# Training, Test and Validation Sets in ML

This article is about description for those who need to know what is the actual difference between the dataset split between the **Training and Test sets** in Machine Learning while training and classifying models.

## What is the Training Data?

All the machine learning algorithms learn from data by finding relationships, developing understanding, making decisions, and building its confidence by using the training data we provide to a machine learning model. And this is to be noted that a machine learning model will perform based on what training data we have given to a model. Better the training data we will provide, better the model will perform.

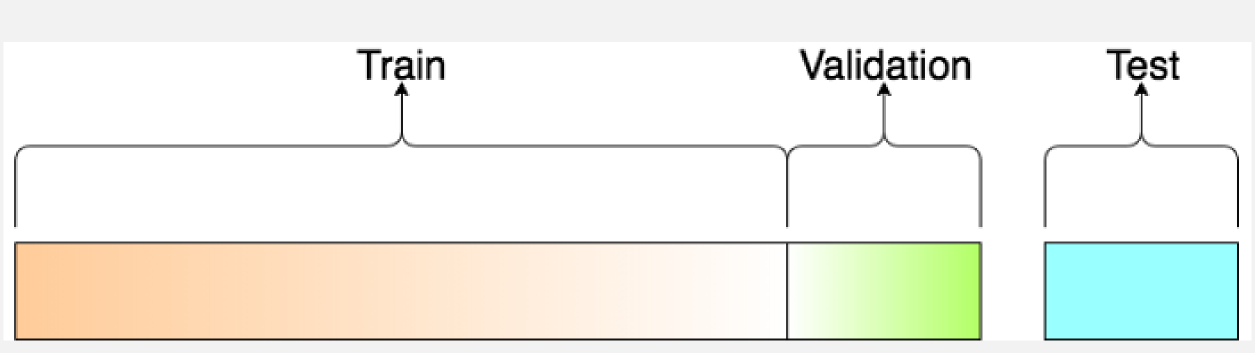
## What is Test Data?

Once a machine learning model is trained by using a training set, then the model is evaluated on a test set. The test data provides a brilliant opportunity for us to evaluate the model. The test set is only used once our machine learning model is trained correctly using the training set. Generally, a test set is only taken from the same dataset from where the training set has been received.

## Validation Set

Besides the Training and Test sets, there is another set which is known as a Validation Set. Validation Set is used to evaluate the model's hyperparameters. Our machine learning model will go through this data, but it will never learn anything from the validation set. A Data Scientist uses the results of a Validation set to update higher level hyperparameters.

## How to Split the Dataset into Training and Test sets



Generally, a dataset should be split into Training and Test sets with a ratio of 80 per cent Training set and 20 per cent Test set. This split of the Training and Test sets is ideal.

## When to use A Validation Set with Training and Test sets

Now, as we know, sometimes the data needs to be split into three rather than only training and test sets. So, the question arises when to use a Validation Set?

- Some models need substantial data to be trained with. However, in some cases models with very few hyperparameters will be easy to validate and prepare, in such instances we need to split the data into three sets. Still, the ratio of the validation set should be less if we have few hyperparameters.

- If our model has many hyperparameters, then obviously we need to increase the proportion of validation set.
- In some cases, when our model will not have any hyperparameter, in such cases, we will not need a Validation Set.

## What are Hyperparameters in Training and Test Sets?

A model has one or more parameters that determine what it will predict given a new instance. A machine learning algorithm tries to find optimal values for these parameters such that the model generalizes well to new cases. A hyperparameter is a parameter of the machine learning algorithm itself, not of the model.