Laporan I praktikum Data Mining

DATA MINING DENGAN PYTHON

Disusun untuk memenuhi

tugas praktikum mata kuliah Data Mining

Oleh :

MUHAMMAD FARID
2108108010028

JURUSAN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SYIAH KUALA

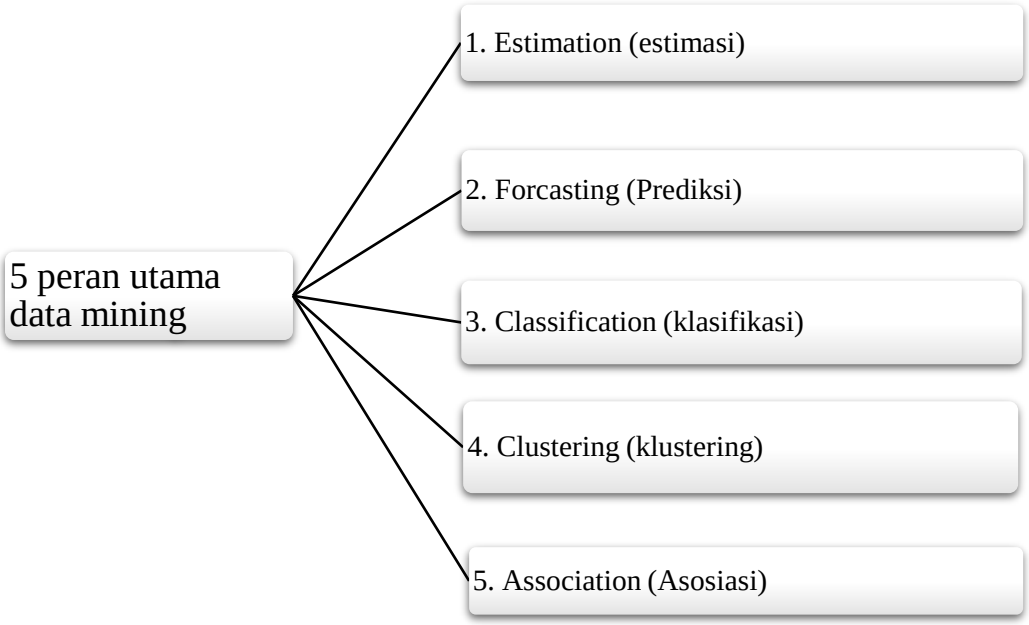
2023

SOAL :

1. Sebutkan 5 peran utama data mining!
2. Sebutkan tahapan utama proses data mining
3. Cari dataset yang dapat digunakan untuk Classification, Clustering, dan Regression (masing-masing satu dataset). Tampilkan 20 data pertama, sumbernya, dan berikan penjelasan detail mengenai jumlah observasi, deskripsi variabel, dan lainnya (jelaskan selengkap-lengkapnya).

PENYELESAIAN :

1.



2. Tahapan utama dalam data mining adalah sebagai berikut:
    - **Pemahaman terhadap data**: Tahap pertama dalam data mining adalah memahami data yang ada. Ini melibatkan pemahaman tentang jenis data yang tersedia, bagaimana data tersebut dikumpulkan, dan bagaimana data tersebut akan digunakan.
    - **Pra-pemrosesan data**: Tahap berikutnya adalah pra-pemrosesan data. Ini melibatkan menghapus data yang tidak relevan, menyeimbangkan kelas, menangani data yang hilang atau rusak, dan mengubah data mentah menjadi format yang cocok untuk proses mining.
    - **Pemilihan fitur:** Tahap ini melibatkan memilih variabel atau fitur yang paling relevan untuk diproses dalam analisis data. Pemilihan fitur membantu untuk meningkatkan efektivitas dan efisiensi proses mining.
    - **Pemodelan:** Tahap ini melibatkan pembuatan model atau algoritma untuk mengekstrak pola dari data. Beberapa teknik yang sering digunakan dalam pemodelan adalah clustering, klasifikasi, regresi, dan asosiasi.

- **Evaluasi model:** Tahap berikutnya adalah mengevaluasi model yang telah dibuat. Evaluasi ini dapat dilakukan dengan menggunakan teknik seperti validasi silang, confusion matrix, atau ROC curve.
- **Implementasi model:** Setelah model telah dievaluasi dan disempurnakan, model tersebut dapat diimplementasikan pada data baru untuk tujuan prediksi atau analisis.
- **Interpretasi dan visualisasi hasil**: Tahap akhir adalah interpretasi dan visualisasi hasil. Hasil mining yang diperoleh dapat diinterpretasikan dan divisualisasikan untuk mempermudah pemahaman tentang data dan hasil analisis.

