MODEL DATA -TRAINING SET

VALIDATION DATA-TESTING SET

Prepare Model & Validation Data

- Split the data into two -
  - Model / Training data : To build the model
  - Validation / Test data : To test the model

- Gives an estimate of the predictive model's performance

- Gives an insight on how the model will generalize to an independent dataset

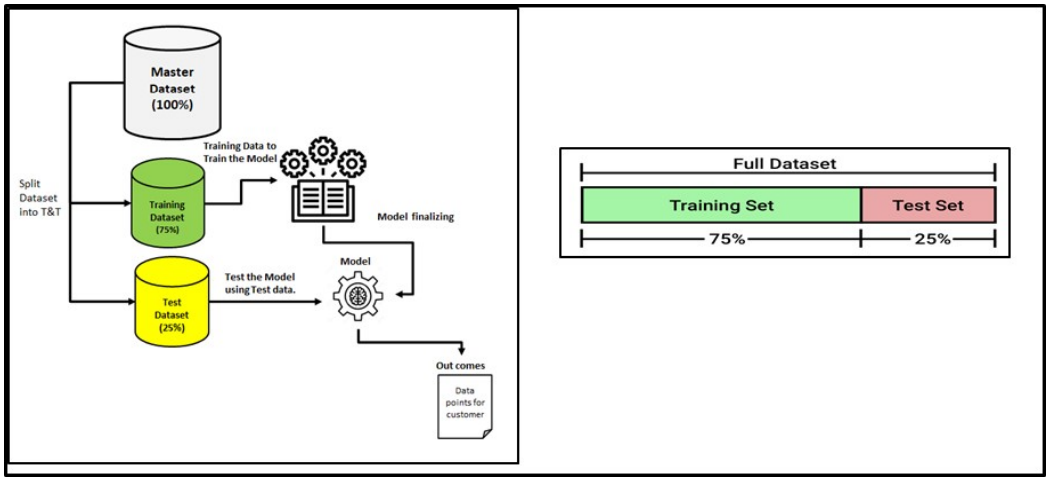
- Model data to Validation data ratio -
  70:30
  80:20

Training and Testing :

Training dataset - Is used to make sure the machine learns patterns of the data, applied to train or fit, your model. For example, you use the training set to find the optimal weights, or coefficients

Testing dataset - Is used to see how well the machine can predict new answers based on its training.

The train-test split procedure is used to estimate the ML performance of algorithms when they are used to make predictions on data that is not used to train the model, an unbiased evaluation of the final model.
You shouldn't use it for fitting or validation.



Data Split into Training/Testing Set

We split a dataset into training data and test data in the machine learning.

The split range is usually 70%-30% between training and testing stages from the given data set.

- A major amount of data would be spent on to train your model
- The rest of the amount can be spent to evaluate your test model.
- But you cannot mix/reuse the same data for both Train and Test purposes
- If you evaluate your model on the same training data, your model could be very overfitted. Then there is a question of whether models can predict new data.
- Therefore, you should have separate training and test subsets of your dataset.

Random State :

Random state ensures that the splits that you generate are reproducible. The random state that you provide is used as a seed to the random number generator.

This ensures that the random numbers are generated in the same order.