# DIFFERENT TYPES OF 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 :

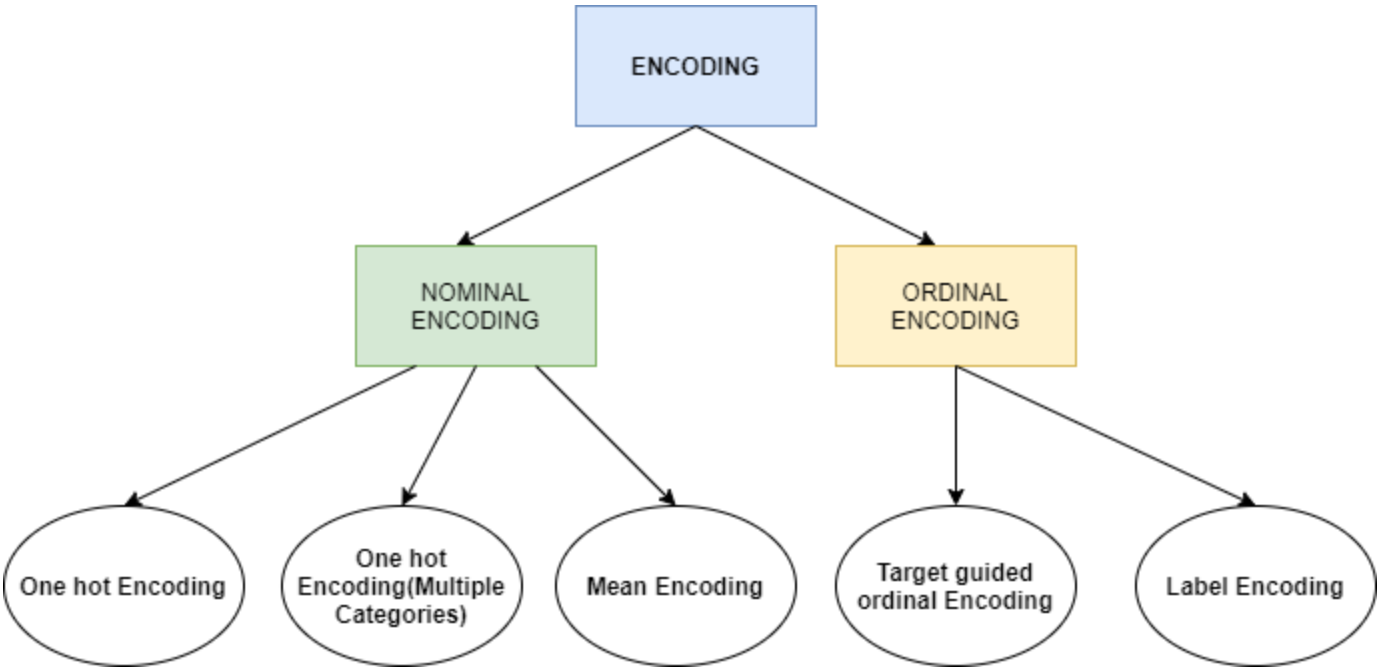
Ordinal categories are those in which we have to worry about the rank. These categories can be rearranged based on ranks.

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).

Now that we have discussed about the type of categorical variables, let's see the different types of encoding:

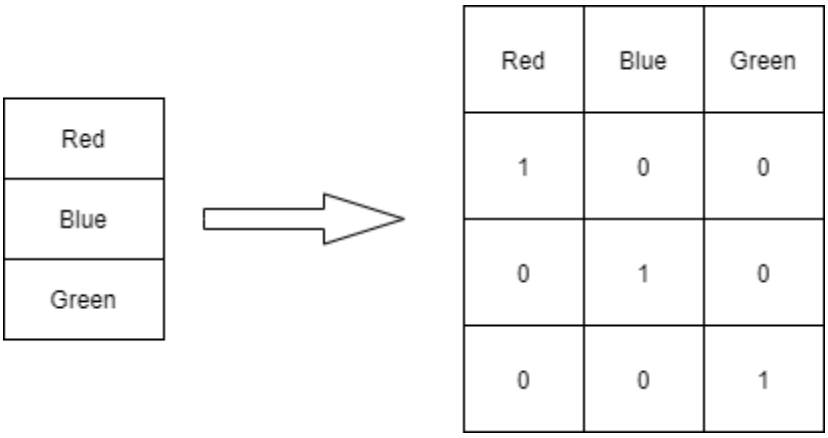
1. Nominal Encoding
2. Ordinal Encoding




# 1. ONE HOT ENCODING

This method is applied to nominal categorical variables.

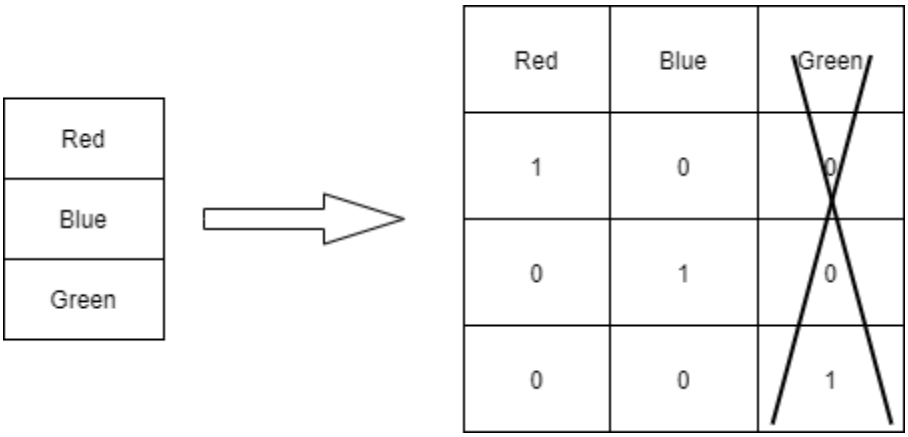
Example, suppose we have a column containing 3 categorical variables, then in one hot encoding 3 columns will be created each for a categorical variable.




