MODUL DATA MINING

Data Preparation

Pada modul ini dijelaskan mengenai contoh penyiapan data dengan berbagai cara.
Diharapkan setelah mempelajari modul ini, mahasiswa mampu memahami tujuan transformasi dan mengimplementasikan pada kasus transformasi data.

EPS 3

DAFTAR ISI

DATA PREPARATION

Data preparation/penyiapan data merupakan tahap yang dilaksanakan setelah data understanding/pemahaman data. Setelah diketahui kondisi mengenai data yang akan digunakan dalam tahap modeling, maka pada tahap ini akan dilakukan penyiapan data sehingga data tersebut siap untuk masuk dalam tahap modeling. Pada tahap ini akan dilakukan pemilihan data (atribut dan/atau instance), transformasi data, agregasi, merge atribut, dan sebagainya menyesuaikan dengan kebutuhan data yang diharapkan. Pada modul ini akan dicontohkan transformasi data dan pemilihan fitur.

A. Transformasi Data

1. Panggil library yang akan digunakan

```
import pandas as pd
```

2. Panggil data yang akan digunakan

```
data = pd.read_csv('G:/train.csv')
```

3. Cek isi data, apakah sesuai dengan yang diharapkan

```
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

Ekstraksi dan pengelompokkan

4. Pada data tersebut, tidak ada atribut yang menjelaskan mengenai apakah penumpang merupakan staf/officer atau penumpang yang memililiki gelar bangsawan atau tidak. Kita dapat mengetahui title penumpang tersebut dari atribut nama penumpang.
   Cek pada atribut nama penumpang ada title apa saja?
   Cek apakah pada atribut nama penumpang terdapat data yang kosong/null?

5. Setelah diketahui daftar titlenya apa saja, maka kita dapat mengekstrak data title penumpang ini dan menyimpannya dalam atribut yang berbeda.
   Pecah atribut nama sesuai dengan kondisi isi dari atribut tersebut dan masukkan data hasil pecahan tersebut ke dalam kolom baru 'Title' yang terletak di posisi ke 12 alias setelah kolom terakhir.

```
data.insert(value=data.Name.map(lambda name: name.split(",")[1].split(".")[0].strip()),loc=12,column="Title")
```

6. Kemudian lakukan transformasi title penumpang ke dalam kelompok title yang sudah ditentukan.

```
title_map={"Capt":"Officer",
      "Col":"Officer",
      "Major":"Officer",
      "Johkheer":"Royalty",
      "Don":"Royalty",
      "Sir":"Royalty",
      "Dr":"Royalty",
      "Rev":"Officer",
      "The Countess":"Royalty",
      "Dona":"Royalty",
      "Mme":"Mrs",
      "Mlle":"Miss",
      "Ms":"Mrs",
      "Mr":"Mr",
      "Mrs":"Mrs",
      "Miss":"Miss",
      "Master":"Master",
      "Lady":"royalty"}
data["Title"]=data.Title.map(title_map)
```

7. Maka isi dari atribut title akan berubah sesuai dengan transformasi ke dalam kelompok title yang sudah ditentukan.
   Cek data apakah berubah?

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	Mr
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	Mrs
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	Miss
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	Mrs
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	Mr
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q	Mr
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S	Mr
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S	Master
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S	Mrs
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C	Mrs
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S	Miss

8. Cek jumlah dari masing-masing kelompok title. Berapa orang yang termasuk officer, royal, dan lainnya?

9. Pada data tersebut tidak ada atribut yang menjelaskan bahwa penumpang merupakan dewasa atau anak-anak. Sedangkan kelompok usia seperti ini dapat menambah informasi mengenai kondisi penumpang. (analisis ini dilakukan pada data understanding).
   Oleh karenanya, kita perlu menambahkan kolom yang menjelaskan kelompok usia tersebut. Kelompok usia dapat diekstrak dari atribut umur.
   Cek apakah pada atribut umur ada yang kosong/null?

```
data["Age"].isnull().sum()
```

```
177
```

10. Karena pada atribut umur banyak yang kosong, maka untuk kelompok umur ini sebaiknya kita ambil dari title. Karena kita dapat menentukan secara umum kelompok umur seseorang dilihat titlenya.
    Buat fungsi untuk mengelompokkan kelompok umur dengan menggunakan aturan kondisi.

