MACHINE
LEARNING

Data Encoding - Nominal or OHE:

Colab:

https://colab.research.google.com/drive/1KoTlBwW__aZfcSgASOHQplGTN8VlAS9R#scrollTo=7hshG505BUKL

What is Data Encoding?

In machine learning, algorithms can only work with numbers. So if your dataset has categorical data (like colors, names, or labels) or textual data, you need to convert it into a numerical format. This process is called data encoding.

Types of Data

Before encoding, we need to know what type of data we're dealing with:

1. **Numerical Data** – Already numbers (e.g., age, salary, temperature).
    - Often scaled or normalized, but not "encoded" in the categorical sense.
2. **Categorical Data** – Non-numeric, can be:
    - **Nominal**: No inherent order (e.g., color: red, blue, green)
    - **Ordinal**: Has a clear order (e.g., education level: high school < bachelor < master < PhD)

Encoding Techniques

A. Label Encoding

- Converts each category into an integer.
- Example: Color → Red=0, Blue=1, Green=2
- Suitable for **ordinal data** because it preserves order.
- **Caution:** For nominal data, the algorithm might assume an order, which is **wrong**.

B. One-Hot Encoding

- Converts categories into binary vectors (0 or 1).
- Example: Color →

    Red → [1, 0, 0]

    Blue → [0, 1, 0]

    Green → [0, 0, 1]

- Best for **nominal data**.

- In Python, you can do this using pandas.get_dummies() or sklearn.OneHotEncoder.

C. Ordinal Encoding

- Assigns numbers based on order.
- Example: Education → High School=0, Bachelor=1, Master=2, PhD=3
- Use when the categories have a meaningful ranking.

D. Binary Encoding

- Converts categories into binary numbers (useful for many categories to reduce dimensionality compared to one-hot encoding).

- Example: 5 categories → need 3 bits:

0 → 000

1 → 001

2 → 010

3 → 011

4 → 100

E. Frequency or Count Encoding

- Replace each category with the number of times it appears.
- Example: Color → Red appears 5 times → 5, Blue 3 times → 3

4. When to Use Which Encoding

Data Type	**Technique**	**Notes**
Nominal	One-Hot, Binary	Avoid label encoding if order is meaningless
Ordinal	Label / Ordinal	Preserves order
High-cardinality	Frequency / Binary	Prevents too many dimensions