3. Dataset

Classification

```
In [14]: data_classification_farid = pd.read_csv("mushrooms.csv", sep =";")
In [15]: data_classification_farid.head(n=20)
Out[15]:
```

	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring-type	spore-print-color	population
0	p	x	s	n	t	p	f	c	n	k	...	s	w	w	p	w	o	p	k	s
1	e	x	s	y	t	a	f	c	b	k	...	s	w	w	p	w	o	p	n	n
2	e	b	s	w	t	l	f	c	b	n	...	s	w	w	p	w	o	p	n	n
3	p	x	y	w	t	p	f	c	n	n	...	s	w	w	p	w	o	p	k	s
4	e	x	s	g	f	n	f	w	b	k	...	s	w	w	p	w	o	e	n	a
5	e	x	y	y	t	a	f	c	b	n	...	s	w	w	p	w	o	p	k	n
6	e	b	s	w	t	a	f	c	b	g	...	s	w	w	p	w	o	p	k	n
7	e	b	y	w	t	l	f	c	b	n	...	s	w	w	p	w	o	p	n	s
8	p	x	y	w	t	p	f	c	n	p	...	s	w	w	p	w	o	p	k	v
9	e	b	s	y	t	a	f	c	b	g	...	s	w	w	p	w	o	p	k	s
10	e	x	y	y	t	l	f	c	b	g	...	s	w	w	p	w	o	p	n	n
11	e	x	y	y	t	a	f	c	b	n	...	s	w	w	p	w	o	p	k	s
12	e	b	s	y	t	a	f	c	b	w	...	s	w	w	p	w	o	p	n	s
13	p	x	y	w	t	p	f	c	n	k	...	s	w	w	p	w	o	p	n	v
14	e	x	f	n	f	n	f	w	b	n	...	f	w	w	p	w	o	e	k	a
15	e	s	f	g	f	n	f	c	n	k	...	s	w	w	p	w	o	p	n	y
16	e	f	f	w	f	n	f	w	b	k	...	s	w	w	p	w	o	e	n	a
17	p	x	s	n	t	p	f	c	n	n	...	s	w	w	p	w	o	p	k	s
18	p	x	y	w	t	p	f	c	n	n	...	s	w	w	p	w	o	p	n	s
19	p	x	s	n	t	p	f	c	n	k	...	s	w	w	p	w	o	p	n	s

20 rows × 23 columns

Sumber data : https://archive.ics.uci.edu/ml/datasets/mushroom

Interpretasi :

Dataset “mushrooms” diatas merupakan data klasifikasi dan termasuk algoritma supervised. Data diatas memiliki 8124 baris dan 23 kolom (22 feature, 1 label). Kumpulan data ini mencakup deskripsi sampel hipotetis yang sesuai dengan 23 spesies jamur insang di Keluarga Agaricus dan Lepiota Setiap spesies diidentifikasi sebagai pasti dapat dimakan, pasti beracun, atau tidak dapat

dimakan dan tidak direkomendasikan. Kelas yang terakhir ini digabungkan dengan yang beracun.

Informasi Atribut:

- bentuk topi : lonceng = b,kerucut = c,cembung = x,datar = f, tombol-tombol=k,cekung=s
- tutup-permukaan: berserat = f, alur = g, bersisik = y, halus = s
- warna topi: coklat = n, buff = b, kayu mani s= c,abu-abu = g, hijau = r, merah muda = p,ungu = u,merah = e, putih = w, kuning = y
- memar: memar = t, no = f
- bau: almond = a,anise = l,creosote = c, fish y = y, foul = f, apek = m,none = n, pedas = p, pedas = s
- lampiran insang: terpasang = a, turun = d, bebas = f, berlekuk = n
- jarak insang: dekat = c, ramai = w, jauh = d
- ukuran insang: lebar = b,sempit = n
- warna insang: hitam=k,coklat=n,buff=b,coklat=h,abu-abu=g,hijau=r,
- oranye=o,merah muda=p,ungu=u,merah=e,putih=w, kuning = y
- bentuk tangkai: membesar=e,runcing=t
- tangkai-akar: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
- batang-permukaan-di atas-cincin: berserat=f, bersisik=y, halus=k, halus=s
- tangkai-permukaan-bawah-cincin: berserat=f,bersisik=y,halus=k,halus=s
- tangkai-warna-di atas-cincin: coklat=n,buff=b,kayu manis=c,abu-abu=g,oranye=o,merah muda=p,merah=e,putih=w,kuning=y
- batang-warna-di bawah-cincin: coklat=n,buff=b,kayu manis=c,abu-abu=g,oranye=o,merah muda=p,merah=e,putih=w,kuning=y
- tipe cadar: parsial=p,universal=u
- warna cadar: coklat=n,oranye=o,putih=w,kuning=y
- nomor dering: tidak ada = n, satu = o, dua = t
- ring-type: cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z

