In Machine Learning (ML), **encoding** refers to the process of converting **categorical data (non-numeric data)** into a **numeric format** that machine learning algorithms can understand and process. Most ML algorithms require numerical input, so encoding is a crucial preprocessing step.

Types of Encoding in ML

1. Label Encoding

- Converts each category into a unique number.

✅Good for ordinal data (data with an order).

❌Not ideal for nominal data (no order), as it introduces false relationships.

One-Hot Encoding

- Creates a **binary column** for each category.

✅Best for nominal data.

- ❌Can increase dimensionality if there are many categories.

✅What is Dummy Encoding?

Dummy Encoding is essentially a form of **One-Hot Encoding**, but:

- It creates **binary columns** for **each category** in a feature.
- **Drops one column** to avoid the **dummy variable trap** (perfect multicollinearity in linear models).

Difference: One-Hot vs Dummy Encoding

Encoding Type	Creates All Columns	Drops One Column
One-Hot Encoding	Yes	No
Dummy Encoding	No	Yes

When to Use `drop_first=True`?

- When using **linear models** (like regression, logistic regression) to avoid **multicollinearity**.
- For **tree-based models** (e.g., Random Forest, XGBoost), it's okay to use full one-hot encoding.