Q1. What is data encoding? How is it useful in data science?

Encoding is the process of converting the data or a given sequence of characters, symbols, alphabets etc., into a specified format, for the secured transmission of data. Machine learning models can only work with numerical values. For this reason, it is necessary to transform the categorical values of the relevant features into numerical ones. This process is called feature encoding.

Q2. What is nominal encoding? Provide an example of how you would use it in a real-world scenario.

Nominal encoding is a technique used in machine learning and data analysis to convert categorical data into a numerical format. It is also known as one-hot encoding, where each category is assigned a unique numerical value. The purpose of this technique is to enable the machine learning algorithms to interpret the categorical data effectively.

We can use this encoding technique in various real-world scenarios such as text classification, customer segmentation, and image classification. For example, in text classification, we can convert the text data into numerical format using nominal encoding to classify the text into different categories such as sentiment analysis or topic classification.

Q3. In what situations is nominal encoding preferred over one-hot encoding? Provide a practical example.

Nominal encoding and one-hot encoding are essentially the same thing, and the terms are often used interchangeably. However, nominal encoding can be used to represent categorical variables as ordered or numerical values, whereas one-hot encoding always represents categorical variables as binary vectors.

There are some situations where nominal encoding may be preferred over one-hot encoding. One such situation is when the categorical variable has a large number of categories, and one-hot encoding would result in a high-dimensional feature space. In such cases, nominal encoding can be used to reduce the dimensionality of the feature space while still preserving the information in the data.

Q4. Suppose you have a dataset containing categorical data with 5 unique values. Which encoding technique would you use to transform this data into a format suitable for machine learning algorithms? Explain why you made this choice.

I will use OneHotEncoding because no. of unique values are 5 only . In general, one-hot encoding is the most commonly used method for nominal variables. It is simple to understand and implement, and it works well with most machine learning models.

Q5. In a machine learning project, you have a dataset with 1000 rows and 5 columns. Two of the columns are categorical, and the remaining three columns are numerical. If you were to use nominal encoding to transform the categorical data, how many new columns would be created? Show your calculations.

In [18]:

```
import seaborn as sns
import pandas as pd
from sklearn.preprocessing import OneHotEncoder
tips = sns.load_dataset("tips")
```

In [19]:

```
df = pd.DataFrame(tips.head(100))
```

In [20]:

```
df
```

Out[20]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
...	...	...	...	...	...	...	...
95	40.17	4.73	Male	Yes	Fri	Dinner	4
96	27.28	4.00	Male	Yes	Fri	Dinner	2
97	12.03	1.50	Male	Yes	Fri	Dinner	2
98	21.01	3.00	Male	Yes	Fri	Dinner	2
99	12.46	1.50	Male	No	Fri	Dinner	2

100 rows × 7 columns

In [21]:

```
encorder = OneHotEncoder()
encorder.fit_transform(df[['smoker','day']])
```

Out[21]:

```
<100x6 sparse matrix of type '<class 'numpy.float64'>'
        with 200 stored elements in Compressed Sparse Row format>
```

here 100 row and 6 category in colums so number of colums is created is 100 * 6 = 600

Q6. You are working with a dataset containing information about different types of animals, including their species, habitat, and diet. Which encoding technique would you use to transform the categorical data into a format suitable for machine learning algorithms? Justify your answer.

We use this categorical data encoding technique when the features are nominal(do not have any order). In one hot encoding, for each level of a categorical feature, we create a new variable. Each category is mapped with a binary variable containing either 0 or 1. Here, 0 represents the absence, and 1 represents the presence of that category.

These newly created binary features are known as Dummy variables. The number of dummy variables depends on the levels present in the categorical variable. This might sound complicated. Let us take an example to understand this better. Suppose we have a dataset with a category animal, having different animals like Dog, Cat, Sheep, Cow, Lion. Now we have to one-hot encode this data.

Q7.You are working on a project that involves predicting customer churn for a telecommunications company. You have a dataset with 5 features, including the customer's gender, age, contract type, monthly charges, and tenure. Which encoding technique(s) would you use to transform the categorical data into numerical data? Provide a step-by-step explanation of how you would implement the encoding.

OneHotEncoding will be used as there is only one categorical data (gender) and the rest are numerical. steps in encoding will be: The following step-by-step example shows how to perform one-hot encoding for this exact dataset in Python