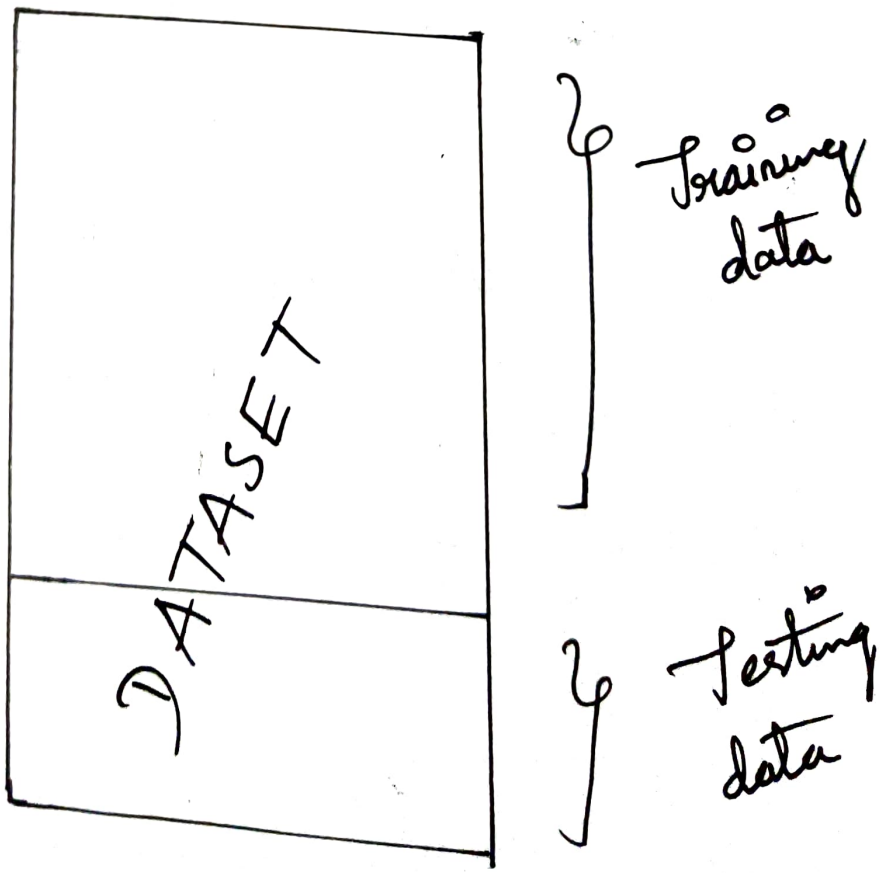
Before getting into the different model algorithm, we need to understand the most important question "Why do we need to split the data ?".

We create a predictive model to be able to guess the outcome for the unseen data.

In order to measure how a model performs on a new instance, we keep some part of data "unseen" by model.



Train data :-

→ The train data is used to train the model.

This will help the model to make predictions of it

Test data :-

→ The unseen data is used to test the model.

By using this data we can understand the learning of model.

We compare the train and test data for preventing the model from underfitting & overfitting.

70-30 % is the split percentage of test and train respectively (Generally they will consider this split percentage. But some situations makes them to change).

The training and testing split can be in Python by using Scikit-learn library.

(Refer Python Notebook for understanding sklearn).