```
def passenger_type (row):
    if row['Age'] < 2 :
        return 'Infant'
    elif (row['Age'] >= 2 and row['Age'] < 12):
        return 'Child'
    elif (row['Age'] >= 12 and row['Age'] < 18):
        return 'Youth'
    elif (row['Age'] >= 18 and row['Age'] < 65):
        return 'Adult'
    elif row['Age'] >= 65:
        return 'Senior'
    elif row['Title'] == 'Master':
        return 'Child'
    elif row['Title'] == 'Miss':
        return 'Child'
    elif row['Title'] == 'Mr' or row['Title'] == 'Mrs':
        return 'Adult'
    else:
        return 'Unknown'
```

11. Kemudian panggil fungsi tersebut untuk mengektrak kelompok umur, dan hasilnya akan dimasukkan dalam atribut dengan nama ‘passeger_type’

```
data['PassengerType'] = data.apply(lambda row: passenger_type(row),axis=1)
```

Cek data, apakah atribut bertambah?

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title	PassengerType
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	Mr	Adult
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	Mrs	Adult
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	Miss	Adult
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	Mrs	Adult
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	Mr	Adult
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q	Mr	Adult

12. Cek jumlah dari masing-masing kelompok umur. Berapa orang yang senior, dewasa, dan anak-anak?

Encoding

Encoding merupakan teknik yang dilakukan untuk mengubah data alphanumeric menjadi numerik. Contoh: {apel, mangga, lemon} -> {1, 2, 3} ; 1 merupakan apel, 2 merupakan mangga, dan 3 merupakan lemon. Encoding dilakukan salah satu alasannya adalah terkadang algoritma tidak mengenali bentuk alphanumeric sehingga data harus diubah menjadi bentuk numerik.

Encoding ini dilakukan dengan berbagai teknik. Python menyediakan library untuk melakukan proses encoding ini. Namun sebelum itu, berikut ini adalah cara manual untuk melakukan encoding.

13. Cara manual adalah dengan melakukan mapping, seperti yang dilakukan pada cara pengelompokkan kelompok title sebelumnya.
    Berikut ini adalah cara untuk melakukan mapping atribut jenis kelamin. Dengan laki-laki : 1 dan perempuan : 0

```
sex_map={"male":1,"female":0}
data["Sex"]=data["Sex"].map(sex_map)
```

14. Cek apakah isian atribut jenis kelamin berubah?

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title	PassengerType
0	1	0	3	Braund, Mr. Owen Harris	1	22.0	1	0	A/5 21171	7.2500	NaN	S	Mr	Adult
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	38.0	1	0	PC 17599	71.2833	C85	C	Mrs	Adult
2	3	1	3	Heikkinen, Miss. Laina	0	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	Miss	Adult
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.0	1	0	113803	53.1000	C123	S	Mrs	Adult
4	5	0	3	Allen, Mr. William Henry	1	35.0	0	0	373450	8.0500	NaN	S	Mr	Adult
5	6	0	3	Moran, Mr. James	1	NaN	0	0	330877	8.4583	NaN	Q	Mr	Adult

15. Coba lakukan mapping ini pada PassengerType
    Dengan aturan mapping sebagai berikut:
    Unknown : 0
    Infant : 1
    Child : 2
    Youth : 3
    Adult : 4
    Senior : 5

16. Cara manual ini sayangnya memiliki kelemahan yaitu jika isi data diluar dari yang didefinisikan dalam mapping, maka data tersebut tidak dapat dikonversi.
    Misalnya dalam kelompok umur ada data yang kelompok umurnya adalah super senior, karena pada aturan mapping tidak didefinisikan maka super senior tidak dapat dikonversi.
    Selain itu, jika melakukan mapping manual seperti ini, ini hanya diperuntukkan jika tipe atributnya adalah ordinal.
    Jika tetap memaksa untuk mapping jenis kelamin seperti cara diatas (tahap 13), maka nantinya pada saat pemodelan, model akan melihat adanya order atau urutan. Misalnya diubah aturannya dengan laki-laki adalah 1 dan perempuan adalah 2. Maka model akan mengenali bahwa perempuan sama dengan 2 laki-laki.
    Sehingga cara encoding harus memperhatikan tipe atribut.

Ordinal Encoding

17. Jika tipe atribut adalah atribut yaitu memperhatikan adanya order atau urutan, maka cara mapping manual sebelumnya dapat digunakan. Selain itu, jika menggunakan cara mapping manual, maka dipastikan bahwa mapping sudah fix dan tidak ada kemungkinan adanya kelompok baru.

    Cara selain mapping manual adalah menggunakan library sklearn.

