EXPERIMENT NO: 3A

Data Preprocessing and Encoding for Machine Learning

Aim:

To preprocess a dataset by handling missing values, encoding categorical variables, and preparing it for analysis or machine learning.

Algorithm:

1. **Load Data: Read CSV file into a DataFrame.**
2. **Handle Missing Values**
3. **Encode Categorical Data**
4. **Combine Data: Concatenate dummy variables with other relevant columns.**
5. **Output: Print the cleaned and encoded dataset.**

Program:

```
[3]: import numpy as np
     import pandas as pd
     df=pd.read_csv("C:/Users/vijay/Downloads/pre_process_datasample (1).csv")
     df.Country.fillna(df.Country.mode()[0],inplace=True)
     df.Age.fillna(df.Age.median(),inplace=True)
     df.Salary.fillna(round(df.Salary.mean()),inplace=True)
     country_dummies=pd.get_dummies(df.Country)
     updated_dataset=pd.concat([country_dummies,df.iloc[:,[1,2,3]]],axis=1)
     updated_dataset.Purchased.replace(['No','Yes'],[0,1],inplace=True)
     print(updated_dataset)
```

```
   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:

A cleaned and transformed dataset where all missing values are replaced, categorical variables are converted into numeric form, and the Purchased column is ready for modeling. Each row now has numeric values only, suitable for machine learning or statistical analysis.