Modul Praktikum

Kecerdasan Buatan

2

Dataset dan EDA

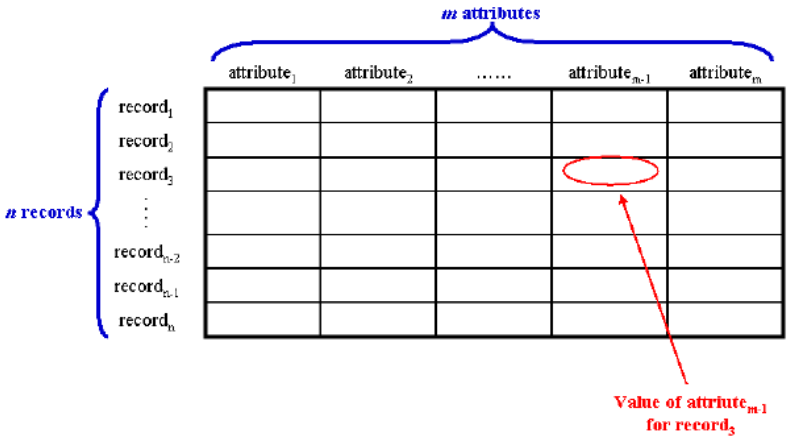
Pengertian Dataset

Dataset diartikan sebagai kumpulan data atau dokumen yang berisi satu atau lebih catatan (record). Setiap kelompok record ini disebut sebagai dataset dan memiliki peran untuk menyimpan informasi seperti catatan medis, asuransi, program, dan sistem data institusi. Dataset dapat berbentuk fakta, angka, nama, atau bahkan deskripsi. Contoh dataset seperti data nilai praktikum kecerdasan buatan mahasiswa Informatika tahun 2022, data sentimen pemilu 2024 pada media sosial, data penjualan beras kota samarinda tahun 2017 – 2022, dan lain-lain. Untuk memahami lebih dalam terkait dengan dataset kami sajikan contoh gambar dari dataset di bawah ini.

	name	custname	Sex	Age	Marital	tanggunan	Occupation	Wilayah	Keterangan Waktu	Target
2	MJK_SUPER_N	YENNI HARDJANTO	P	46	kawin	1	Pegawai Swasta	Barat	Siang	Potensial
3	MJK_SUPER_N	LUKMAN HAKIM	L	47	kawin	3	Pegawai Swasta	Barat	Malam	Potensial
4	MJK_SUPER_N	JAMBARIAL	L	47	kawin	1	Wiraswasta	Barat	Sore	Potensial
5	MJK_SUPER_N	WAGINO	L	46	kawin	2	Polisi	Barat	Sore	Potensial
6	MJK_SUPER_N	WINARTO	L	61	belum kawin	0	Wiraswasta	Barat	Siang	Potensial
7	MJK_SUPER_N	SUHERI	L	59	kawin	2	Pegawai Swasta	Barat	Pagi	Potensial
8	MJK_SUPER_N	ADE IMAM SUHARYANTO	L	57	kawin	1	Wiraswasta	Barat	Sore	Potensial
9	MJK_SUPER_N	ANDRI FERIKA	L	61	kawin	1	Pegawai Swasta	Barat	Pagi	Potensial
10	MJK_SUPER_N	EFFY ZULKIFLIE	P	47	kawin	1	Wiraswasta	Barat	Sore	Potensial
11	MJK_SUPER_N	NURJANNAH	P	57	kawin	3	Pegawai Swasta	Barat	Sore	Potensial
12	MJK_SUPER_N	ENDANG SRI SOESILOWAT	P	44	kawin	3	Pegawai Swasta	Barat	Siang	Tidak Potensial
13	MJK_SUPER_N	HARY AGUS WIBOWO	L	38	belum kawin	0	Pegawai Swasta	Barat	Siang	Tidak Potensial
14	MJK_SUPER_N	SOEPRAHADI	L	40	belum kawin	0	Wiraswasta	Barat	Sore	Tidak Potensial
15	MJK_SUPER_N	HAMONANGAN SIRAIT	L	45	kawin	2	Wiraswasta	Barat	Siang	Tidak Potensial
16	MJK_SUPER_N	CHERYL ASMARADEWI	L	38	kawin	2	Pegawai Swasta	Barat	Sore	Tidak Potensial
17	MJK_SUPER_N	HARUN RASYID	L	40	belum kawin	0	Pendidik	Barat	Siang	Tidak Potensial
18	MJK_SUPER_N	HADI SUWARNO	L	47	belum kawin	0	Pegawai Swasta	Barat	Sore	Potensial
19	MJK_SUPER_N	ZAINAL ARIFIN S	L	47	belum kawin	0	Pegawai Swasta	Barat	Sore	Potensial

