Practical No 1

Step 1

In [15]:
```
import pandas as pd
import numpy as np
```

Step 2

Description

The Iris dataset is a classic dataset in machine learning, consisting of 150 samples of iris flowers.

It contains four features: sepal length, sepal width, petal length, and petal width, all measured in centimeters.

These features are used to classify the flowers into three species: setosa, versicolor, and virginica.

The dataset is widely used for testing classification algorithms.

Source

https://www.kaggle.com/datasets/uciml/iris

Step 3

In [41]:
```
df = pd.read_csv('IRIS.csv')
```

In [42]:
```
df
```

Out[42]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	NaN	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	NaN	3.0	5.1	1.8	Iris-virginica

150 rows × 5 columns

Step 4

In [43]:
```
df.columns
```

Out[43]:
```
Index(['sepal_length', 'sepal_width', 'petal_length', 'petal_width',
       'species'],
      dtype='object')
```

In [46]:
```
df['species'].unique()
```

Out[46]:
```
array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

In [47]:
```
df['species'].value_counts()
```

Out[47]:
```
species
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
Name: count, dtype: int64
```

In [49]:
```
df.isnull().sum()
```

Out[49]:
```
sepal_length    11
sepal_width      2
petal_length    11
petal_width      0
species          0
dtype: int64
```

In [65]:
```
df['sepal_length'] = df['sepal_length'].fillna(df['sepal_length'].mean())
print(f"Mean of 'sepal_length': {df['sepal_length'].mean()}")
```

```
df['sepal_width'] = df['sepal_width'].fillna(df['sepal_width'].mean())
print(f"Mean of 'sepal_width': {df['sepal_width'].mean()}")

df['petal_length'] = df['petal_length'].fillna(df['petal_length'].mean())
print(f"Mean of 'petal_length': {df['petal_length'].mean()}")

df['petal_width'] = df['petal_width'].fillna(df['petal_width'].mean())
print(f"Mean of 'petal_width': {df['petal_width'].mean()}")
```

```
Mean of 'sepal_length': 5.848201438848921
Mean of 'sepal_width': 3.0486486486486486
Mean of 'petal_length': 3.776258992805756
Mean of 'petal_width': 1.1986666666666668
```

In [51]:
```
df.isnull().sum()
```

Out[51]:
```
sepal_length    0
sepal_width     0
petal_length    0
petal_width     0
species         0
dtype: int64
```

In [52]:
```
df.sample(5)
```

Out[52]:

	sepal_length	sepal_width	petal_length	petal_width	species
115	6.4	3.2	3.776259	2.3	Iris-virginica
79	5.7	2.6	3.500000	1.0	Iris-versicolor
11	4.8	3.4	1.600000	0.2	Iris-setosa
15	5.7	4.4	1.500000	0.4	Iris-setosa
81	5.5	2.4	3.700000	1.0	Iris-versicolor

In [63]:
```
df.describe(include='all')
```

Out[63]:

	sepal_length	sepal_width	petal_length	petal_width	species
count	150.000000	150.000000	150.000000	150.000000	150
unique	NaN	NaN	NaN	NaN	3
top	NaN	NaN	NaN	NaN	Iris-setosa
freq	NaN	NaN	NaN	NaN	50
mean	5.848201	3.048649	3.776259	1.198667	NaN
std	0.809536	0.427963	1.704143	0.763161	NaN
min	4.300000	2.000000	1.000000	0.100000	NaN
25%	5.100000	2.800000	1.600000	0.300000	NaN
50%	5.848201	3.000000	4.200000	1.300000	NaN
75%	6.400000	3.300000	5.100000	1.800000	NaN
max	7.900000	4.400000	6.900000	2.500000	NaN

In [64]:
```
df.dtypes
```

Out[64]:
```
sepal_length    float64
sepal_width     float64
petal_length    float64
petal_width     float64
species          object
dtype: object
```

In [38]:
```
df.shape
```

Out[38]:
```
(150, 5)
```

Step 5

In [58]:
```
df.dtypes
```

Out[58]:
```
sepal_length    float64
sepal_width     float64
petal_length    float64
petal_width     float64
species          object
dtype: object
```

In [59]:
```
# Summarize the variable types based on data types
print("\nSummary of Variables:")
for col in df.columns:
    if df[col].dtype == 'object':
        print(f"{col}: Character (String)")
    elif df[col].dtype == 'int64':
        print(f"{col}: Integer")
    elif df[col].dtype == 'float64':
        print(f"{col}: Numeric")
```

```
    elif df[col].dtype == 'bool':
        print(f"{col}: Logical (Boolean)")
    else:
        print(f"{col}: Unknown")
```

```
Summary of Variables:
sepal_length: Numeric
sepal_width: Numeric
petal_length: Numeric
petal_width: Numeric
species: Character (String)
```

Step 6

In [33]:
```
df1 = pd.read_csv('IRIS.csv')
```

In [34]:
```
df1.head()
```

Out[34]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

In [35]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```

In [36]:
```
df1['species'] = le.fit_transform(df1['species'])
```

In [37]:
```
df1.sample(5)
```

Out[37]:

	sepal_length	sepal_width	petal_length	petal_width	species
19	5.1	3.8	NaN	0.3	0
2	4.7	3.2	1.3	0.2	0
96	5.7	2.9	4.2	1.3	1
20	5.4	3.4	1.7	0.2	0
111	6.4	2.7	5.3	1.9	2