MACHINE LEARNING

PRACTICE LAB

PREPROCESSING

Nikki Evana Blessy. N

2341459

Implement the concept of data preprocessing on your assigned domain. Submit both pdf and .ipynb file

```
[4]  import pandas as pd

     # Import the dataset
     from google.colab import files
     uploaded=files.upload()

     import pandas as pd
     df=pd.read_csv('World-happiness-report-2024.csv')
```

```
Choose Files  World-happ...rt-2024.csv
• World-happiness-report-2024.csv(text/csv) - 8618 bytes, last modified: 12/3/2024 - 100% done
Saving World-happiness-report-2024.csv to World-happiness-report-2024.csv
```

```
# Display the first few rows of the dataset
print("Dataset Preview:")
print(df.head())
```

```
Dataset Preview:
   Ladder score  upperwhisker  lowerwhisker  Log GDP per capita  \
0         7.741         7.815         7.667               1.844
1         7.583         7.665         7.500               1.908
2         7.525         7.618         7.433               1.881
3         7.344         7.422         7.267               1.878
4         7.341         7.405         7.277               1.803

   Social support  Healthy life expectancy  Freedom to make life choices  \
0           1.572                    0.695                         0.859
1           1.520                    0.699                         0.823
2           1.617                    0.718                         0.819
3           1.501                    0.724                         0.838
4           1.513                    0.740                         0.641

   Generosity  Perceptions of corruption  Dystopia + residual
0       0.142                      0.546                2.082
1       0.204                      0.548                1.881
2       0.258                      0.182                2.050
3       0.221                      0.524                1.658
4       0.153                      0.193                2.298
```

```python
# Display basic information about the dataset
print("\nDataset Info:")
print(df.info())
```

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 143 entries, 0 to 142
Data columns (total 10 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   Ladder score                  143 non-null    float64
 1   upperwhisker                  143 non-null    float64
 2   lowerwhisker                  143 non-null    float64
 3   Log GDP per capita            140 non-null    float64
 4   Social support                140 non-null    float64
 5   Healthy life expectancy       140 non-null    float64
 6   Freedom to make life choices  140 non-null    float64
 7   Generosity                    140 non-null    float64
 8   Perceptions of corruption     140 non-null    float64
 9   Dystopia + residual           140 non-null    float64
dtypes: float64(10)
memory usage: 11.3 KB
None
```

```
# Display summary statistics
print("\nSummary Statistics:")
print(df.describe())
```

```
Summary Statistics:
       Ladder score  upperwhisker  lowerwhisker  Log GDP per capita  \
count    143.000000    143.000000    143.000000          140.000000
mean       5.527580      5.641175      5.413972            1.378807
std        1.170717      1.155008      1.187133            0.425098
min        1.721000      1.775000      1.667000            0.000000
25%        4.726000      4.845500      4.606000            1.077750
50%        5.785000      5.895000      5.674000            1.431500
75%        6.416000      6.507500      6.319000            1.741500
max        7.741000      7.815000      7.667000            2.141000

       Social support  Healthy life expectancy  Freedom to make life choices  \
count      140.000000               140.000000                    140.000000
mean         1.134329                 0.520886                      0.620621
std          0.333317                 0.164923                      0.162492
min          0.000000                 0.000000                      0.000000
25%          0.921750                 0.398000                      0.527500
50%          1.237500                 0.549500                      0.641000
75%          1.383250                 0.648500                      0.736000
max          1.617000                 0.857000                      0.863000

       Generosity  Perceptions of corruption  Dystopia + residual
count  140.000000                 140.000000           140.000000
mean     0.146271                   0.154121             1.575914
std      0.073441                   0.126238             0.537459
min      0.000000                   0.000000            -0.073000
25%      0.091000                   0.068750             1.308250
50%      0.136500                   0.120500             1.644500
75%      0.192500                   0.193750             1.881750
max      0.401000                   0.575000             2.998000
```

```
# Perform basic analysis (Example: Check for missing values)
print("\nMissing Values:")
print(df.isnull().sum())
```

```
Missing Values:
Ladder score                      0
upperwhisker                      0
lowerwhisker                      0
Log GDP per capita                3
Social support                    3
Healthy life expectancy           3
Freedom to make life choices      3
Generosity                        3
Perceptions of corruption         3
Dystopia + residual               3
dtype: int64
```