MODUL DATA MINING

Data Understanding

Pada modul ini dijelaskan mengenai contoh pemahaman data dengan berbagai metode.

Diharapkan setelah mempelajari modul ini, mahasiswa mampu memahami tujuan dari pemahaman data dan mengimplementasikan pada kasus yang memerlukan ekplorasi data.

EPS 2

DAFTAR ISI

DATA UNDERSTANDING

Data understanding/ pemahaman data merupakan tahap yang dilaksanakan setelah tujuan dan cakupan proyek data mining ditetapkan pada tahap pemahaman bisnis. Dalam tahap pemahaman data ini dilakukan berbagai kegiatan yaitu

1. Pengumpulan data
   Berdasarkan tujuan dilakukannya proyek data mining, maka ditentukan kebutuhan data. Data yang dibutuhkan tersebut dikumpulkan dan dilakukan pendokumentasian mengenai sumber dan jenis-jenis data maupun variable/atribut dari data tersebut.
2. Eksplorasi data
   Eksplorasi data yang dilakukan dapat menggunakan metode statistik maupun visual. Metode statistik yang dapat digunakan misalnya melihat total data, nilai rata-rata, nilai minimal maupun maksimal, dan sebagainya. Disamping metode tersebut, dapat juga digunakan metode visual yang menggunakan diagram sebagai representasi data. Diagram yang digunakan sebaiknya disesuaikan dengan tujuan eksplorasi data, misalnya diagram batang (bar chart) digunakan untuk membandingkan nilai, diagram garis (line/ trend chart) digunakan untuk melihat kenaikan atau penurunan nilai dalam rentang waktu tertentu, maupun diagram lingkaran (pie chart) yang digunakan untuk melihat proporsi data.
3. Verifikasi kualitas data.
   Kualitas dari data perlu diketahui pada tahap ini. Hal ini dikarenakan kualitas data masukan akan berdampak pada hasil akhir. Pada tahap ini dilakukan verifikasi mengenai kualitas data, jika kualitas data perlu diperbaiki, maka perbaikannya dapat dilakukan pada tahap selanjutnya. Salah satu yang mempengaruhi kualitas data adalah kelengkapan sehingga jika terjadi data yang kosong (null) maka pada tahap selanjutnya dibuat perencanaan scenario untuk mengatasi data kosong tersebut.

Pada modul ini hanya akan dibahas mengenai pengecekan tipe data, eksplorasi, dan verifikasi kualitas data (missing value).

A. Load Data dan Library

1. Import library yang diperlukan

```
import pandas as pd
from pandas.tools.plotting import scatter_matrix
pd.set_option('display.width', 500)
pd.set_option('display.max_columns', 100)
pd.set_option('display.notebook_repr_html', True)
```

2. Load dataset ‘train.csv’. Path disesuaikan dengan lokasi penyimpanan file tersebut.

```
data = pd.read_csv('[redacted]/train.csv')
```

   Cek isi dataset.

```
data.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

B. Dokumentasi Tipe Data

Pendokumentasian tipe data dari masing-masing variable/atribut penting dilakukan. Hal ini dikarenakan operasi-operasi tertentu hanya dapat diterapkan terhadap tipe data tertentu saja. Misalnya operasi matematika hanya dapat diterapkan pada variable/atribut yang bersifat numerik misalnya integer atau float dan tidak dapat diterapkan pada tipe data string atau object.

1. Cek jumlah data (baris data) dan jumlah variabel/atribut

```
data = pd.DataFrame(data)
print (data.shape)
```

```
(891, 12)
```

Dari hasil tersebut diartikan bahwa total baris data adalah sebanyak 891 dan jumlah variabel/atributnya adalah 12 atribut.

2. Cek tipe data pada seluruh variable/atribut

```
print (data.dtypes)
```

```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

3. Cek tipe data pada salah satu variable/atribut

```
print (data['Age'].dtypes)
```

```
float64
```

Jika diperhatikan, tipe data selain numerik (integer dan float) yaitu string dicantumkan sebagai object. Hal ini dikarenakan data yang di-load dibaca menggunakan library pandas sehingga dikenali dalam bentuk dataframe. Silahkan pelajari mengenai dataframe yang digunakan oleh library pandas.

C. Exploratory Data Analysis (EDA)

1. Pendekatan statistik sederhana

```
data.describe()
```

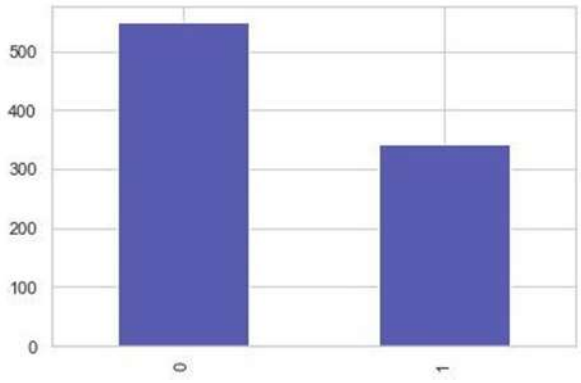
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Tugas 1:
Terjemahkan hasil dari fungsi describe yang baru saja anda lakukan.