Struktur Dataset

Dataset memiliki struktur yang kompleks seperti berikut.



Dengan keterangan sebagai berikut

1. Attribute adalah sebuah atau sekumpulan kolom pada dataset untuk pengelompokan pada semua record dan label.
2. Record adalah sebuah atau sekumpulan baris nilai yang akan membentuk sebuah keputusan atau label.
3. Label adalah keputusan atau penamaan pada setiap nilai record.

Tipe Dataset

Dataset juga memiliki tipe-tipe untuk membedakan antara satu dataset dengan dataset lainnya, dan mempelajari hal ini juga berguna untuk menentukan algoritma Machine Learning/Deep Learning yang tepat sesuai dataset yang digunakan sehingga dapat mencapai akurasi yang optimal. Tipe-tipe dataset diantaranya adalah sebagai berikut :

1. Data *Categorical*

   Data *Categorical* adalah data yang memiliki kategori (data yang lebih dari 2 variabel dependen). pengkategoriannya sesuai terhadap setiap variable independen. Contoh data *categorical* adalah pada gambar dibawah ini,

Obs	Cholesterol	Sex	BP_Status
1	194	Male	Normal
2	200	Female	High
3	233	Male	High
4	192	Female	Optimal
5	209	Female	Normal
6	200	Female	High
7	184	Female	Normal
8	228	Female	High
9	150	Female	Normal
10	221	Male	Normal

2. Data *Numeric*

   Data Numeric adalah data kategori yang di dalamnya terdapat mayoritas data independen terdiri atas angka atau *numeric*. Contoh data *numeric* adalah pada gambar dibawah ini,

Case	Attributes				Decision
	Length	Height	Width	Weight	Quality
1	4.7	1.8	1.7	1.7	high
2	4.5	1.4	1.8	0.9	high
3	4.7	1.8	1.9	1.3	high
4	4.5	1.8	1.7	1.3	medium
5	4.3	1.6	1.9	1.7	medium
6	4.3	1.4	1.7	0.9	low
7	4.5	1.6	1.9	0.9	very-low
8	4.5	1.4	1.8	1.3	very-low

3. Data *Text*

   Data *text* adalah data kategori yang mayoritas datanya independen terdiri atas kumpulan teks yang umumnya adalah *sentiment*, dan lain-lain. Contoh data teks adalah pada gambar di bawah ini,

Loves the German bakeries in Sydney. Together with my imported honey it feels like home	Positive
@VivaLaLauren Mine is broken too! I miss my sidekick	Negative
Finished fixing my twitter...I had to unfollow and follow everyone again	Negative
@DinahLady I too, liked the movie! I want to buy the DVD when it comes out	Positive
@frugaldougal So sad to hear about @OscarTheCat	Negative
@Mofette briliant! May the fourth be with you #starwarsday #starwars	Positive
Good morning thespians a bright and sunny day in UK, Spring at last	Positive
@DowneyisDOWNEY Me neither! My laptop's new, has dvd burning/ripping software but I just can't copy the files somehow!	Negative

4. Data Citra

   Data Citra adalah data yang di mana seluruh isi data nya adalah citra atau gambar, biasanya data citra ini terdiri atas beberapa folder untuk mengategorikan data citra. Contoh dari data citra adalah pada gambar di bawah ini,

