MODUL PRAKTIKUM

MATA KULIAH DATA MINING

PERTEMUAN 02

SEMESTER GENAP

TAHUN AJARAN 2024 -2025

Disusun oleh:

Dwi Welly Sukma Nirad S.Kom, M.T

Aina Hubby Aziira M.Eng

Nurul Afani

Rizka Kurnia Illahi

DEPARTEMEN SISTEM INFORMASI

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS ANDALAS

TAHUN 2025

IDENTITAS PRAKTIKUM

IDENTITAS MATA KULIAH

Kode mata kuliah	JSI62122
Nama mata kuliah	Data Mining
CPMK yang dibebankan pada praktikum	CPMK-01 Mahasiswa mampu menjelaskan definisi data, tahap pra-pengolahan data, serta teknik-teknik merepresentasikan data (CP-1).
Materi Praktikum Pertemuan 02	Pengenalan library: pandas, numpy, scikit-learn
	Membaca dan menulis file (txt, csv, excel)
	Manipulasi data menggunakan pandas
	Operasi matematika dengan numpy

IDENTITAS DOSEN DAN ASISTEN MAHASISWA

Nama Dosen Pengampu	1. Dwi Welly Sukma Nirad S.Kom, M.T 2. Aina Hubby Aziira M.Eng
Nama Asisten Mahasiswa (Kelas A)	1. 2211523034 - Muhammad Fariz 2. 2211521012 - Rizka Kurnia Illahi 3. 2211521010 - Dhiya Gustita Aqila 4. 2211522013 - Benni Putra Chaniago 5. 2211521017 - Ghina Anfasha Nurhadi 6. 2211523022 - Daffa Agustian Saadi 7. 2211521007 - Annisa Nurul Hakim 8. 2211522021 - Rifqi Asverian Putra 9. 2211521009 - Miftahul Khaira 10. 2211521015- Nurul Afani 11. 2211523028 - M.Faiz Al-Dzikro

Nama Asisten Mahasiswa (Kelas B)	1. 2211523034 - Muhammad Fariz 2. 2211521012 - Rizka Kurnia Illahi 3. 2211521010 - Dhiya Gustita Aqila 4. 2211522013 - Benni Putra Chaniago 5. 2211521017 - Ghina Anfasha Nurhadi 6. 2211523022 - Daffa Agustian Saadi 7. 2211521007 - Annisa Nurul Hakim 8. 2211522021 - Rifqi Asverian Putra 9. 2211521009 - Miftahul Khaira 10. 2211521015- Nurul Afani 11. 2211523028 - M.Faiz Al-Dzikro

DAFTAR ISI

OPERASI FILE DAN IMPORT LIBRARY

A. PENGENALAN LIBRARY

Jupyter notebook memungkinkan adanya penggunaan package / library pada inputnya. Library ini berguna mempermudah dalam menggunakan fungsi yang diinginkan. Ada beberapa library atau biasa disebut modul yang sering digunakan yaitu :

1. Pandas

Pandas atau python data analysis library adalah library yang digunakan untuk bisa membaca dan mengakses file / data yang dimasukkan (baik itu file txt, csv, tsv sebagainya).

Install Pandas :

```
pip install pandas
```

Rumus umum memanggil library pandas adalah : **import pandas**

Contoh :

```
import pandas as pd
pd.read_csv('data.csv')
```

Maka akan muncul :

[2]:

	1	2	3	4	5
0	6	7	8	9	10
1	11	12	13	14	15
2	16	17	18	19	20
3	21	22	23	24	25
4	26	27	28	29	30
5	31	32	33	34	35
6	36	37	38	39	40
7	41	42	43	44	45
8	46	47	48	49	50
9	51	52	53	54	55
10	56	57	58	59	60
11	61	62	63	64	65
12	66	67	68	69	70
13	71	72	73	74	75

Bisa juga dideklarasikan sebagai berikut :

```
import pandas
pandas.read_csv('data.csv')
```

[3]:

	1	2	3	4	5
0	6	7	8	9	10
1	11	12	13	14	15
2	16	17	18	19	20
3	21	22	23	24	25
4	26	27	28	29	30
5	31	32	33	34	35
6	36	37	38	39	40
7	41	42	43	44	45
8	46	47	48	49	50
9	51	52	53	54	55
10	56	57	58	59	60
11	61	62	63	64	65
12	66	67	68	69	70
13	71	72	73	74	75

2. Numpy

Numpy atau numerical python adalah library yang digunakan untuk membuat operasi vektor – matriks. Dengan menggunakan numpy kita bisa melakukan agregat pada program, analisis data dan lain sebagainya.