2. Pendekatan visual
   - Distribusi data pada suatu variable/atribut

```
data['Survived'].value_counts().plot(kind='bar')
data['Survived'].value_counts()
```

```
0    549
1    342
Name: Survived, dtype: int64
```



Dari diagram diatas, dapat diamati bahwa jumlah yang meninggal adalah 549, yang selamat 342.

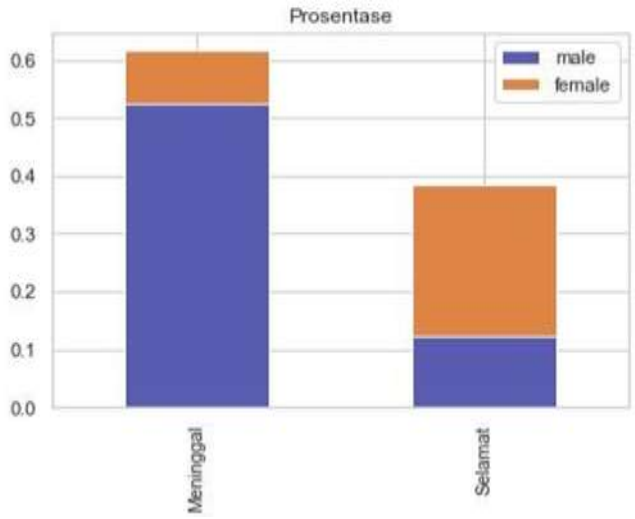
   - Perbandingan antar variable/atribut
     Buat fungsi untuk menghitung dan menampilkan diagramnya.

```
def survival_stacked_bar(variable):
    died=data[data['Survived']==0][variable].value_counts()/len(data['Survived']==0)
    survived=data[data['Survived']==1][variable].value_counts()/len(data['Survived']==1)
    dataset=pd.DataFrame([died,survived])
    dataset.index=['Meninggal', 'Selamat']
    dataset.plot(kind='bar',stacked=True,title='Prosentase')
    return dataset.head()
```

Panggil fungsi tersebut.

```
survival_stacked_bar('Sex')
```

	male	female
Meninggal	0.525253	0.090909
Selamat	0.122334	0.261504



Tugas 2 :

Buat sebanyak mungkin diagram yang menggambarkan relasi berbagai macam kolom sesuai interpretasi anda. Cantumkan screenshot script dan hasilnya, kemudian jelaskan makna dari diagram tersebut.

D. Kualitas Data (missing value)

1. Data perlu dicek apakah terdapat data yang kosong atau tidak.

```
data.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
```

2. Maka selanjutnya perlu dilakukan pembuatan scenario terhadap data kosong ini.
   Penanganan data kosong dapat dilakukan dengan cara:
   a. Menghapus atribut yang memiliki nilai null
   b. Menghapus data baris yang memiliki nilai null
   c. Mengisi dengan nilai baru dengan cara:
      - Mengisi manual

- Mengisi dengan menggunakan metode imputation (mean, median, modus, dll)

Hal penting yang perlu diingat dalam penanganan missing value adalah seberapa penting data kosong tersebut mempengaruhi hasil. Misalnya jika atribut umur (Age) mempengaruhi selamat atau tidaknya seorang penumpang maka atribut umur sebaiknya tidak dihapus. Nilai null-nya dapat ditangani dengan cara lainnya misalnya baris yang kosong yang dihapus atau nilai yang kosong tersebut diisi.

Hal ini perlu merujuk pada deskripsi masing-masing varibel/atribut. Selain itu, juga perlu dipertimbangkan keterkaitan/korelasi antara varibel/atribut dengan variabel tujuan (misalnya: apakah atribut umur mempengaruhi selamat/tidaknya penumpang)

Tugas 3:

- Jelaskan kelebihan dan kekurangan dari masing-masing metode penanganan missing value yang telah dijelaskan diatas
- Buat scenario penanganan missing value terhadap data kosong yang telah ditemukan tersebut.

LATIHAN MAHASISWA

1. Silahkan ikuti dan praktekkan setiap tahapan yang dijelaskan pada modul.
2. Kerjakan soal yang tercantum pada tahapan tersebut.
   Jawaban sebaiknya dilengkapi dengan screenshot script dan hasil dari running script tersebut baru diikuti dengan penjelasan.
3. Hasil dari soal no 2 dituliskan dalam bentuk laporan mandiri dalam satu file PDF.
   Format laporan:
   Subyek file " Modul2-DM-[KELAS]-[NPM] .PDF "