**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**

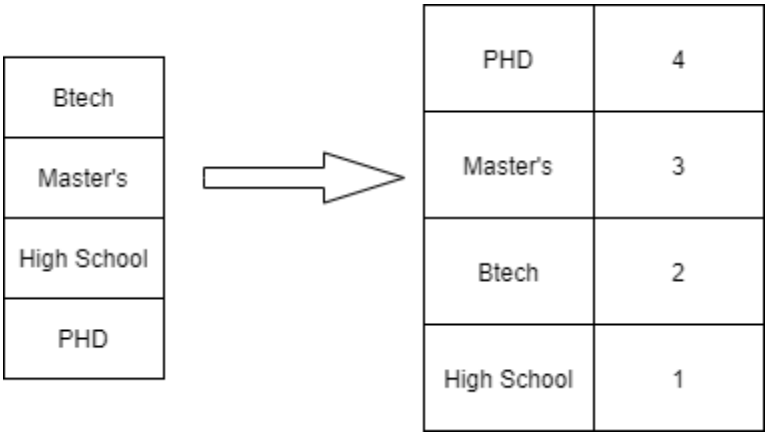
**DISADVANTAGE**

Suppose we have a column which has 100 categorical variables. Now if we try to convert the categorical variables into dummy variable then we will get 99 columns. This will increase the dimension of the overall dataset which will lead to curse of dimensionality.

So basically, if there is a lot of categorical variables in a column then we should not apply this technique.

# 2. LABEL ENCODING

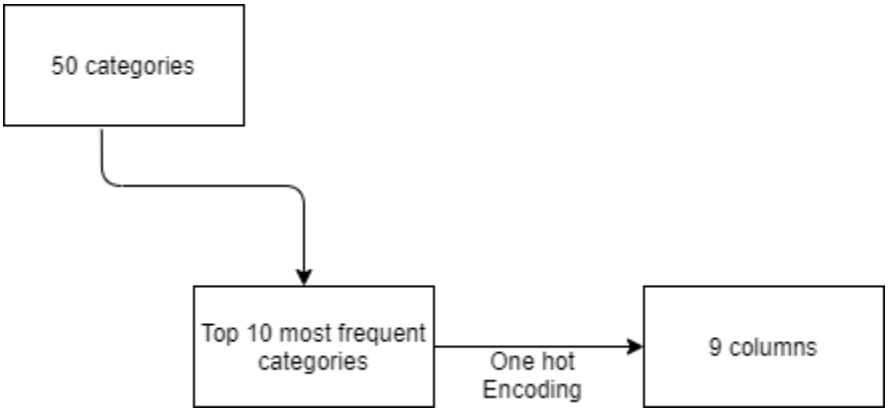
This technique will be used only for Ordinal categories. Ranks are provided based on the importance of the category. Below table illustrates that PHD is considered as the highest degree, so the highest label is given to it and so on.




**Label Encoding**

# 3. ONE HOT ENCODING (MULTIPLE CATEGORIES) — NOMINAL CATEGORIES

In this method, we will consider only those categories which has the most number of repetitions and we will consider the top 10 repeating categories and apply one hot-encoding to only those categories.




**One hot encoding-multiple categories**

The above technique was for Nominal variables. What shall we do if such kind of scenario arises for Ordinal variable. Let's see how to handle such scenario.

# 4. 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.

| Column | O/P |
|---|---|
| Btech | 1 |
| PHD | 1 |
| Master's | 0 |
| High School | 0 |
| PHD | 1 |
| Btech | 0 |
| Master's | 0 |
| High School | 0 |
| High School | 1 |

Mean of Btech = 0.5

Mean of PHD = 1

Mean of Master's = 0

Mean of High School = 0.33

| Column | Rank |
|---|---|
| Btech | 2 |
| PHD | 1 |
| Master's | 4 |
| High School | 3 |
| PHD | 1 |
| Btech | 2 |
| Master's | 4 |
| High School | 3 |
| High School | 3 |

**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.

To overcome this limitation for Nominal variables we use another technique called **Mean Encoding**

# 5. MEAN ENCODING

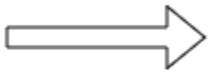
In this method, we will convert the categories into their mean values based on the output.

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:

| Pincode | O/P |
|---|---|
| 753001 | 1 |
| 753002 | 1 |
| 753003 | 0 |
| 753001 | 0 |
| 753004 | 1 |
| 753002 | 0 |
| 753002 | 1 |
| 753001 | 0 |
| 753003 | 1 |



| Column | Mean |
|---|---|
| 753001 | 0.33 |
| 753002 | 0.66 |
| 753003 | 0.5 |
| 753004 | 1 |