5. Data *Time Series*
   Data *Time Series* adalah data dengan setiap elemen data yang ada diukur pada setiap waktu, contoh data *time series* adalah pada gambar di bawah ini

Date	Ozone ($\mu g/m^3$)	Temperature (°C)	Relative humidity (%)	*n* deaths
1 Jan 2002	4.59	−0.2	75.7	199
2 Jan 2002	4.88	0.1	77.5	231
3 Jan 2002	4.71	0.9	81.3	210
4 Jan 2002	4.14	0.5	85.4	203
5 Jan 2002	2.01	4.3	93.5	224
6 Jan 2002	2.4	7.1	96.4	198
7 Jan 2002	4.08	5.2	93.5	180
8 Jan 2002	3.13	3.5	81.5	188
9 Jan 2002	2.05	3.2	88.3	168
10 Jan 2002	5.19	5.3	85.4	194
11 Jan 2002	3.59	3.0	92.6	223
12 Jan 2002	12.87	4.8	94.2	201

Tipe File pada Dataset

Tipe atau jenis file pada dataset terdiri atas beberapa ekstensi yang ada pada Microsoft Excel, Google Spreadsheet (untuk data *categorical* dan data *time series*), ataupun gambar (untuk data citra). Rincian dan penjelasan dari tipe file pada dataset adalah sebagai berikut :

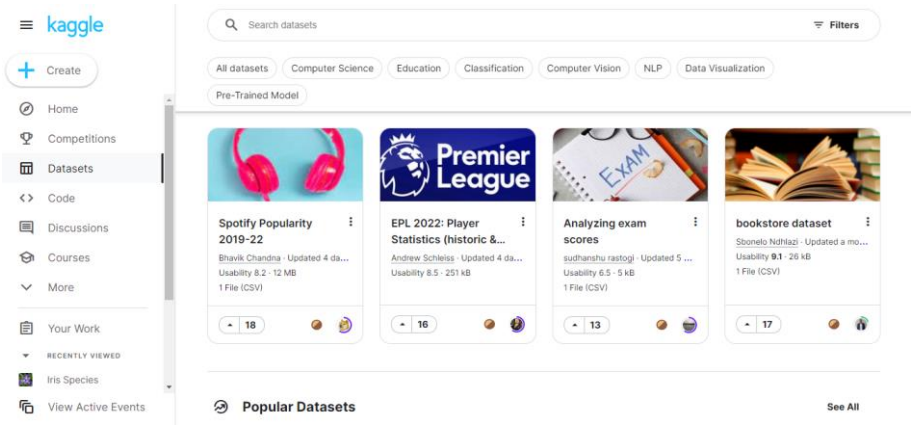
1. CSV (*Comma Separated Value*)
   File *Comma Separated Value* (CSV) adalah file teks yang dipisahkan yang menggunakan koma untuk memisahkan nilai.
2. Xlsx/Xls
   File Xlsx/Xls digunakan untuk menyimpan dan mengelola data seperti angka, rumus, teks, dan bentuk gambar.
3. JPG/JPEG
   JPG/JPEG adalah format gambar terkompresi yang banyak digunakan pada citra. JPG/JPEG ini adalah format citra yang paling umum digunakan dalam kamera digital.
4. PNG
   PNG adalah singkatan dari *Portable Graphics Format*. Ini adalah format citra raster terkompresi (setingkat diatas JPG/JPEG)

Sumber Dataset

Dataset bisa didapatkan dengan cara berbayar ataupun tidak berbayar. Tetapi ada beberapa situs yang dapat dikunjungi, dimana situs tersebut menyediakan dataset yang terbuka yang rinciannya adalah sebagai berikut :

1. Kaggle

   Kaggle adalah platform komunitas online bagi para ilmuwan data dan penggemar machine learning. Kaggle memungkinkan pengguna untuk berkolaborasi dengan pengguna lain, menemukan dan memublikasikan kumpulan data, menggunakan notebook terintegrasi GPU, dan bersaing dengan ilmuwan data lainnya untuk memecahkan tantangan ilmu data.