```
from sklearn import preprocessing
```

    Pada library ini terdapat dua jenis encoding yang disediakan untuk encoding tipe ordinal yaitu

    a. ordinal encoder (https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OrdinalEncoder.html)

b. label encoder (https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html)

Categorical Encoding

18. Jika tipe atribut adalah categorical, alias nominal namun tidak ada urutan, misalnya atribut jenis kelamin. Maka salah satu cara yang dapat dilakukan untuk encoding adalah dengan menggunakan **one-hot encoding.**
    Teknik one-hot encoding ini dapat menggunakan library Get Dummies yang disediakan oleh pandas atau oneHotEncoder yang disediakan oleh sklearn (https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html)

19. Untuk mencoba teknik teknik encoding silahkan cek link: https://medium.com/@adiptamartulandi/data-preprocessing-pada-machine-learning-handling-categorical-data-ucupstory-6e409dbfd0a0.
    Buka github-nya.

20. Coba teknik encoder untuk nominal dan ordinal pada data titanic dengan masing-masing tipe atribut pilih satu teknik saja.

B. Pemilihan Fitur/Attribut

Masih ingat tentang curse of dimensionality? Pada bagian ini kita akan memilih dan menentukan atribut mana saja yang akan kita gunakan pada tahap selanjutnya. Pemilihan atribut ini dapat menggunakan berbagai cara.

Salah satu cara untuk memilih atribut adalah dengan teknik korelasi. Korelasi yang dapat dihitung dengan python adalah tiga jenis korelasi yaitu: pearson, spearman, dan kendall. Penghitungan korelasi dapat telah disediakan pada library Pandas atau Numpy.

Berikut ini contoh hitung korelasi menggunakan pandas, dengan deafult adalah pearson correlation.

1. Panggil library untuk visualisasi

```
import seaborn as sns
```

2. Buat fungsi untuk hitung korelasi

```
def titanic_corr(data):
    correlation = data.corr()
    sns.heatmap(correlation, annot=True, cbar=True, cmap="RdYlGn")
```

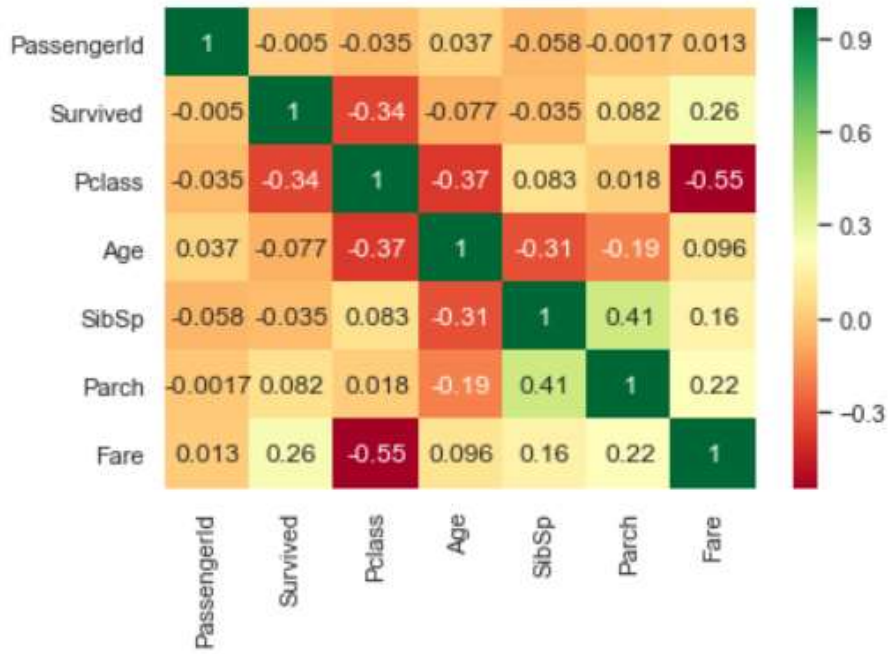
3. Panggil dan jalankan fungsi tersebut.

```
titanic_corr(data)
```

4. Jika ingin menampilkan nilai hasil korelasi tanpa bentuk heatmap, maka tuliskan code berikut.

```
data.corr()
```

5. Hasilnya adalah sebagai berikut (visualisasi heatmap). Jika diperhatikan, atribut yang dapat dihitung korelasinya adalah atribut numeric saja.
   Korelasi ini dihitung sebelum melakukan proses encoding, untuk melakukan segala jenis operasi statistik sebaiknya atribut sudah dalam bentuk numerik (dengan cara encoding).



6. Coba hitung korelasi setelah atribut telah di-encoding.
   Apa maksud hasil korelasi tersebut?

7. Untuk passangerId, yang sifatnya adalah unik per instance-nya maka atribut tersebut dihapus atau tidak?

8. Berikut ini adalah cara untuk menghapus atribut.

```
data.drop(["PassengerId","Age","Parch","Name","Ticket","Embarked","Cabin"], inplace=True,axis=1)
```

9. Berikut ini adalah cara untuk untuk menghabus instance/baris

```
data = data.drop([556, 759, 822], axis = 0)
```

10. Silahkan hapus atribut yang tidak anda gunakan dari data anda. Jelaskan mengapa anda menghapus atribut tersebut.

LATIHAN MAHASISWA

1. Silahkan dicoba setiap tahapannya.
2. Kerjakan soal yang tercantum pada tahapan tersebut.
3. Buat video dari pengerjaan nomor 2.
   Pada video jelaskan tahapan dan maksud dari setiap tahapan tersebut.
   Penjelasan menggunakan voice over, bukan tulisan yang tercantum pada video.