Darshan UNIVERSITY

योग: कर्मसु कौशलम्

Data Mining

Lab - 5

Data Preprocessing

Name: Smit Maru

Enrollment No: 23010101161

1) First, you need to read the titanic dataset from local disk and display Last five records

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
df = pd.read_csv(r'titanic.csv', encoding='ISO-8859-1')
print(df)
```

```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                Name     Sex   Age  SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                              Heikkinen, Miss. Laina  female  26.0      0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                            Allen, Mr. William Henry    male  35.0      0
..                                                ...     ...   ...    ...
886                              Montvila, Rev. Juozas    male  27.0      0
887                       Graham, Miss. Margaret Edith  female  19.0      0
888           Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                              Behr, Mr. Karl Howell    male  26.0      0
890                                Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     ...               ...      ...   ...      ...
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

2) Handle Missing Values in data set [use dropna(), fillna(), and interpolate]

In [3]:
```
data_withfillna = df.copy()
mean_age = data_withfillna['Age'].mean()
data_withfillna['Age'] = data_withfillna['Age'].fillna(mean_age)
data_withfillna.head(5)
```

Out[3]:

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

In [4]:

```
data_withfillna = df.copy()
median_age = data_withfillna['Age'].median()
data_withfillna['Age'] = data_withfillna['Age'].fillna(median_age)
data_withfillna.head(5)
```

Out[4]:

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

In [5]:

```
data_withfillna = df.copy()
mode_age = data_withfillna['Age'].mode()
data_withfillna['Age'] = data_withfillna['Age'].fillna(mode_age)
data_withfillna.head(5)
```

Out[5]:

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

3) Apply Scaling to AGE attribute with min max, decimal scaling and z score.

In [6]:
```
min_age = data_withfillna['Age'].min()
max_age = data_withfillna['Age'].max()
data_withfillna['Age_MinMax'] = (data_withfillna['Age'] - min_age) / (max_age - min
data_withfillna['Age_MinMax'].head(5)
```

Out[6]:
```
0    0.271174
1    0.472229
2    0.321438
3    0.434531
4    0.434531
Name: Age_MinMax, dtype: float64
```

In [7]:
```
max_abs_age = data_withfillna['Age'].abs().max()
j = len(str(int(max_abs_age)))
data_withfillna['Age_DecimalScaling'] = data_withfillna['Age'] / (10 ** j)
data_withfillna['Age_DecimalScaling'].head(5)
```

Out[7]:
```
0    0.22
1    0.38
2    0.26
3    0.35
4    0.35
Name: Age_DecimalScaling, dtype: float64
```

In [8]:
```
mean = data_withfillna['Age'].mean()
sd = data_withfillna['Age'].std()
data_withfillna['Age_ZScore'] = (data_withfillna['Age'] - mean) / sd
data_withfillna['Age_ZScore'].head(5)
```

Out[8]:
```
0   -0.530005
1    0.571430
2   -0.254646
3    0.364911
4    0.364911
Name: Age_ZScore, dtype: float64
```