2. Mnist

   Dataset MNIST adalah singkatan dari Modified National Institute of Standards and Technology dataset. Ini adalah kumpulan data dari 60.000 gambar skala abu-abu 28x28 piksel persegi kecil dari angka tunggal tulisan tangan antara 0 dan 9.


THE MNIST DATABASE

of handwritten digits

Yann LeCun, Courant Institute, NYU
Corinna Cortes, Google Labs, New York
Christopher J.C. Burges, Microsoft Research, Redmond

The MNIST database of handwritten digits, available from this page, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.

It is a good database for people who want to try learning techniques and pattern recognition methods on real-world data while spending minimal efforts on preprocessing and formatting.

Four files are available on this site:

train-images-idx3-ubyte.gz: training set images (9912422 bytes)
train-labels-idx1-ubyte.gz: training set labels (28881 bytes)
t10k-images-idx3-ubyte.gz: test set images (1648877 bytes)
t10k-labels-idx1-ubyte.gz: test set labels (4542 bytes)

please note that your browser may uncompress these files without telling you. If the files you downloaded have a larger size than the above, they have been uncompressed by your browser. Simply rename them to remove the .gz extension. Some people have asked me "my application can't open your image files". These files are not in any standard image format. You have to write your own (very simple) program to read them. The file format is described at the bottom of this page.

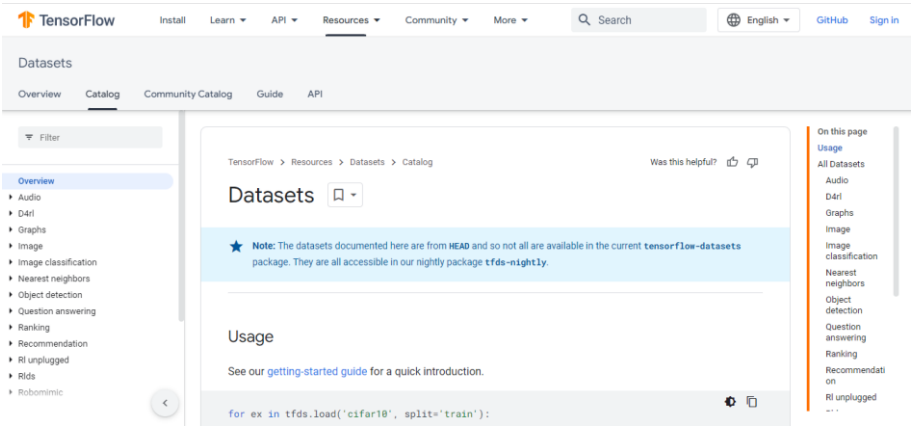
The original black and white (bilevel) images from NIST were size normalized to fit in a 20x20 pixel box while preserving their aspect ratio. The resulting images contain grey levels as a result of the anti-aliasing technique used by the normalization algorithm. the images were centered in a 28x28 image by computing the center of mass of the pixels, and translating the image so as to position this point at the center of the 28x28 field.

With some classification methods (particuarly template-based methods, such as SVM and K-nearest neighbors), the error rate improves when the digits are centered by bounding box rather than center of mass. If you do this kind of pre-processing, you should report it in your publications.

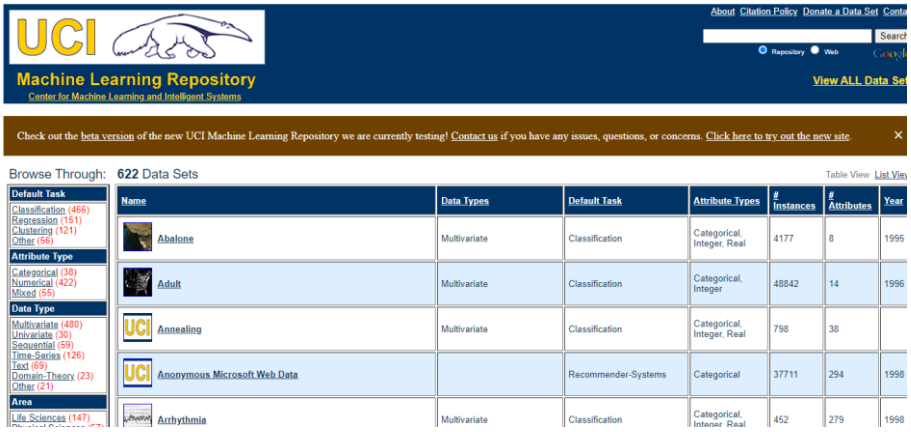
The MNIST database was constructed from NIST's Special Database 3 and Special Database 1 which contain binary images of handwritten digits. NIST originally designated SD-3 as their training set and

