Preprocessing Steps in Machine Learning

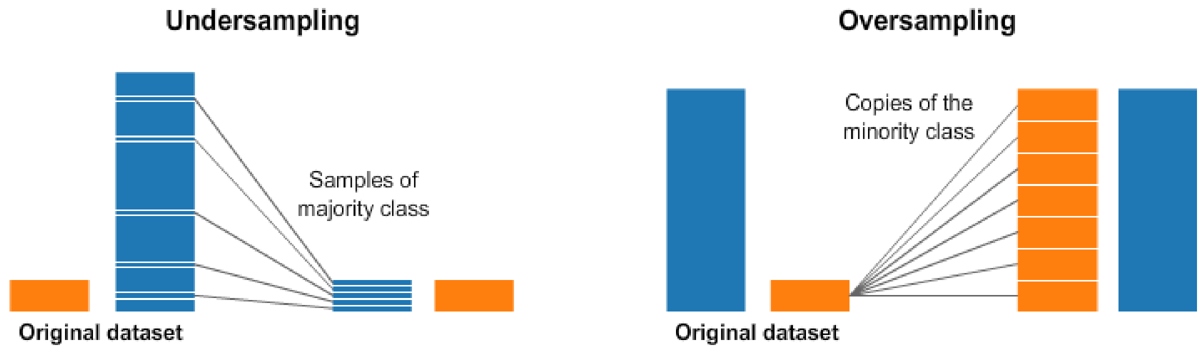
After gathering the data, we have to prepare it for the use in our machine learning training. We have to clean the data and transform it in a way that model can understand. To analyze the data, Exploratory Data Analysis should be done.

1. Duplicate Values:

Mostly we remove duplicate values so model would not give advantage or bias to that specific value.

2. Imbalanced Data:

An Imbalanced dataset is one where the number of instances of a class(es) are significantly higher than another class(es), thus leading to an imbalance and creating rarer class(es). We should use undersampling or oversampling methods.



3. Missing Values:

When analyzing dataset, we can see some values are missing. It is important to identify and handle missing values so that our model can learn better. There are some ways to handle missing values.

-Eliminate missing values: If our data is big enough and proportion of missing values are relatively small, we can consider dropping them.

-Filling with mean, mode or median: We can fill missing values with proper statistical metric. These metrics can be mean, mode or median. This method works for numeric data.

4. Outlier Detection:

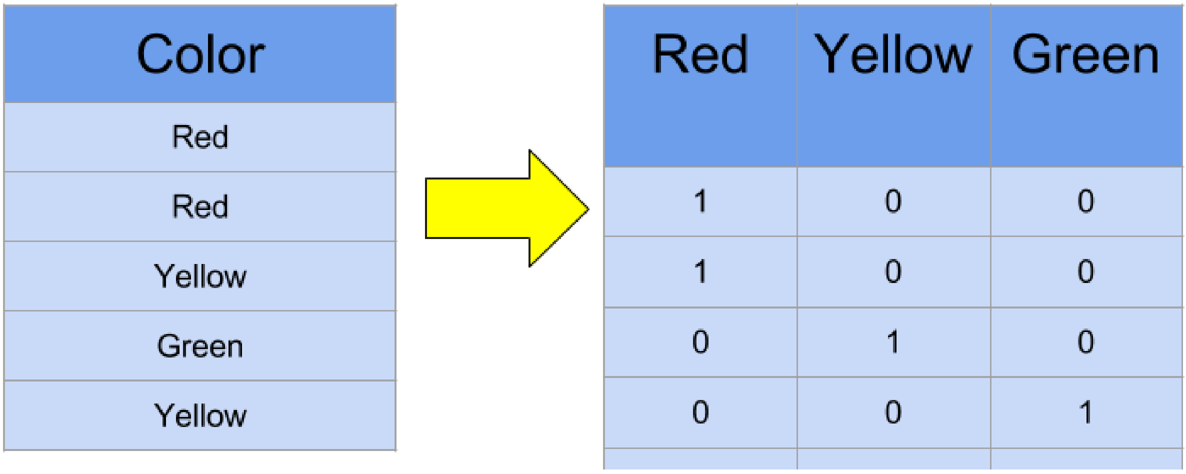
Outliers are extreme values that deviate from other observations on data. Outliers in a dataset may be occurred because of data entry errors, measurement errors etc. Removing them would be an option, but in some cases, we should not drop them because they may be not outliers, they may affect model performance. We can detect these outliers by using Standard Deviation, Box Plots/IQR Calculation, Isolation Forest…