- spora-cetak-warna: hitam=k,coklat=n,buff=b,coklat=h,hijau=r, orange=o,ungu=u,putih=w,kuning=y
- populasi: berlimpah=a, berkerumun=c, banyak=n, tersebar=s, beberapa=v, soliter=y
- habitat: rerumputan=g,daun=l,padang rumput=m,jalur=p, perkotaan=u,limbah=w,kayu=d

Regression

```
In [12]: data_regression_farid = pd.read_csv("C:/Users/ASUS/Downloads/AirQualityUCI.csv", sep =";")

In [13]: data_regression_farid.head(n=20)

Out[13]:
```

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH
0	10/03/2004	18.00.00	2.6	1360	150	11.9	1046	166	1056	113	1692	1268	13.6	48.9
1	10/03/2004	19.00.00	2.0	1292	112	9.4	955	103	1174	92	1559	972	13.3	47.7
2	10/03/2004	20.00.00	2.2	1402	88	9.0	939	131	1140	114	1555	1074	11.9	54.0
3	10/03/2004	21.00.00	2.2	1376	80	9.2	948	172	1092	122	1584	1203	11.0	60.0
4	10/03/2004	22.00.00	1.6	1272	51	6.5	836	131	1205	116	1490	1110	11.2	59.6
5	10/03/2004	23.00.00	1.2	1197	38	4.7	750	89	1337	96	1393	949	11.2	59.2
6	11/03/2004	00.00.00	1.2	1185	31	3.6	690	62	1462	77	1333	733	11.3	56.8
7	11/03/2004	01.00.00	1.0	1136	31	3.3	672	62	1453	76	1333	730	10.7	60.0
8	11/03/2004	02.00.00	0.9	1094	24	2.3	609	45	1579	60	1276	620	10.7	59.7
9	11/03/2004	03.00.00	0.6	1010	19	1.7	561	-200	1705	-200	1235	501	10.3	60.2
10	11/03/2004	04.00.00	-200.0	1011	14	1.3	527	21	1818	34	1197	445	10.1	60.5
11	11/03/2004	05.00.00	0.7	1066	8	1.1	512	16	1918	28	1182	422	11.0	56.2
12	11/03/2004	06.00.00	0.7	1052	16	1.6	553	34	1738	48	1221	472	10.5	58.1
13	11/03/2004	07.00.00	1.1	1144	29	3.2	667	98	1490	82	1339	730	10.2	59.6
14	11/03/2004	08.00.00	2.0	1333	64	8.0	900	174	1136	112	1517	1102	10.8	57.4
15	11/03/2004	09.00.00	2.2	1351	87	9.5	960	129	1079	101	1583	1028	10.5	60.6
16	11/03/2004	10.00.00	1.7	1233	77	6.3	827	112	1218	98	1446	860	10.8	58.4
17	11/03/2004	11.00.00	1.5	1179	43	5.0	762	95	1328	92	1362	671	10.5	57.6
18	11/03/2004	12.00.00	1.6	1236	61	5.2	774	104	1301	95	1401	664	9.5	66.8
19	11/03/2004	13.00.00	1.9	1286	63	7.3	869	146	1162	112	1537	799	8.3	76.4

Sumber data : https://archive.ics.uci.edu/ml/datasets/Air+Quality

Interpretasi :