3. Tensorflow

   [TensorFlow Datasets] adalah kumpulan dataset yang siap digunakan, dengan TensorFlow atau framework Python ML lainnya, seperti Jax. Semua kumpulan data diekspos dengan syntax tf.data.Datasets.



4. UCI ML



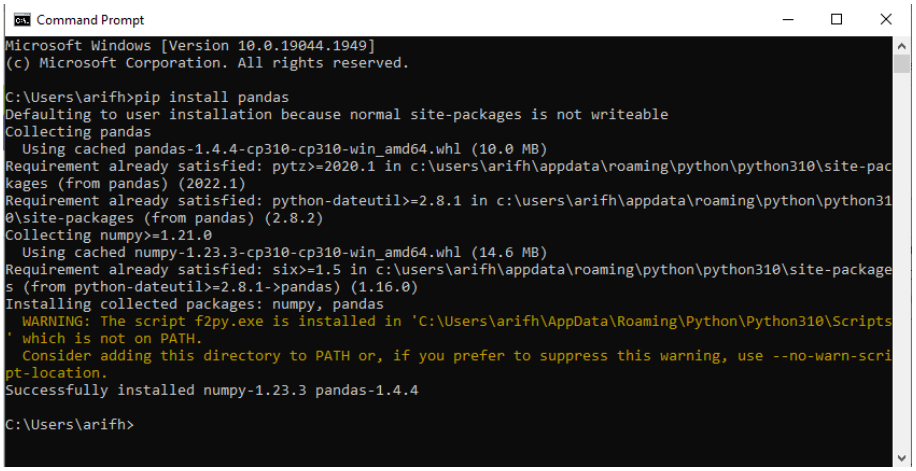
Pandas

Pandas merupakan library python yang open source dan mudah digunakan untuk membuat tabel, mengubah dimensi data , mengecek data dan lainnya. Pandas sering digunakan untuk memanipulasi data, dan membersihkan data mentah ke dalam sebuah bentuk yang bisa untuk diolah dan lainnya.

1. Install Pandas

   Anda dapat menginstall pandas melalui command prompt sebagai berikut

   - Menggunakan Venv >> `pip install pandas`
   - Menggunakan Conda >> `conda install pandas`



2. Jika sudah berhasil melakukan instalasi Pandas, kita dapat menggunakannya untuk melakukan menampilkan data dengan cara import modul tersebut pada proyek yang telah kita buat.

```
import pandas as pd

df = pd.read_csv("House_Rent_Dataset.csv")
df.head(5)
```

Output :

```
[2]: import pandas as pd

     df = pd.read_csv("House_Rent_Dataset.csv")
     df.head(5)
```

[2]:

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
0	2022-05-18	2	10000	1100	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner
1	2022-05-13	2	20000	800	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
2	2022-05-16	2	17000	1000	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
3	2022-07-04	2	10000	800	1 out of 2	Super Area	Dumdum Park	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
4	2022-05-09	2	7500	850	1 out of 2	Carpet Area	South Dum Dum	Kolkata	Unfurnished	Bachelors	1	Contact Owner

3. Menampilkan 5 data terakhir dari dataset

```
df = pd.read_csv("House_Rent_Dataset.csv")
df.tail()
```

Output :

```
[3]: import pandas as pd

     df = pd.read_csv("House_Rent_Dataset.csv")
     df.tail()
```

[3]:

