# **Encoding**

Encoding is a technique of converting categorical variables into numerical values so that it could be easily fitted to a machine learning model.

Before getting into the details, let’s understand about the different types of categorical variables.  
  
**Nominal categorical variable:**

Nominal categorical variables are those for which we do not have to worry about the arrangement of the categories.

Example,

i. suppose we have a gender column with categories as Male and Female.  
ii. We can also have a state column in which we have different states like NY, FL, NV, TX  
So here we don’t have to worry about the arrangement of the categories.

## **Ordinal Categorical variable :**

Ordinal categories are those in which we have to worry about the rank. These categories can be rearranged based on ranks.

Example,

i. Suppose in a dataset there is an education column which we will use to predict the salary of the person. The education column has categories like ‘bachelors’,’masters’,’PHD’. Based on the above categories we can rearrange this and assign ranks to each category. Based on the education level ‘PHD’ will get the highest rank (PHD-1, masters-2, bachelors-3).

## **Types of Encoding techniques : Now that we have discussed about the type of categorical variables, let’s see the different types of encoding:**

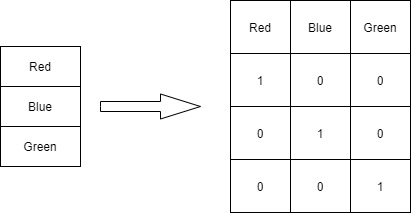
## **Nominal Encoding**

## **Ordinal Encoding**

## **1. One Hot Encoding**

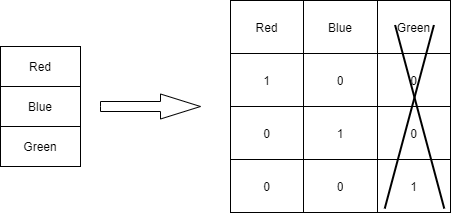
This method is applied to nominal categorical variables.

Example, suppose we have a column containing 3 categorical variables, then in one hot encoding 3 columns will be created each for a categorical variable.

One Hot Encoding

### **Dummy Variable Trap**

We can skip the last column ‘Green’ as 0,0 signifies green. This means, suppose we have ‘n’ columns, then the one hot encoding should create ‘n-1’ columns.

Dummy Variable Trap

### **Advantages of One-Hot Encoding**

1. **Model Compatibility**
2. **Simple Implementation**
3. **No Assumptions About Order**

### **Disadvantages of One-Hot Encoding**

1. **High Dimensionality**
2. **Scalability Issues**
3. **Redundancy**
4. **Lack of Ordinal Information**

## **2. Label Encoding**

## This technique will be used only for Ordinal categories. Ranks are provided based on the importance of the category. Below table illustrates that PHD is considered as the highest degree, so the highest label is given to it and so on.

## 

### **Advantages of Label Encoding**

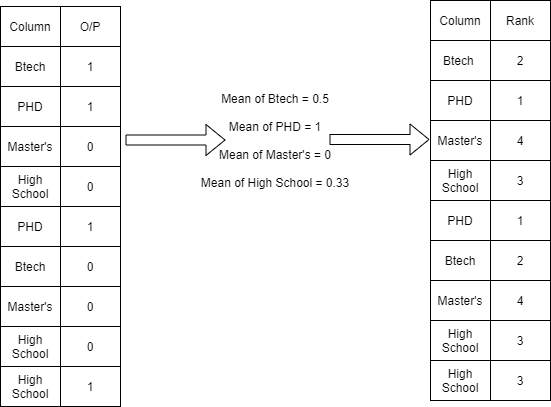
1. **Simplicity and Efficiency (lower memory usage)**
2. **Scalability**
3. **Compatibility**

### **Disadvantages of Label Encoding**

1. **Potential for Bias:**
   * **Unintended Relationships:** The numerical values assigned to categories can introduce bias into the model if the algorithm assumes a relationship that doesn't exist. For example, assigning 0, 1, 2 to categories A, B, C might suggest a relationship where B is somehow between A and C, which might not be true.

## **3. Target guided ordinal categories**

In this method, we calculate the mean of each categorical variable based on the output and then rank them. Below table illustrates this.

Target Encoding

We can apply this technique but cant do this with nominal as we dont know the order in case of nominal variables unlike in the case of Ordinal where we know the order of variables.

### **Advantages of Target-Guided Ordinal Encoding**

1. **Reduces Cardinality:**
2. **Improves Model Performance**

### **Disadvantages of Target-Guided Ordinal Encoding**

1. **Complexity and Computation**
2. **Bias Introduction**
3. **Dependence on Target Variable**

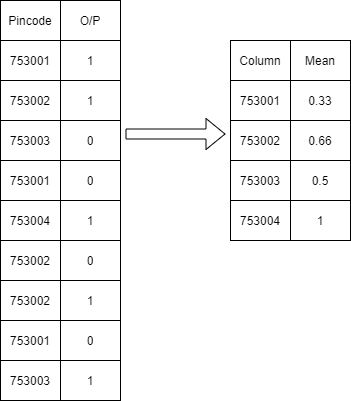
## **5. Mean Encoding**

In this method, we will convert the categories into their mean values based on the output.

This type of approach will be applicable where we have a lot of categorical variables for a particular column.

Example, suppose we have a column as pincode which contains all the pincodes of a city. It will contain many pincodes with multiple occurances. To encode we can use this technique which will convert all the pincodes into their mean values based on the output column.

Below table will illustrate the approach:



### **Advantages of Mean Encoding**

## **Reduces Dimensionality**

## **Improves Model Performance**

## **Maintains Interpretability**

### **Disadvantages of Mean Encoding**

## **Complexity and Computation**

## **Bias Introduction**

## **Dependence on Target Variable**

## 