install numpy :

```
pip install numpy
```

Rumus umum memanggil library numpy adalah : **import numpy**

contoh :

```
import numpy as np
a = np.array([[1,2,3],[4,5,6]])
b = np.array([[3,4,5],[6,7,8]])
c = a*b
c
```

maka akan muncul :

```
[2]:

array([[ 3,  8, 15],
       [24, 35, 48]])
```

Jika library numpy tidak diimport seperti ini:

```
#import numpy as np
a = [[1,2,3],[4,5,6]]
b = [[3,4,5],[6,7,8]]
c = a+b
c
```

maka akan muncul :

```
[7]:
[[1, 2, 3], [4, 5, 6], [3, 4, 5], [6, 7, 8]]
```

Jadi ketika operasi matematika dijalankan pada array – matriks, operasi matematika tidak akan jalan sesuai dengan keinginan. Jika c dideklarasikan sebagai a*b maka hasilnya akan jadi error.

3. Scikit Learn

Scikit-learn adalah pustaka Python yang dirancang untuk memfasilitasi proses data mining dengan menyediakan berbagai algoritma dan alat untuk analisis data, termasuk klasifikasi, regresi, clustering, dan pengolahan data. Fitur Utama Scikit-Learn:

1. Klasifikasi: Mengelompokkan data ke dalam kategori yang telah ditentukan. Contoh algoritma: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees.
2. Regresi: Memprediksi nilai kontinu berdasarkan data input. Contoh algoritma: Linear Regression, Ridge Regression.
3. Clustering: Mengelompokkan data ke dalam grup berdasarkan kesamaan. Contoh algoritma: K-Means, DBSCAN.
4. Pengolahan Data: Menyediakan alat untuk preprocessing data, seperti normalisasi, pengisian nilai yang hilang, dan encoding variabel kategorikal.
5. Model Selection dan Evaluasi: Menyediakan alat untuk memilih model yang tepat dan melakukan validasi silang, yang membantu dalam mengevaluasi kinerja model.
6. Pipeline: Memudahkan pengelolaan alur kerja machine learning dengan menggabungkan preprocessing dan model dalam satu objek.

Install scikit-learn

```
pip install scikit-learn
```

Rumus umum memanggil library pandas adalah : **import sklearn**

B. MEMBACA DAN MENULIS FILE

Dalam bahasa pemrograman python, sebuah file dapat dimanipulasi untuk berbagai kebutuhan. Kebutuhan tersebut tidak terlepas dari keinginan untuk memperoleh informasi di dalamnya. Ada beberapa hal yang bisa dilakukan dalam operasi file pada python diantaranya membuka, membaca, menulis, dan menambah file. Hal yang dilakukan tersebut merupakan teknik dasar untuk pemrosesan file.

Ada 2 tipe file yang dikelompokkan pada bahasa python :

1. File teks : file ini berisikan teks; contoh : file txt, csv, tsv, json, md, dll
2. File binary : file ini berisikan non-teks yang hanya diproses oleh program tertentu; contoh : jpg, jpeg, exe, mkv, m4a, dll.

a. Membaca file

Dalam membaca file python, fungsi yang digunakan adalah open(... ,"r") dalam membaca dan menulis file. Contoh :

- **`baca_file = open("teks.txt","r")`** *//berfungsi ketika file teks.txt berada pada direktori yang sama dengan script python.ipynb.*
- **`baca_file = open("C:teks.txt","r" )`** *//berfungsi ketika file teks.txt tidak berada pada direktori yang sama dengan script python.ipynb.*
- untuk perintah menampilkan file tersebut dengan cara : print(baca_file.read())

Ada beberapa mode dalam python untuk menentukan hak akses atas file :

1. r : untuk membaca file (read)
2. w : untuk menulis file (write)
3. a : untuk menambah data file (append)
4. r+ : untuk membaca sekaligus menulis data file

Dalam membaca file binary, tinggal menambahkan kode b setelah mode python tersebut : rb, wb, ab, atau r+b.

b. Menulis File

Ada dua metode yang bisa kita gunakan untuk menulis file :

1. `write()`: parameternya teks (string)
2. `writelines()`: parameternya teks dalam bentuk list.

Method `write()` akan menulis semua teks, sedangkan `writelines()` akan menulis per baris.

C. MANIPULASI DATA DENGAN PANDAS

Berikut beberapa cara manipulasi data dalam DataFrame dengan Pandas, diantaranya yaitu:

1. Membaca data

   Pandas menyediakan fungsi untuk membaca data dari berbagai sumber, seperti `read_csv()` untuk file CSV, `read_excel()` untuk file Excel, dan `read_sql()` untuk database SQL.