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
4741	2022-05-18	2	15000	1000	3 out of 5	Carpet Area	Bandam Kommu	Hyderabad	Semi-Furnished	Bachelors/Family	2	Contact Owner
4742	2022-05-15	3	29000	2000	1 out of 4	Super Area	Manikonda, Hyderabad	Hyderabad	Semi-Furnished	Bachelors/Family	3	Contact Owner
4743	2022-07-10	3	35000	1750	3 out of 5	Carpet Area	Himayath Nagar, NH 7	Hyderabad	Semi-Furnished	Bachelors/Family	3	Contact Agent
4744	2022-07-06	3	45000	1500	23 out of 34	Carpet Area	Gachibowli	Hyderabad	Semi-Furnished	Family	2	Contact Agent
4745	2022-05-04	2	15000	1000	4 out of 5	Carpet Area	Suchitra Circle	Hyderabad	Unfurnished	Bachelors	2	Contact Owner

4. Untuk menampilkan Info dari dataset

```
df = pd.read_csv("House_Rent_Dataset.csv")
df.info()
```

Output :

```
[4]: import pandas as pd

     df = pd.read_csv("House_Rent_Dataset.csv")
     df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4746 entries, 0 to 4745
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Posted On         4746 non-null   object
 1   BHK               4746 non-null   int64
 2   Rent              4746 non-null   int64
 3   Size              4746 non-null   int64
 4   Floor             4746 non-null   object
 5   Area Type         4746 non-null   object
 6   Area Locality     4746 non-null   object
 7   City              4746 non-null   object
 8   Furnishing Status 4746 non-null   object
 9   Tenant Preferred  4746 non-null   object
 10  Bathroom          4746 non-null   int64
 11  Point of Contact  4746 non-null   object
dtypes: int64(4), object(8)
memory usage: 445.1+ KB
```

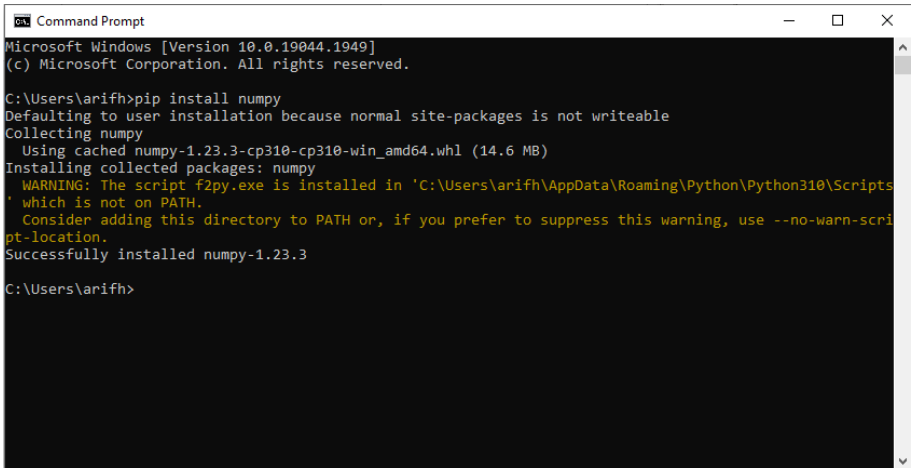
NumPy

NumPy merupakan salah satu library Python yang berfungsi untuk proses komputasi numerik. NumPy memiliki kemampuan untuk membuat objek berdimensi array. Array merupakan sekumpulan variabel yang memiliki tipe data yang sama. Kelebihan dari NumPy adalah dapat memudahkan operasi komputasi pada data, cocok untuk melakukan akses secara acak, dan elemen array merupakan sebuah nilai yang independen sehingga penyimpanannya dianggap sangat efisien.