5. Bucketing (Binning):

Data binning, bucketing is a data pre-processing method used to minimize the effects of small observation errors (noisy data). The original data values are divided into small intervals known as bins and then they are replaced by a general value calculated for that bin.

6. Feature Encoding:

Feature encoding is basically transforming data so it can be easily accepted and understood by machine learning model as input. We have to transform our categorical values to numerical values since machine learning models can only work with numerical values.



Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

-For Nominal data: One-Hot Encoding method can be used.

-For Ordinal data: Label encoding method can be used.

7. Feature Scaling:

It is a method to bring independent variables of a dataset within a specific range. Scaling limits the range of variables so we can compare them better.

Standardization: It transforms data to have a mean of zero and a standard deviation of 1.

$$X_{new} = \frac{X - \mu}{\sigma}$$

Normalization: It transforms data to have a values between 0 and 1.

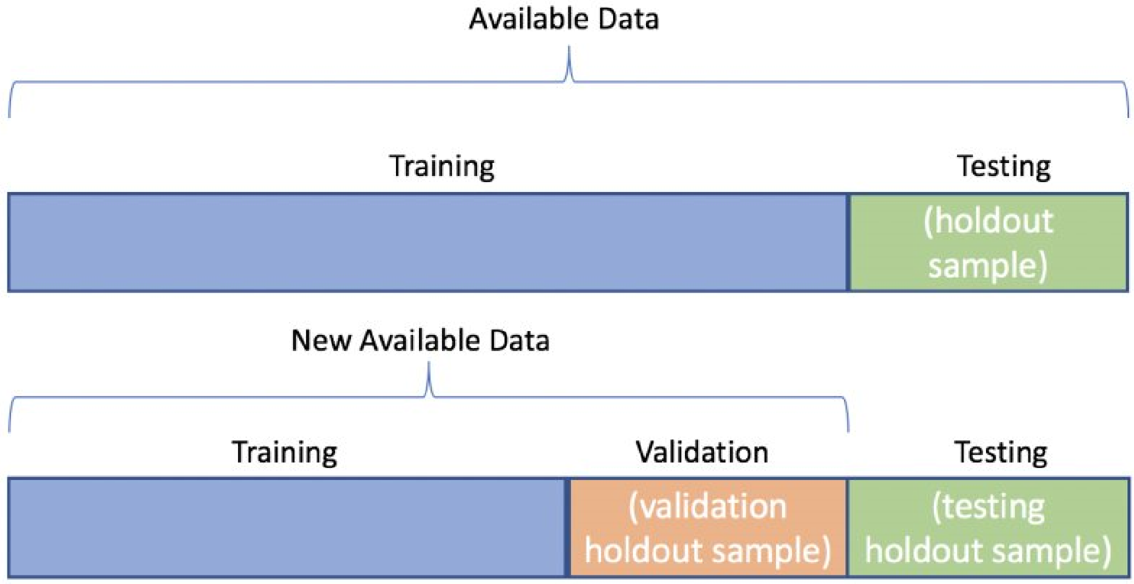
$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

8. Splitting Dataset:

We have to split our dataset into 2 (Train, Test) or 3 as (Train, Validation, Test) sets so we can fit, validate and test our data during machine learning process.

Common split ratios are:

-70/30 (Train/Test)
-60/20/20 (Train/Test/Validation)


Available Data
Training
Testing
(holdout sample)
New Available Data
Training
Validation
Testing
(validation holdout sample)
(testing holdout sample)