Name: Shivam Pandey
Roll no:47
Experiment no.2
To Apply data Cleaning Techniques
Date of Performance: 13-02-2024
Date of Submission: 13-02-2024

Academic Year: 2023-24 **Semester: VIII**
Class / Branch: BE Computer **Subject: Applied Data Science Lab**

Experiment No. 2

1. Aim: To apply data cleaning techniques.

Dataset: In this experiment, a fictitious data containing 10 observations and 4 variables is used. The dataset contains Country, Age, Salary, and purchased columns. The dataset has categorical variables and missing values in these columns.

2. Software used: Google Colaboratory / Jupyter Notebook

3. Theory :-

Data cleaning is just the collective name to a series of actions we perform on our data in the process of getting it ready for analysis.

Some of the steps in data cleaning are:

• Handling missing values

• Encoding categorical features

• Outliers detection

• Transformations etc.

Handling missing values is a key part of data preprocessing and hence, it is of utmost importance for data scientists/machine learning engineers to learn different techniques in relation imputing / replacing numerical or categorical missing values with appropriate value based on appropriate strategies. That's primarily the reason we need to convert categorical columns to numerical columns so that a machine learning algorithm understands it. This process is called categorical encoding.

SimpleImputer is a class found in package sklearn.impute. It is used to impute / replace the numerical or categorical missing data related to one or more features with appropriate values.

Typically, any structured dataset includes multiple columns – a combination of numerical as well as categorical variables. A machine can only understand the numbers. It cannot understand the text. That's essentially the case with Machine Learning algorithms too. There are multiple ways of handling Categorical variables.
The two most widely used techniques:

• Label Encoding

• One-Hot Encoding

Label Encoding is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.

One-Hot Encoding is the process of creating dummy variables. It simply creates additional features based on the number of unique values in the categorical feature. Every unique value in the category will be added as a feature.

4. Program

CODE:

```
import pandas as pd

from sklearn.impute import SimpleImputer

from sklearn.preprocessing import LabelEncoder, OneHotEncoder

df=pd.read_csv('/content/CountryAgeSalary - CountryAgeSalary.csv')

print("Original Dataset:")

print(df)

print("\n")



# Data Cleaning Techniques



# Handling missing values

# Using SimpleImputer to replace missing values with the mean

imputer = SimpleImputer(strategy='mean')

df[['Age', 'Salary']] = imputer.fit_transform(df[['Age', 'Salary']])
```

```
# Encoding categorical features

# Using LabelEncoder for 'Country' and 'Purchased' columns

label_encoder = LabelEncoder()

df['Country'] = label_encoder.fit_transform(df['Country'])

df['Purchased'] = label_encoder.fit_transform(df['Purchased'])

# Display the cleaned DataFrame

print("Cleaned Dataset:")

print(df)
```

OUTPUT:

```
Original Dataset:
   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


Cleaned Dataset:
   Country        Age        Salary  Purchased
0        0  44.000000  72000.000000          0
1        2  27.000000  48000.000000          1
2        1  30.000000  54000.000000          0
3        2  38.000000  61000.000000          0
4        1  40.000000  63777.777778          1
5        0  35.000000  58000.000000          1
6        2  38.777778  52000.000000          0
7        0  48.000000  79000.000000          1
8        1  50.000000  83000.000000          0
9        0  37.000000  67000.000000          1
```

5.Conclusion :-

In this experiment, data cleaning techniques were applied to a fictitious dataset containing 10 observations and 4 variables: Country, Age, Salary, and Purchased. Missing values were handled using the mean strategy via SimpleImputer, ensuring data completeness. Categorical features were encoded using LabelEncoder, converting text-based categories into numerical values for better compatibility with machine learning algorithms. These preprocessing steps have rendered the dataset ready for further analysis or machine learning model development, ensuring improved data quality and interpretability.