ML | Dummy variable trap in Regression Models

Before learning about the dummy variable trap, let's first understand what actually dummy variable is.

Dummy Variable in Regression Models:
In statistics, especially in regression models, we deal with various kinds of data. The data may be quantitative (numerical) or qualitative (categorical). The numerical data can be easily handled in regression models but we can't use categorical data directly, it needs to be transformed in some way.

For transforming categorical attributes to numerical attributes, we can use the label encoding procedure (label encoding assigns a unique integer to each category of data). But this procedure is not alone that suitable, hence, ***One hot encoding*** is used in regression models following label encoding. This enables us to create new attributes according to the number of classes present in the categorical attribute i.e if there are *n* number of categories in categorical attribute, *n* new attributes will be created. These attributes created are called ***Dummy Variables***. Hence, dummy variables are "proxy" variables for categorical data in regression models.
These dummy variables will be created with one-hot *encoding* and each attribute will have a value of either 0 or 1, representing the presence or absence of that attribute.

Dummy Variable Trap:
The Dummy variable trap is a scenario where there are attributes that are highly correlated (Multicollinear) and one variable predicts the value of others. When we use *one-hot encoding* for handling the categorical data, then one dummy variable (attribute) can be predicted with the help of other dummy variables. Hence, one dummy variable is highly correlated with other dummy variables. Using all dummy variables for regression models leads to a ***dummy variable trap***. So, the regression models should be designed to exclude one dummy variable.

For Example –
Let's consider the case of gender having two values *male* (0 or 1) and *female* (1 or 0). Including both the dummy variable can cause redundancy because if a person is not male in such case that person is a female, hence, we don't need to use both the variables in regression models. This will protect us from the dummy variable trap.