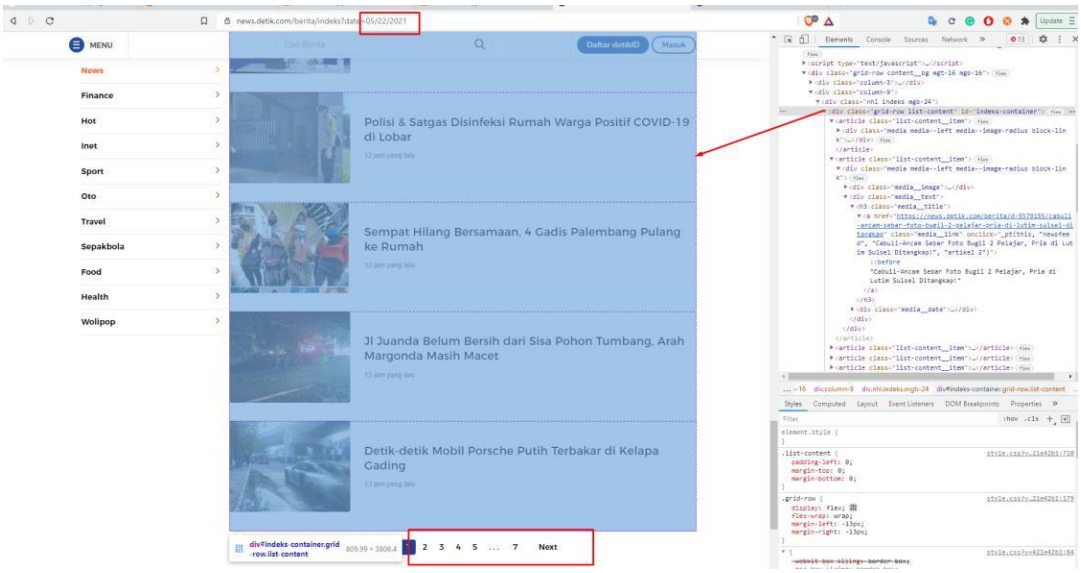
Setelah mengetahui struktur html maka langkah selanjutnya adalah membuat kode yang memparsing halaman web tersebut. Setelah melakukan request pada URL tertentu (baris 9), maka selanjutnya adalah menganalisa hasil request menggunakan Beautiful Soup (baris 10). Untuk mencari Judul maka perlu menemukan element berdasarkan tag H1 dengan class detail__title (baris 12), tanggal dan author pada baris 13 dan 14 . Selanjutnya adalah pembersihan konten dilakukan dengan menghapus semua elemen tabel pada teks.

```
1  import requests
2  from bs4 import BeautifulSoup
3  import pandas as pd
4  import numpy as np
5
6  #ambil berita detik
7  def getBeritaDetik(url):
8      B = { }
9      response = requests.get(url)
10     soup = BeautifulSoup(response.text, 'html.parser')
11     #ambil elemen-elemen berita
12     B['judul'] = soup.find('h1', {'class': 'detail__title'}).text.replace('\n', "").strip()
13     B['tanggal'] = soup.find('div', {'class': 'detail__date'}).text.replace('\n', "").strip()
14     B['author'] = soup.find('div', {'class': 'detail__author'}).text.replace('\n', "").strip()
15     berita = soup.find('div', {'class': 'detail__body-text'})
16     text_berita = berita.text
17     #bersihkan isi berita
18     blah = berita.find_all("table")
19     for x in blah:
20         text_berita = text_berita.replace(x.text,'').replace('\n', "").strip()
21         #print(x.text)
22     B['berita'] = text_berita
23     return(B)
24
```

Method tersebut dapat digunakan untuk mengambil 1 halaman berita contohnya:

4. Mengambil index daftar berita pertanggal

Setelah berhasil mengambil data per berita, maka langkah selanjutnya adalah mengambil seluruh berita. Index berita tersedia pada alamat https://news.detik.com/berita/indek dan dapat dilihat berdasarkan tanggal. Daftar berita pada tanggal tertentu dipisahkan berdasarkan halaman- halaman, karena jumlah berita pada tanggal tersebut cukup banyak. Hasil analisa menunjukkan bahwa kita dapat memberikan parameter date dan nomer halaman. Struktur halaman indek dapat dilihat pada Gambar di bawah ini.



