What is Data Encoding:

**Data encoding** is the process of converting data from one form to another, usually for the purpose of transmission, storage, or analysis

Types of Encoding

1. Label Encoding: **Label Encoding** is a technique that is used to convert categorical columns into numerical ones so that they can be fitted by machine learning models which only take numerical data.

**Example Of Label Encoding**

Suppose we have a column*Height* in some dataset that has elements as Tall, Medium, and short. To convert this categorical column into a numerical column we will apply label encoding to this column.

|  |  |
| --- | --- |
| Height | Height |
| Tall | 0 |
| Medium | 1 |
| Short | 2 |

1. ONE HOT ENCODING: **One Hot Encoding** is a method for converting categorical variables into a binary format. It creates new binary columns (0s and 1s) for each category in the original variable. Each category in the original column is represented as a separate column, where a value of 1 indicates the presence of that category, and 0 indicates its absence.

| **Fruit** | **Categorical value of fruit** | **Price** |
| --- | --- | --- |
| apple | 1 | 5 |
| mango | 2 | 10 |
| apple | 1 | 15 |
| orange | 3 | 20 |

The output after applying one-hot encoding on the data is given as follows,

| **Fruit\_apple** | **Fruit\_mango** | **Fruit\_orange** | **price** |
| --- | --- | --- | --- |
| 1 | 0 | 0 | 5 |
| 0 | 1 | 0 | 10 |
| 1 | 0 | 0 | 15 |
| 0 | 0 | 1 | 20 |

1. Ordinal Encoding: Ordinal encoding is a technique used to convert categorical data into numerical values, which is essential for machine learning models that require numeric input. This method is particularly useful for categorical variables that have a natural order or ranking
2. Frequency/count encoding: [**Frequency encoding (or count encoding)** is a method used in machine learning to handle categorical data. It involves replacing each category with the count of how often it appears in the dataset](https://www.bing.com/ck/a?!&&p=997aeaa318eed99f520d1c7e21da10a3075a725a8380afcf154980bd4537f071JmltdHM9MTczMzI3MDQwMA&ptn=3&ver=2&hsh=4&fclid=3fa05d36-e562-6906-03fe-49b2e4b96831&psq=frequency+or+count+encoding&u=a1aHR0cHM6Ly9sZXRzZGF0YXNjaWVuY2UuY29tL2ZyZXF1ZW5jeS1lbmNvZGluZy8&ntb=1)**1**

Example of Frequency or count encoding:

Lets the categorical dataset of fruits is given

Fruit

Apple

Banana

Apple

Orange

Banana

Banana

Apple

Grapes

Orange

When we will apply frequency encoding then the the values of fruit column will be replaced by values of fruit\_count column.

Fruit Fruit\_Count

Apple 3

Banana 3

Apple 3

Orange 2

Banana 3

Banana 3

Apple 3

Grapes 1

Orange 2

WHY WE DO ENCODING:

1. Algorithm Compatibility: Most ML algorithms prefer numerical values over categorical values.
2. Efficiency and Performance: Numerical data are faster to work with and they take less storage
3. Feature representation: For example the word is meal , in sindhi it is mani, in urdu it is roti and in balochi it is tukr, but if we represent/encode meal with 1 we kill all these biases and it represents all features. This means that no problem which language you use but algorithm knows only 1.
4. Memory Usage: Due to replacing the categories with numbers the memory space is reduced.
5. Light data: By doing encoding, data becomes light.
6. Our computers feel easy to work with numerical values as compared to categorical values.
7. Reduces computation because if we Remove “Najaf” with 1 then it becomes easy and takes less space.
8. Less computational expenses
9. It takes more power to work with texts as compared to numbers.