1. Install NumPy

   Sama dengan Pandas, anda dapat menginstall NumPy melalui command prompt sebagai berikut :

   - Menggunakan Venv >> `pip install numpy`
   - Menggunakan Conda >> `conda install numpy`



2. Import Numpy dan membuat array dengan NumPy

```
import numpy as np

arr = np.array([1, 2, 3, 4])
arr
```

Output :

3. Indexing dan slicing array

   Menunjuk array pada indeks ke-1

```
import numpy as np

arr = np.array([1, 2, 3, 4, 5, 6, 7, 8])
arr[1]
```

Output :

Menunjuk array pada indeks ke-2 hingga ke-5

```
import numpy as np

arr = np.array([1, 2, 3, 4, 5, 6, 7, 8])
arr[2:5]
```

Output :

Menunjuk array pada indeks ke-4 terakhir

```
import numpy as np

arr = np.array([1, 2, 3, 4, 5, 6, 7, 8])
arr[4:]
```

Output :

Step pada array

```
import numpy as np

arr = np.array([1, 2, 3, 4, 5, 6, 7, 8])
arr[0:8:2]
```

Output :

4. Tipe data NumPy

   Tipe data NumPy ada beberapa jenis yang dapat di

   - Strings - digunakan untuk data teks yang ciri-cirinya diberikan di bawah tanda kutip. misalnya "ABD"
   - Integer - digunakan untuk bilangan bulat. misalnya -1, -2, -3
   - Float - digunakan untuk bilangan asli. misalnya 1.2, 42.42
   - Boolean - digunakan untuk True dengan False.
   - Complex - digunakan untuk mewakili bilangan kompleks. misalnya 1.0 + 2.0j, 1.5 + 2.5j

Menampilkan Tipe data NumPy

```
import numpy as np

arr = np.array([1, 2, 3, 4, 5, 6, 7, 8])
arr.dtype
```

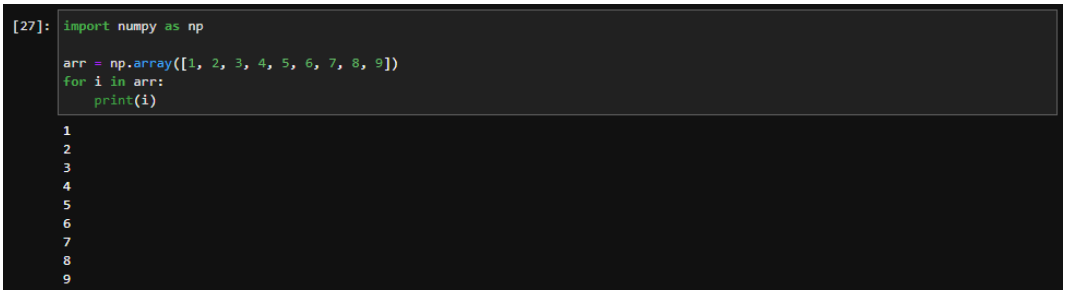
Output :

5. Iterasi

```
import numpy as np

arr = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9])
for i in arr:
    print(i)
```

Output :



6. Pengurutan

```
import numpy as np

arr = np.array([3, 2, 9, 5, 4, 7, 6, 8, 1])
np.sort(arr)
```

Output :

7. Pencarian

Pencarian dengan NumPy dengan menggunakan np.where(kondisi)

```
import numpy as np

arr = np.array([3, 2, 9, 5, 4, 7, 6, 8, 1])
np.where(arr==5)
```

Output :


```
import numpy as np

arr = np.array([3, 2, 9, 5, 4, 7, 6, 8, 1])
np.where(arr%2==0)
```

Output :

Pengertian EDA

Bayangkan ketika anda memutuskan untuk menonton film yang belum pernah Anda dengar. Anda pasti akan mengarah pada sebuah kondisi dimana anda ingin mencari tahu dan bagaimana alur dari film itu apakah dia menarik atau tidak menarik untuk ditonton .Lalu, anda akan melihat beberapa spoiler baik dari media Instagram, youtube, facebook, dan lain-lain sebagai penilaian kuat terhadap sebuah film, dan anda akan mengetahui alur, peringkat, dan ulasan film telah dibuat oleh penonton lain. Apa pun tindakan investigasi yang anda ambil sebelum akhirnya menonton atau tidak menonton film tersebut, tidak lain adalah apa yang oleh para data scientist dalam istilah mereka disebut *Exploratory Data Analysis* (EDA).