2. Mengakses data tertentu pada data frame

   Pandas menyediakan fungsi untuk mengakses kolom atau baris tertentu pada data frame. Metode `.loc` dipakai untuk mengakses baris berdasarkan label indeks, sementara `.iloc` digunakan untuk mengakses baris berdasarkan posisi indeks.

3. Mengelompokkan data

   Pengelompokan digunakan untuk mengelompokkan data menggunakan beberapa kriteria dari kumpulan data kita. Pandas menyediakan fungsi seperti `groupby()` untuk mengelompokkan data.

4. Menggabungkan DataFrame

   Pandas menyediakan fungsi seperti `merge()` dan `concat()` untuk menggabungkan beberapa DataFrame.

D. OPERASI MATEMATIKA DENGAN NUMPY

Terdapat beberapa jenis operator pada Python diantaranya yaitu:

1. Operator Matematika dan String

a. + (tambah)

- Menambahkan dua objek.
- 3 + 5 menghasilkan 8
- 'a' + 'b' menghasilkan 'ab'.

b. - (kurang)
   - Mengurangkan operand kedua dari operand pertama. Jika hanya satu operand,
   - diasumsikan nilai operand pertama adalah 0.
   - 50 - 24 menghasilkan 26.
   - Tidak berlaku untuk string, akan menghasilkan error unsupported operand.

c. * (perkalian)
   - Mengembalikan hasil perkalian angka atau mengembalikan string yang diulang
   - sejumlah tertentu.
   - 2 * 3 menghasilkan 6.

d. ** (pangkat)
   - Mengembalikan operand pertama pangkat operand kedua.
   - 3 ** 4 menghasilkan 81 (sama dengan 3 * 3 * 3 * 3).

e. / (pembagian)
   - Mengembalikan hasil pembagian operand pertama dengan operand kedua (float).
   - 13 / 3 menghasilkan 4.333333333333333.

f. // (pembagian habis dibagi / div)
   - Mengembalikan hasil pembagian operand pertama dengan operand kedua (bilangan
   - bulat), kecuali jika salah satu operand adalah float, akan menghasilkan float.
   - 13 // 3 menghasilkan 4.

g. • % (modulo)
   - Mengembalikan sisa bagi.
   - 13 % 3 menghasilkan 1.

2. Operator Perbandingan

a. < atau operator.lt (less than)
   - Menjalankan perbandingan apakah operand pertama lebih kecil dari operand kedua.
   - 5 < 3 menghasilkan False and 3 < 5 menghasilkan True.
   - Perbandingan dapat berisi lebih dari dua operand, misalnya 3 < 5 < 7 menghasilkan True.

b. > atau operator.gt (greater than)
   - Menjalankan perbandingan apakah operand pertama lebih besar dari operand kedua.
   - 5 > 3 menghasilkan True.

c. <= atau operator.le (less than or equal to)
   - Menjalankan perbandingan apakah operand pertama lebih kecil atau sama dengan operand kedua.
   - x = 3; y = 6;
     - maka x <= y menghasilkan True.

d. >= atau operator.ge (greater than or equal to)
   - Menjalankan perbandingan apakah operand pertama lebih besar atau sama dengan operand kedua.
   - x = 4; y = 3;
     - maka x >= y menghasilkan True.

e. == atau operator.eq (equal to)
   - Menjalankan perbandingan apakah operand pertama sama dengan operand kedua.
   - x = 2; y = 2;
     - maka x == y menghasilkan True.
   - x = 'str'; y = 'stR';
     - maka x == y menghasilkan False.
   - x = 'str'; y = 'str';
     - maka x == y menghasilkan True.

f. != atau operator.ne (not equal to)

- Menjalankan perbandingan apakah operand pertama tidak sama dengan operand kedua.
- x = 2; y = 3;
    - maka x != y returns True.

3. Operator Logika (Boolean)

a. not (boolean NOT)
- Jika x bernilai True, fungsi akan mengembalikan nilai False.
- Jika x bernilai False, fungsi akan mengembalikan nilai True.
- x = True;
    - not x akan mengembalikan nilai False.

b. and (boolean AND)
- x AND y akan mengembalikan nilai False jika x bernilai False, atau fungsi akan mengembalikan nilai y.
- x = False; y = True;
    - x AND y, Fungsi akan mengembalikan nilai False karena x bernilai False.
    - Dalam kasus ini, Python tidak akan mengevaluasi nilai y karena apapun nilai y tidak akan mempengaruhi hasil. Hal ini dinamakan short-circuit evaluation.

c. or (boolean OR)
- x OR y, Jika x bernilai True, fungsi akan mengembalikan nilai True, atau fungsi akan mengembalikan nilai dari y.
- x = True; y = False;
    - x OR y, fungsi akan mengembalikan nilai True.
    - Dalam kasus ini, Python juga menggunakan short-circuit evaluation karena apapun nilai y tidak akan mempengaruhi hasil.

