3.EDA-Data Cleaning

AIM:

- Handling missing values: detection, filling, and dropping
- Removing duplicates and unnecessary data
- Data type conversion and ensuring consistency
- Normalize data (e.g., standardization, min-max scaling).

PROGRAM:

```
import pandas as pd

df = pd.read_csv('/content/Iris.csv')  # Replace with your filename
df.head()

# Count missing values in each column
print(df.isnull().sum())

# Drop rows with any missing values
df = df.dropna()

# Drop columns with all missing values
df = df.dropna(axis=1, how='all')

# Check for duplicates
print(df.duplicated().sum())

# Remove duplicates
df = df.drop_duplicates()

# Check datatypes
print(df.dtypes)

from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
df[['SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthCm']]
= scaler.fit_transform(df[['SepalLengthCm', 'SepalWidthCm',
'PetalLengthCm', 'PetalWidthCm']])
from sklearn.preprocessing import StandardScaler
```

```
scaler = StandardScaler()
df[['SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthCm']]
= scaler.fit_transform(df[['SepalLengthCm', 'SepalWidthCm',
'PetalLengthCm', 'PetalWidthCm']])
from google.colab import files

df.to_csv('cleaned_data.csv', index=False)
files.download('cleaned_data.csv')
```

OUTPUT:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

```
Id              0
SepalLengthCm   0
SepalWidthCm    0
PetalLengthCm   0
PetalWidthCm    0
Species         0
dtype: int64
```

```
Id               int64
SepalLengthCm  float64
SepalWidthCm   float64
PetalLengthCm  float64
PetalWidthCm   float64
Species         object
dtype: object
```

RESULT:

Thus the program was written and executed successfully.