Exploratory Data Analysis (EDA) mengacu pada proses kritisasi dalam melakukan penyelidikan awal pada data untuk menemukan pola, anomali, menguji hipotesis dan untuk memeriksa asumsi dengan bantuan statistik ringkasan dan representasi grafis.

Praktik dalam Exploratory Data Analysis (EDA)

Pada praktik dalam EDA, kita akan menggunakan dataset yang akan dibagikan oleh asisten laboratorium di setiap sesi praktikum pada modul ini. Tetapi, sebagai contohnya kita akan menggunakan house rent dataset pada Kaggle dengan link https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset. Untuk menampilkan dataset dengan Pandas Python dapat dilakukan dengan sintaks sebagai berikut :

```
import pandas as pd


dataFrame = pd.read_csv('House_Rent_Dataset.csv')

dataFrame.head()
```

Output :

```
[5]: import pandas as pd

     dataFrame = pd.read_csv('House_Rent_Dataset.csv')
     dataFrame.head()
```

[5]:

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
0	2022-05-18	2	10000	1100	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner
1	2022-05-13	2	20000	800	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
2	2022-05-16	2	17000	1000	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
3	2022-07-04	2	10000	800	1 out of 2	Super Area	Dumdum Park	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
4	2022-05-09	2	7500	850	1 out of 2	Carpet Area	South Dum Dum	Kolkata	Unfurnished	Bachelors	1	Contact Owner

Untuk mengetahui banyak baris dan kolom kita dapat menggunakan

```
dataFrame.shape
```

Output :

```
[8]: dataFrame.shape

[8]: (4746, 12)
```

Untuk menemukan apakah sebuah atau sekelompok kolom tersebut berisi nilai null atau tidak, kita dapat menggunakan

```
dataFrame.info()
```

Output :

```
[13]: dataFrame.info()
      <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 4746 entries, 0 to 4745
      Data columns (total 12 columns):
       #   Column             Non-Null Count  Dtype
      ---  ------             --------------  -----
       0   Posted On          4746 non-null   object
       1   BHK                4746 non-null   int64
       2   Rent               4746 non-null   int64
       3   Size               4746 non-null   int64
       4   Floor              4746 non-null   object
       5   Area Type          4746 non-null   object
       6   Area Locality      4746 non-null   object
       7   City               4746 non-null   object
       8   Furnishing Status  4746 non-null   object
       9   Tenant Preferred   4746 non-null   object
       10  Bathroom           4746 non-null   int64
       11  Point of Contact   4746 non-null   object
      dtypes: int64(4), object(8)
      memory usage: 445.1+ KB
```

Dalam Pandas, untuk mendapatkan berbagai statistik ringkasan seperti mengembalikan hitungan, rata-rata, simpangan baku, nilai minimum dan maksimum, serta kuantil data kita dapat menggunakan

```
DataFrame.describe()
```

Output :

```
[14]: dataFrame.describe()
```

[14]:

	BHK	Rent	Size	Bathroom
count	4746.000000	4.746000e+03	4746.000000	4746.000000
mean	2.083860	3.499345e+04	967.490729	1.965866
std	0.832256	7.810641e+04	634.202328	0.884532
min	1.000000	1.200000e+03	10.000000	1.000000
25%	2.000000	1.000000e+04	550.000000	1.000000
50%	2.000000	1.600000e+04	850.000000	2.000000
75%	3.000000	3.300000e+04	1200.000000	2.000000
max	6.000000	3.500000e+06	8000.000000	10.000000

Untuk mengetahui jumlah data pada label semisal “Rent”, kita dapat menggunakan

```
dataFrame["Rent"].value_counts()
```

Output :

```
[18]: dataFrame["Rent"].value_counts()

[18]: 15000    275
      10000    248
      12000    238
      20000    175
      8000     162
              ...
      4600       1
      79500      1
      76000      1
      45002      1
      5800       1
      Name: Rent, Length: 243, dtype: int64
```