E. Cara Menangani Null Value

1. Drop Kolom

Menghapus kolom dari dataset yang mengandung banyak nilai null. Ini dilakukan ketika kolom tersebut tidak memberikan kontribusi signifikan dalam analisis atau ketika mayoritas

nilainya null. Langkah ini dapat membantu mengurangi kompleksitas data dan meningkatkan kinerja model. Namun, harus dilakukan dengan hati-hati karena dapat menghilangkan informasi penting.

```
print("Dataframe asli:")
print(df)
```

```
Dataframe asli:
     A    B   C
0  1.0  5.0   9
1  2.0  NaN  10
2  NaN  7.0  11
3  4.0  8.0  12
```

```
df_cleaned_col = df.dropna(axis=1)
```

```
print("\nDataframe setelah drop kolom yang mengandung nilai null:")
print(df_cleaned_col)
```

```
Dataframe setelah drop kolom yang mengandung nilai null:
    C
0   9
1  10
2  11
3  12
```

```
print("Dataframe asli:")
print(df)
```

```
Dataframe asli:
     A    B   C
0  1.0  5.0   9
1  2.0  NaN  10
2  NaN  7.0  11
3  4.0  8.0  12
```

```
df_cleaned = df.drop('B', axis=1)
print(df_cleaned)
```

```
     A   C
0  1.0   9
1  2.0  10
2  NaN  11
3  4.0  12
```

2. Drop Baris yang Berisikan Null Value

Menghapus baris dari dataset yang mengandung nilai null. Ini dilakukan ketika jumlah baris dengan nilai null relatif sedikit dibandingkan dengan keseluruhan dataset atau ketika keberadaan nilai null dapat mempengaruhi hasil analisis secara signifikan. Namun, kehati-hatian juga diperlukan karena dapat mengurangi jumlah data yang tersedia untuk analisis.

```
print("Dataframe asli:")
print(df)
```

```
Dataframe asli:
     A    B   C
0  1.0  5.0   9
1  2.0  NaN  10
2  NaN  7.0  11
3  4.0  8.0  12
```

```
df_cleaned_row = df.dropna(axis=0)
```

```
print("\nDataframe setelah drop baris yang mengandung nilai null:")
print(df_cleaned_row)
```

```
Dataframe setelah drop baris yang mengandung nilai null:
     A    B   C
0  1.0  5.0   9
3  4.0  8.0  12
```

```
print("Dataframe asli:")
print(df)
```

```
Dataframe asli:
     A    B   C
0  1.0  5.0   9
1  2.0  NaN  10
2  NaN  7.0  11
3  4.0  8.0  12
```

```
df_cleaned2 = df.dropna(subset=['B'])
print(df_cleaned2)
```

```
     A    B   C
0  1.0  5.0   9
2  NaN  7.0  11
3  4.0  8.0  12
```

```
print("Dataframe asli:")
print(df)
```

```
Dataframe asli:
     A    B   C
0  1.0  5.0   9
1  2.0  NaN  10
2  NaN  7.0  11
3  4.0  8.0  12
```

```
df.dropna(axis=0, inplace=True)
```

```
print(df)
```

```
     A    B   C
0  1.0  5.0   9
3  4.0  8.0  12
```

3. Imputation

Mengganti nilai null dengan nilai yang diestimasi. Imputasi dapat dilakukan dengan berbagai metode, seperti mengganti nilai null dengan mean, median, modus, nilai dari observasi sebelum atau sesudahnya, atau menggunakan model prediksi. Pendekatan ini memungkinkan untuk mempertahankan ukuran dataset yang sama sambil mempertahankan informasi yang mungkin tersirat dalam data yang tersedia. Namun, imputasi juga memperkenalkan ketidakpastian ke dalam data, terutama jika teknik imputasi yang dipilih tidak sesuai dengan distribusi data atau tidak memperhitungkan struktur data dengan baik.

```
data['product_category_name'].fillna(data['product_category_name'].mode()[0], inplace=True)
data['product_category_weight'].fillna(data['product_category_weight'].mean(), inplace=True)
data['product_category_name'].fillna(data['product_category_name'].median(), inplace=True)
```

REFERENSI

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

Belajar Python: Cara Membaca Dan menulis file di python. Petani Kode. (2017, August 9). https://www.petanikode.com/python-file/

7 Jenis operator python, Fungsi, Dan Contohnya 2025. RevoU. (n.d.). https://www.revou.co/panduan-teknis/operator-python

Dokumentasi Resmi Python https://docs.python.org/3/.