EXPERIMENT 4

Aim:

To perform data cleaning on a cricketer dataset by detecting and handling missing, duplicate, and inappropriate data using Python Pandas for data preprocessing.

Algorithm:

1. Import the required libraries – Pandas and NumPy.
2. Read the dataset using read_csv() and display it.
3. Check for duplicate records using duplicated().
4. Display the dataset information using info().
5. Remove duplicate rows using drop_duplicates().
6. Reset the DataFrame index using reset_index().
7. Replace invalid values (negative data) with NaN.
8. Identify inconsistent text data and correct it.
9. Fill missing data using mean or median values.
10. Display the final cleaned dataset.

Code:

```
import pandas as pd

import numpy as np

df = pd.read_csv("Cricketer_Imperfect.csv")

df
```

Output:

PlayerID	Name	Age	Country	Matches_Played	Runs	Wickets	Batting_Avg	Bowling_Avg	Role
1	Virat Kohli	35	India	254	12340	4	59.3	0.0	Batsman
2	Rashid Khan	25	Afghanistan	90	1300	230	18.2	21.4	Bowler
3	Ben Stokes	31	England	120	4500	170	42.5	32.1	All-Rounder

PlayerID	Name	Age	Country	Matches_Played	Runs	Wickets	Batting_Avg	Bowling_Avg	Role
4	Steve Smith	33	Australia	150	-500	0	61.7	0.0	Batsman
5	Jasprit Bumrah	29	India	67	50	108	12.5	24.3	Bowler
6	Kane Williamson	32	New Zealand	150	6700	0	55.2	0.0	Batsman
7	Shakib Al Hasan	34	Bangladesh	-10	6500	300	38.1	31.0	All-Rounder
8	Trent Boult	32	New Zealand	100	900	220	15.3	25.4	Bowler
9	AB de Villiers	38	South Africa	228	9577	2	53.5	0.0	Batsman
9	AB de Villiers	38	South Africa	228	9577	2	53.5	0.0	Batsman
10	Mitchell Starc	32	Australia	97	550	200	12.5	23.5	Bowler

---

df.duplicated()

Output:

0 False

1 False

2 False

3 False

4 False

5 False

6 False

7 False

8 False

9 True

10 False

dtype: bool

---

df.info()

Output:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 11 entries, 0 to 10

Data columns (total 10 columns):

#	Column	Non-Null Count	Dtype
0	PlayerID	11 non-null	int64
1	Name	11 non-null	object
2	Age	11 non-null	int64
3	Country	11 non-null	object
4	Matches_Played	11 non-null	int64
5	Runs	11 non-null	int64
6	Wickets	11 non-null	int64
7	Batting_Avg	11 non-null	float64
8	Bowling_Avg	11 non-null	float64
9	Role	11 non-null	object

dtypes: float64(2), int64(5), object(3)

memory usage: 1012.0+ bytes

---

df.drop_duplicates(inplace=True)

df

Output:

PlayerID	Name	Age	Country	Matches_Played	Runs	Wickets	Batting_Avg	Bowling_Avg	Role
1	Virat Kohli	35	India	254	12340	4	59.3	0.0	Batsman
2	Rashid Khan	25	Afghanistan	90	1300	230	18.2	21.4	Bowler

PlayerID	Name	Age	Country	Matches_Played	Runs	Wickets	Batting_Avg	Bowling_Avg	Role
3	Ben Stokes	31	England	120	4500	170	42.5	32.1	All-Rounder
4	Steve Smith	33	Australia	150	-500	0	61.7	0.0	Batsman
5	Jasprit Bumrah	29	India	67	50	108	12.5	24.3	Bowler
6	Kane Williamson	32	New Zealand	150	6700	0	55.2	0.0	Batsman
7	Shakib Al Hasan	34	Bangladesh	-10	6500	300	38.1	31.0	All-Rounder
8	Trent Boult	32	New Zealand	100	900	220	15.3	25.4	Bowler
9	AB de Villiers	38	South Africa	228	9577	2	53.5	0.0	Batsman
10	Mitchell Starc	32	Australia	97	550	200	12.5	23.5	Bowler

---

df.loc[df['Runs'] < 0, 'Runs'] = np.nan

df.loc[df['Matches_Played'] < 0, 'Matches_Played'] = np.nan

df

Output:

PlayerID	Name	Age	Country	Matches_Played	Runs	Wickets	Batting_Avg	Bowling_Avg	Role
1	Virat Kohli	35	India	254.0	12340.0	4.0	59.3	0.0	Batsman
2	Rashid Khan	25	Afghanistan	90.0	1300.0	230.0	18.2	21.4	Bowler
3	Ben Stokes	31	England	120.0	4500.0	170.0	42.5	32.1	All-Rounder

PlayerID	Name	Age	Country	Matches_Played	Runs	Wickets	Batting_Avg	Bowling_Avg	Role
4	Steve Smith	33	Australia	150.0	NaN	0.0	61.7	0.0	Batsman
5	Jasprit Bumrah	29	India	67.0	50.0	108.0	12.5	24.3	Bowler
6	Kane Williamson	32	New Zealand	150.0	6700.0	0.0	55.2	0.0	Batsman
7	Shakib Al Hasan	34	Bangladesh	NaN	6500.0	300.0	38.1	31.0	All-Rounder
8	Trent Boult	32	New Zealand	100.0	900.0	220.0	15.3	25.4	Bowler
9	AB de Villiers	38	South Africa	228.0	9577.0	2.0	53.5	0.0	Batsman
10	Mitchell Starc	32	Australia	97.0	550.0	200.0	12.5	23.5	Bowler

---

```
df['Runs'] = df['Runs'].fillna(round(df['Runs'].mean()))
df['Matches_Played'] = df['Matches_Played'].fillna(round(df['Matches_Played'].median()))
df
```

Output:

PlayerID	Name	Age	Country	Matches_Played	Runs	Wickets	Batting_Avg	Bowling_Avg	Role
1	Virat Kohli	35	India	254.0	12340.0	4.0	59.3	0.0	Batsman
2	Rashid Khan	25	Afghanistan	90.0	1300.0	230.0	18.2	21.4	Bowler
3	Ben Stokes	31	England	120.0	4500.0	170.0	42.5	32.1	All-Rounder
4	Steve Smith	33	Australia	150.0	4713.0	0.0	61.7	0.0	Batsman

PlayerID	Name	Age	Country	Matches_Played	Runs	Wickets	Batting_Avg	Bowling_Avg	Role
5	Jasprit Bumrah	29	India	67.0	50.0	108.0	12.5	24.3	Bowler
6	Kane Williamson	32	New Zealand	150.0	6700.0	0.0	55.2	0.0	Batsman
7	Shakib Al Hasan	34	Bangladesh	120.0	6500.0	300.0	38.1	31.0	All-Rounder
8	Trent Boult	32	New Zealand	100.0	900.0	220.0	15.3	25.4	Bowler
9	AB de Villiers	38	South Africa	228.0	9577.0	2.0	53.5	0.0	Batsman
10	Mitchell Starc	32	Australia	97.0	550.0	200.0	12.5	23.5	Bowler

Result:

Thus, the Python program to handle missing, duplicate, and inappropriate data using the Pandas library for data preprocessing was executed successfully and the output was verified.