Semua konten terletak pada div dengan kelas list-kontent dan judul dan url terletak pada tag article. Setelah mengambil list url yang ada pada artikel, maka berdasarkan link tersebut kita ambil detail beritanya (baris 21). Adapun source code untuk mengambil list adalah sebagai berikut

```python
import requests
from bs4 import BeautifulSoup

def indexBerita (tanggal,jumlahHalaman):
    daftarBerita = []
    halaman=0
    for halaman in range(0, jumlahHalaman):
        halaman = halaman + 1
        base_url = 'https://news.detik.com/berita/indeks/' + str(halaman)+"?date="+tanggal
        #print(base_url)

        # Request URL and Beautiful Parser
        r = requests.get(base_url)
        soup = BeautifulSoup(r.text, "html.parser")

        berita_container = soup.find('div', {'id': 'indeks-container'})
        berita = berita_container.find_all('article')

        for item in berita:
            x = item.find("a", href=True)
            berita =getBeritaDetik(x['href'])
            daftarBerita.append(berita)
    return daftarBerita

```

Luaran dari method tersebut dapat kita oleh lebih lanjut atau disimpan melalui excel menggunakan pandas

```python
tanggal ='05/22/2021'
jumlahHalaman=1
list_berita =indexBerita(tanggal,jumlahHalaman)
#olah pada dataframe
df = pd.DataFrame(list_berita)
print(df.columns)
print(df)
```

```
Index(['judul', 'tanggal', 'author', 'berita'], dtype='object')
                                        judul  ...                                             berita
0   Menhub: Masyarakat dari Sumatera ke Jawa Wajib...  ...  Jakarta - Menteri Perhubungan (Menhub) Budi Ka...
1   Cabuli-Ancam Sebar Foto Bugil 2 Pelajar, Pria ...  ...  Luwu Timur - Pemuda bernama Adrian (22) di Luw...
2   Mendes Minta Pendamping Desa Gotong-Royong Ban...  ...  Jakarta - Menteri Desa, Pembangunan Daerah Ter...
3         Gempa M 3,6 Terjadi di Waingapu Sumba Timur  ...  Jakarta - Gempa berkekuatan magnitudo (M) 3,6 ...
4   Beri Benih hingga Cold Storage, Mentan Harap P...  ...  Jakarta - Kementerian Pertanian (Kementan) ser...
5   MUI Setuju Saran JK soal Kotak Amal untuk Pale...  ...  Jakarta -  Majelis Ulama Indonesia (MUI) memin...
6   Dalih Ngantuk Pemobil Tabrak Lari Pedagang Mi ...  ...  Jakarta -  Berkat 'kesaktian' teknologi, penge...
7   Mendes Jadikan 2 Desa di Jatim Percontohan Pem...  ...  Jakarta - Menteri Desa, Pembangunan Daerah Ter...
8   KPK Bantah OTT Bupati Nganjuk Ditangani Baresk...  ...  Jakarta - Direktur Sosialisasi dan Kampanye An...
9   Jejak Dokter Asal Sumut Jual Vaksin Corona Ile...  ...  Jakarta -  Polda Sumatera Utara (Sumut) meneta...
10  Cegah Lonjakan COVID-19, Satgas Minta 7 Kota Z...  ...  Jakarta - Satgas Penanganan COVID-19 meminta p...
11  Libur Lebaran, Vaksinasi Lansia di NTB Tetap J...  ...  Jakarta - Vaksinasi lansia tetap dilaksanakan ...
12  Saling Serang Partai Ummat-Ngabalin Buntut Ami...  ...  Jakarta -  Tenaga Ahli Utama Kantor Staf Presi...
13  Tak Diizinkan, Puluhan Pesepeda Gagal Peringat...  ...  Jakarta - Beberapa organisasi kemasyarakatan s...
```

5. Mengubah dataset menjadi Excel

Untuk mengubah menjadi Excel maka cukup menjalankan perintah to_excel pada Pandas Dataframe anda. Contoh penggunaannya adalah berikut

```
df.to_excel("berita.xls")
```

