**Data Pre-processing**

While working on a data set these are the certain rules that should be taken care before undergoing the process

1. **Identifying and handling the missing values**

In data pre-processing, it is pivotal to identify and correctly handle the missing values, failing to do this, you might draw inaccurate and faulty conclusions and inferences from the data. This will hamper your ML project.

1. **Deleting a particular row**

 In this method, you remove a specific row that has a null value for a feature or a column where more than 75% of the values are missing. However, this method is not 100% efficient, and it is recommended that you use it only when the dataset has adequate samples. You must ensure that after deleting the data, there remains no addition of bias.

1. **Calculating the mean** – This method is useful for features having numeric data like age, salary, year, etc. Here, you can calculate the mean, median, or mode of a feature or column or row that contains a missing value and replace the result for the missing value. This method can add variance to the dataset, and any loss of data can be efficiently negated. Hence, it yields better results compared to the first method (omission of rows/columns). Another way of approximation is through the deviation of neighbouring values. However, this works best for linear data.

### ****Encoding the categorical data****

### **While processing a data algorithm’s can be applied only on numerical values i.e. no mathematical function can not be performed on the categorical values, so we use following 2 ways to overcome this problem.**

1. One hot encoding
2. Label encoding

* **One hot Encoding**

When the values of the categorical variable are nominal i.e. no relationship

In that case we make dummy variables.

**Before One-Hot Encoding**

|  |
| --- |
| **Country** |
| France |
| Germany |
| Spain |
| France |
| Spain |

**After One-Hot Encoding**

|  |  |  |
| --- | --- | --- |
| **France** | **Germany** | **Spain** |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 0 | 0 | 1 |

* **Label encoding**

The categorical variable in the data are encoded from string to integers that represent those categories.

**Before Label Encoding**

|  |
| --- |
| **Country** |
| France |
| Germany |
| Spain |
| France |
| Spain |

**After Label Encoding**

|  |
| --- |
| **Country** |
| 1 |
| 2 |
| 3 |
| 1 |
| 3 |

|  |  |
| --- | --- |
| **Country Dictionary** | |
| France | 1 |
| Germany | 2 |
| Spain | 3 |

1. **Feature Scaling**

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

1. **Standardization**

Xstand 
Standardisation 
x — mean(x) 
standard deviation (x) 

Data standardization is the process of rescaling one or more attributes so that they have a mean value of 0 and a standard deviation of 1.

Standardization assumes that your data has a Gaussian (bell curve) distribution. This does not strictly have to be true, but the technique is more effective if your attribute distribution is Gaussian.

1. **Normalisation**

Normalisation 
x 
norm 
max 
— min x 
(x) — min(x) 

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.