**Introduction**:-

Why we split our data into train, validation, and test set?

Let’s understand it through the table:-

|  | **Purpose** | **Yield** | **Used for model training** | **Used for Parameter tuning** |
| --- | --- | --- | --- | --- |
| **Train Set** | To learn patterns from the data. | A model that makes near-expected predictions. | Yes | Yes |
| **Validation Set** | To understand the model behavior and generalizability on unseen data. | Insight on how to tune your model. | No | Yes |
| **Test Set** | To understand how the model would perform in the real world scenario. | A completely unbiased estimate of model performance. | No | No |

Notes:-

1. Never make the mistake of training your model on validation or test data.
2. If you tune your model after looking at the test accuracies, you are technically leaking information and hence cheating.

For more information on train, validation and test set split you can refer to the following links:-

1. <https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7>
2. <https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6>
3. <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>