	A	B	C	D	E	F
		judul	tanggal	author	berita	
0		Menhub: Masyarakat	Sabtu, 22 Mei 2021 23:44	Matius Alfons - detikNev	Jakarta - Menteri Perhubungan (Menhub) Budi Karya Sumadi meminta ag	
1		Cabuli-Ancam Sebar	Sabtu, 22 Mei 2021 23:13	Hermawan Mappiwali - d	Luwu Timur - Pemuda bernama Adrian (22) di Luwu Timur (Lutim), Sulawe	
2		Mendes Minta Penda	Sabtu, 22 Mei 2021 23:09	Jihaan Khoirunnisaa - de	Jakarta - Menteri Desa, Pembangunan Daerah Tertinggal dan Transmigras	
3		Gempa M 3,6 Terjadi	Sabtu, 22 Mei 2021 22:54	Tim detikcom - detikNev	Jakarta - Gempa berkekuatan magnitudo (M) 3,6 terjadi di Waingapu, Su	
4		Beri Benih hingga Co	Sabtu, 22 Mei 2021 22:48	Nadhifa Sarah Amalia - d	Jakarta - Kementerian Pertanian (Kementan) serahkan bantuan benih hor	
5		MUI Setuju Saran JK	Sabtu, 22 Mei 2021 22:40	Tim detikcom - detikNev	Jakarta - Majelis Ulama Indonesia (MUI) meminta agar Pemerintah Indon	
6		Dalih Ngantuk Pemol	Sabtu, 22 Mei 2021 22:23	Tim detikcom - detikNev	Jakarta - Berkat 'kesaktian' teknologi, pengemudi yang menabrak lari pe	
7		Mendes Jadikan 2 De	Sabtu, 22 Mei 2021 22:01	Inkana Putri - detikNews	Jakarta - Menteri Desa, Pembangunan Daerah Tertinggal dan Transmigras	
8		KPK Bantah OTT Bu	Sabtu, 22 Mei 2021 21:50	Matius Alfons - detikNev	Jakarta - Direktur Sosialisasi dan Kampanye Anti-Korupsi Komisi Pembe	
9		Jejak Dokter Asal Su	Sabtu, 22 Mei 2021 21:41	Tim detikcom - detikNev	Jakarta - Polda Sumatera Utara (Sumut) menetapkan empat orang, term	
10		Cegah Lonjakan COV	Sabtu, 22 Mei 2021 21:40	Erika Dyah - detikNews	Jakarta - Satgas Penanganan COVID-19 meminta pemerintah kabupaten/	
11		Libur Lebaran, Vaksi	Sabtu, 22 Mei 2021 21:28	Nadhifa Sarah Amalia - d	Jakarta - Vaksinasi lansia tetap dilaksanakan meski saat Idul Fitri. Vaksi	
12		Saling Serang Partai	Sabtu, 22 Mei 2021 21:01	Tim detikcom - detikNev	Jakarta - Tenaga Ahli Utama Kantor Staf Presiden (KSP) Ali Mochtar Nga	
13		Tak Diizinkan, Puluh	Sabtu, 22 Mei 2021 20:59	Adhyasta Dirgantara - d	Jakarta - Beberapa organisasi kemasyarakatan sipil bersama mahasiswa	
14		Kisah Ipda Tita, Polw	Sabtu, 22 Mei 2021 20:30	Yogi Ernes - detikNews	Serpong - Kasus penganiayaan anak yang dilakukan ayahnya sendiri di S	
15		Tekan COVID-19, Sa	Sabtu, 22 Mei 2021 20:30	Inkana Putri - detikNews	Jakarta - Juru Bicara Satgas Penanganan COVID-19 Prof Wiku Adisasmi	
16		Polisi & Satgas Disin	Sabtu, 22 Mei 2021 20:26	Erika Dyah - detikNews	Jakarta - Bhabinkamtibmas Desa Midang Polsek Gunungsari, Aipda Dew	
17		Sempat Hilang Bersa	Sabtu, 22 Mei 2021 20:16	Prima Syahbana - detikI	Palembang - Empat gadis di Palembang, Sumatera Selatan, yang dilapo	
18		Jl Juanda Belum Ber	Sabtu, 22 Mei 2021 20:10	Adhyasta Dirgantara - d	Depok - Sejumlah pohon tumbang di beberapa titik di Jl Ir H Juanda, Dep	
19		Detik-detik Mobil Por	Sabtu, 22 Mei 2021 19:49	Yogi Ernes - detikNews	Jakarta - Sebuah mobil sport merek Porsche hangus terbakar di Kelapa G	

Terdapat code Python seperti berikut:

mylist = [1, 2, 3, ‘apel’, 4]

mylist.append([‘apel’, 4])

print(mylist)

Apakah hasil output yang dihasilkan?

3. Terdapat code Python seperti berikut:

import numpy as np

myarray = np.array([[2, 2, 2], [3, 3, 3]])

print(myarray.T.shape)

Apakah output yang dihasilkan?

4. Terdapat code Python seperti berikut:

myarray = np.array([[1, 2, 3], [4, 5, 6]])

rownames = ['a', 'b']

colnames = ['satu', 'dua', 'tiga']

mydf = pd.DataFrame(myarray, index=rownames, columns=colnames)

print(mydf.satu[‘a’])

Apakah output yang dihasilkan?

5. Terdapat code Python seperti berikut:

def fungsi(a, b, c):

if b > c:

b -= 1

a += 1

fungsi(a, b, c)

return a + b + c

print(fungsi(4, 6, 1))

Apakah output yang dihasilkan?