Data diatas memiliki 9358 baris dan 15 atribut. Data diatas berisi tanggapan dari perangkat multisensor gas yang ditempatkan di lapangan di kota Italia. Rata-rata respons per jam dicatat bersama dengan referensi konsentrasi gas dari penganalisis bersertifikat. Kumpulan data berisi 9358 contoh tanggapan rata-rata per jam dari susunan 5 sensor kimia oksida logam yang disematkan dalam Perangkat Multisensor Bahan Kimia Kualitas Udara. Perangkat itu terletak di lapangan di area yang sangat tercemar, di permukaan jalan, di dalam kota Italia. Data direkam dari Maret 2004 hingga Februari 2005 (satu tahun) yang mewakili rekaman terlama yang tersedia secara bebas dari respons perangkat sensor kimia kualitas udara yang digunakan di lapangan. Ground Truth rata-rata konsentrasi per jam untuk CO, Hidrokarbon Non Metanik, Benzena, Total Nitrogen Oksida

(NOx), dan Nitrogen Dioksida (NO2) dan disediakan oleh penganalisa bersertifikat referensi yang ditempatkan di lokasi yang sama.

Penjelasan atribut :

- Atribut 1 = Tanggal (DD/MM/YYYY)
- Atribut 2 = waktu (HH.MM.SS)
- Atribut 3 = Konsentrasi CO rata-rata per jam sebenarnya dalam mg/m^3 (penganalisa referensi)
- Atribut 4 = PT08.S1 (timah oksida) respons sensor rata-rata per jam (target CO nominal)
- Atribut 5 = Konsentrasi HidroKarbon Non Metanik keseluruhan rata-rata per jam sebenarnya dalam mikrog/m^3 (penganalisa referensi)
- Atribut 6 = Konsentrasi Benzena rata-rata per jam sebenarnya dalam mikrog/m^3 (penganalisa referensi)
- Atribut 7 = PT08.S2 (titania) respons sensor rata-rata per jam (target NMHC nominal)
- Atribut 8 = Konsentrasi NOx rata-rata per jam sebenarnya dalam ppb (penganalisa referensi)
- Atribut 9 = Respon sensor rata-rata per jam PT08.S3 (tungsten oksida) (menargetkan NOx nominal)
- Atribut 10 = Konsentrasi NO2 rata-rata per jam sebenarnya dalam mikrog/m^3 (penganalisa referensi)
- Atribut 11 = Respons sensor rata-rata PT08.S4 (tungsten oksida) per jam (target NO2 nominal)
- Atribut 12 = Respon sensor rata-rata per jam PT08.S5 (indium oksida) (target O3 nominal)
- Atribut 13 = Suhu dalam Â°C
- Atribut 14 = Kelembaban Relatif (%)
- Atribut 15 = AH Kelembaban Mutlak

- Clustering

```
In [18]: import pandas as pd

In [20]: data_clustering_farid = pd.read_excel("user_knowlagde_modelling.xlsx",sheet_name="user_knowlagde_modelling")

In [21]: data_clustering_farid.head(n=20)
```

Out[21]:

	STG	SCG	STR	LPR	PEG
0	0.00	0.00	0.00	0.00	0.00
1	0.08	0.08	0.10	0.24	0.90
2	0.06	0.06	0.05	0.25	0.33
3	0.10	0.10	0.15	0.65	0.30
4	0.08	0.08	0.08	0.98	0.24
5	0.09	0.15	0.40	0.10	0.66
6	0.10	0.10	0.43	0.29	0.56
7	0.15	0.02	0.34	0.40	0.01
8	0.20	0.14	0.35	0.72	0.25
9	0.00	0.00	0.50	0.20	0.85
10	0.18	0.18	0.55	0.30	0.81
11	0.06	0.06	0.51	0.41	0.30
12	0.10	0.10	0.52	0.78	0.34
13	0.10	0.10	0.70	0.15	0.90
14	0.20	0.20	0.70	0.30	0.60
15	0.12	0.12	0.75	0.35	0.80
16	0.05	0.07	0.70	0.01	0.05
17	0.10	0.25	0.10	0.08	0.33
18	0.15	0.32	0.05	0.27	0.29
19	0.20	0.29	0.25	0.49	0.56

Sumberdata :

https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling

Interpretasi :

Data diatas memiliki 403 baris dan 5 atribut.

Informasi Atribut:

STG (Tingkat waktu studi untuk materi objek tujuan),

SCG (Tingkat jumlah pengulangan pengguna untuk materi objek tujuan)

STR (Tingkat waktu studi pengguna untuk objek terkait dengan objek tujuan)

LPR (Kinerja ujian pengguna untuk objek terkait dengan objek tujuan)

PEG (Kinerja ujian pengguna untuk objek sasaran)