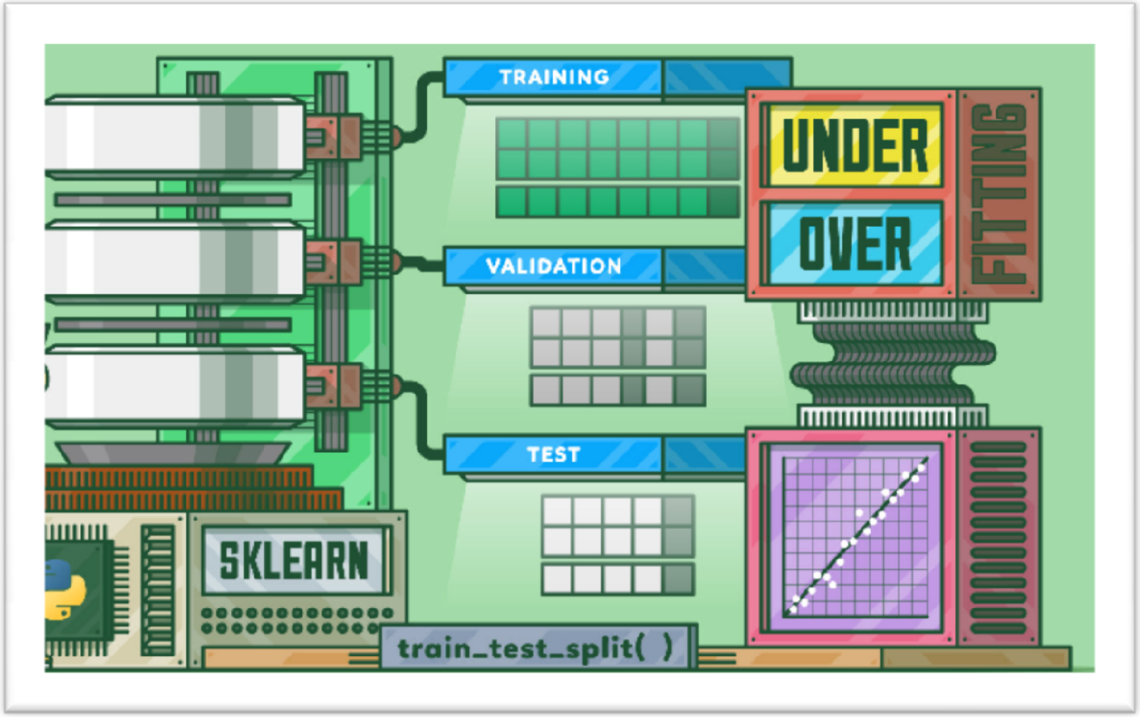
SK LEARN MODEL SELECTION

taroonreddy.com

Split Your Dataset With scikit-learn's train_test_split()



taroonreddy.com

The training set is applied to train, or **fit**, your model. For example, you use the training set to find the optimal weights, or coefficients

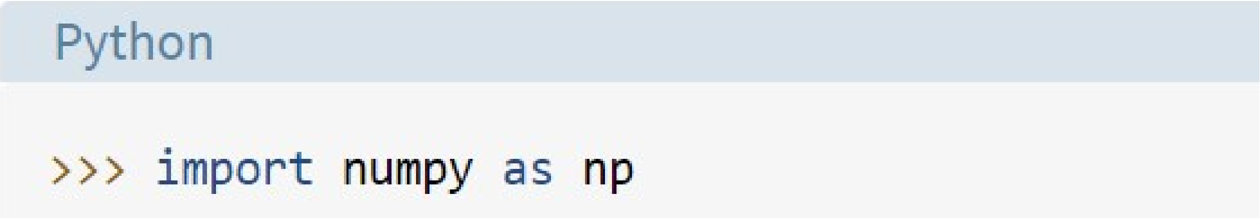
The validation set is used for unbiased model evaluation during hyper parameter tuning.

The test set is needed for an unbiased evaluation of the final model. You shouldn’t use it for fitting or validation.

In less complex cases, when you don’t have to tune hyper parameters, it’s okay to work with only the training and test sets.

Application of train_test_split()

You need to **import** **train_test_split()** and NumPy before you can use them, so you can start with the **import** statements:



```
>>> import numpy as np
```

train_test_split(), you need to provide the sequences that you want to split as well as any optional arguments.

arrays is the sequence of lists, NumPy arrays, pandas DataFrames, or similar array-like objects that hold the data you want to split. All these objects together make up the dataset and must be of the same length.

In supervised machine learning applications, you’ll typically work with two such sequences:
A two-dimensional array with the inputs (x)
A one-dimensional array with the outputs (y)