Praktikum Data Preprocessing

Data Normalization

Ulima Inas Shabrina(2110181048)

Latihan Pertemuan 3
Assignment#

1. dataset ← titanic.csv, dan tampilkan
2. rows, cols ← jumlah baris dan kolom pada dataset, dan tampilkan
3. data ← ambil dataset kolom fitur (Age, Fare), dan tampilkan
4. class ← ambil dataset kolom kelas (Survived)
5. Lakukan pengisian missing value pada fitur Age dengan nilai mean dari masing-masing class

Lakukan normalisasi pada data dengan algoritma berikut, dan tampilkan:

6. Min-Max (0-1)
7. Z-Score
8. Sigmoidal

Menampilkan data titanic.csv

In [8]:
```
import pandas as pd
dataset = pd.read_csv('titanic.csv')

dataset
```

Out[8]:

	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
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Menampilkan jumlah baris dan kolom pada dataset

In [9]:
```
rows, cols = dataset.shape

print('Jumlah Baris', rows)
print('Jumlah Kolom', cols)
```

```
Jumlah Baris 891
Jumlah Kolom 12
```

Menampilkan dataset kolom fitur Age dan Fare

In [10]:
```
data = pd.DataFrame(dataset, columns = ['Age', 'Fare'])

data
```

Out[10]:

	Age	Fare
0	22.0	7.2500
1	38.0	71.2833
2	26.0	7.9250
3	35.0	53.1000
4	35.0	8.0500
...	...	...
886	27.0	13.0000
887	19.0	30.0000
888	NaN	23.4500
889	26.0	30.0000
890	32.0	7.7500

891 rows × 2 columns

Menampilkan dataset kolom Fitur Survivdes

In [21]:
```
class_label = pd.DataFrame(dataset, columns = ['Survived'])

class_label
```

Out[21]:

	Survived
0	0
1	1
2	1
3	1
4	0
...	...
886	0
887	1
888	0
889	1
890	0

891 rows × 1 columns

Pengisian missing value pada fitur Age

In [10]:
```
dataset_fill = dataset.fillna(dataset.groupby('Survived').transform('mean'))
dataset_fill
```

Out[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	30.626179	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Min_Max 0-1

In [18]:
```
data_min_max = pd.DataFrame(dataset, columns = ['Age', 'Fare'])
new_min = 0
new_max = 1

column = 'Age'
minval_age = data_min_max[column].min()
maxval_age = data_min_max[column].max()

data_min_max[column] = ((data_min_max[column] - minval_age)*(new_max - new_min)/(maxval_age - minval_age)) + new_min

column = 'Fare'
minval_fare = data_min_max[column].min()
maxval_fare = data_min_max[column].max()

data_min_max[column] = ((data_min_max[column] - minval_fare)*(new_max - new_min)/(maxval_fare - minval_fare)) + new_min

data_min_max
```

Out[18]:

	Age	Fare
0	0.271174	0.014151
1	0.472229	0.139136
2	0.321438	0.015469
3	0.434531	0.103644
4	0.434531	0.015713
...	...	...
886	0.334004	0.025374
887	0.233476	0.058556
888	NaN	0.045771
889	0.321438	0.058556
890	0.396833	0.015127

Z Score

In [19]:
```
data_z = pd.DataFrame(dataset, columns = ['Age', 'Fare'])

column = 'Age'
data_z[column] = (data_z[column] - data_z[column].mean()) / data_z[column].std()

column = 'Fare'
data_z[column] = (data_z[column] - data_z[column].mean()) / data_z[column].std()

data_z
```

Out[19]:

	Age	Fare
0	-0.530005	-0.502163
1	0.571430	0.786404
2	-0.254646	-0.488580
3	0.364911	0.420494
4	0.364911	-0.486064
...	...	...
886	-0.185807	-0.386454
887	-0.736524	-0.044356
888	NaN	-0.176164
889	-0.254646	-0.044356
890	0.158392	-0.492101

891 rows × 2 columns

Sigmodial

In [22]:
```
import numpy as np
data_norm = pd.DataFrame(dataset, columns = ['Age', 'Fare'])

column = 'Age'
data_norm[column] = (data_norm[column] - data_norm[column].mean()) / data_norm[column].std()
data_sigmodial = (1-np.exp((-data_norm))) / (1+np.exp((-data_norm)))

column = 'Fare'
data_norm[column] = (data_norm[column] - data_norm[column].mean()) / data_norm[column].std()
data_sigmodial = (1-np.exp((-data_norm))) / (1+np.exp((-data_norm)))

data_sigmodial
```

Out[22]:

	Age	Fare
0	-0.258969	-0.245935
1	0.278186	0.374117
2	-0.126640	-0.239544
3	0.180458	0.207203
4	0.180458	-0.238358
...	...	...
886	-0.092637	-0.190857
887	-0.352471	-0.022174
888	NaN	-0.087855
889	-0.126640	-0.022174
890	0.079031	-0.241203

891 rows × 2 columns

Terimakasih