CS23334 -Fundamental Of Data Science

240701463

Data Preprocessing (Missing Values and Encoding)

Exp :3A

Date: 29-07-2025

Aim:

To **understand the importance of Data Preprocessing** by performing two critical steps on a sample dataset: **imputing missing values** and **encoding categorical variables** into a format suitable for machine learning models.

Algorithm:

1. Load the dataset and **identify missing values** in 'Age' and 'Salary' and categorical columns ('Country', 'Purchased').
2. **Impute missing values** in the numerical 'Age' column using the **median**.
3. **Impute missing values** in the numerical 'Salary' column using the **mean**.
4. **Impute missing values** in the categorical 'Country' column using the **mode** (most frequent value).
5. Perform **One-Hot Encoding** on the 'Country' column using pd.get_dummies().
6. Perform **Label Encoding** on the binary 'Purchased' column (Yes=1, No=0).

Code:

```
import numpy as np

import pandas as pd

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

print("--- Initial Data ---\n", df)

df.info()

age_median = df.Age.median()

salary_mean = round(df.Salary.mean())
```

```
country_mode = df.Country.mode()[0]
df.Country.fillna(country_mode, inplace=True)
df.Age.fillna(age_median, inplace=True)
df.Salary.fillna(salary_mean, inplace=True)
print("\n--- Data After Missing Value Imputation (Execution Count 7) ---\n", df)
df.info()
country_dummies = pd.get_dummies(df.Country)
updated_dataset = pd.concat([country_dummies, df.iloc[:,[1,2,3]]], axis=1)
print("\n--- Data After One-Hot Encoding for Country (Execution Count 10) ---\n", updated_dataset)
updated_dataset.Purchased.replace(['No','Yes'],[0,1],inplace=True)
print("\n--- Final Preprocessed Data (Execution Count 15) ---\n", updated_dataset)
```

Output:

	Country	Age	Salary	Purchased
0	France	44.0	72000.0	No
1	Spain	27.0	48000.0	Yes
2	Germany	30.0	54000.0	No
3	Spain	38.0	61000.0	No
4	Germany	40.0	NaN	Yes
5	France	35.0	58000.0	Yes
6	Spain	NaN	52000.0	No
7	France	48.0	79000.0	Yes
8	Germany	50.0	83000.0	No
9	France	37.0	67000.0	Yes

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Country    10 non-null     object
 1   Age        9 non-null      float64
 2   Salary     9 non-null      float64
 3   Purchased  10 non-null     object
dtypes: float64(2), object(2)
memory usage: 448.0+ bytes
```

```
df.Salary.fillna(round(df.Salary.mean()),inplace=True)
```

	Country	Age	Salary	Purchased
0	France	44.0	72000.0	No
1	Spain	27.0	48000.0	Yes
2	Germany	30.0	54000.0	No
3	Spain	38.0	61000.0	No
4	Germany	40.0	63778.0	Yes
5	France	35.0	58000.0	Yes
6	Spain	38.0	52000.0	No
7	France	48.0	79000.0	Yes
8	Germany	50.0	83000.0	No
9	France	37.0	67000.0	Yes

	France	Germany	Spain
0	True	False	False
1	False	False	True
2	False	True	False
3	False	False	True
4	False	True	False
5	True	False	False
6	False	False	True
7	True	False	False
8	False	True	False
9	True	False	False

	France	Germany	Spain	Age	Salary	Purchased
0	True	False	False	44.0	72000.0	No
1	False	False	True	27.0	48000.0	Yes
2	False	True	False	30.0	54000.0	No
3	False	False	True	38.0	61000.0	No
4	False	True	False	40.0	63778.0	Yes
5	True	False	False	35.0	58000.0	Yes
6	False	False	True	38.0	52000.0	No
7	True	False	False	48.0	79000.0	Yes
8	False	True	False	50.0	83000.0	No
9	True	False	False	37.0	67000.0	Yes

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Country    10 non-null     object
 1   Age        10 non-null     float64
 2   Salary     10 non-null     float64
 3   Purchased  10 non-null     object
dtypes: float64(2), object(2)
memory usage: 448.0+ bytes
```

	France	Germany	Spain	Age	Salary	Purchased
0	True	False	False	44.0	72000.0	0
1	False	False	True	27.0	48000.0	1
2	False	True	False	30.0	54000.0	0
3	False	False	True	38.0	61000.0	0
4	False	True	False	40.0	63778.0	1
5	True	False	False	35.0	58000.0	1
6	False	False	True	38.0	52000.0	0
7	True	False	False	48.0	79000.0	1
8	False	True	False	50.0	83000.0	0
9	True	False	False	37.0	67000.0	1

Result:

The experiment successfully demonstrates the initial steps of data preprocessing. The importance of this phase is clear:

1. **Imputation** ensures that **missing data** does not cause errors or bias model training.
2. **Encoding** converts non-numeric **categorical data** (like country names) into a numerical format (like binary columns or 0/1 labels) that **machine learning algorithms** can process effectively.

The final updated_dataset is cleaned and fully digitized, making it ready for model training. Thus the python program was executed successfully, and the output is verified.