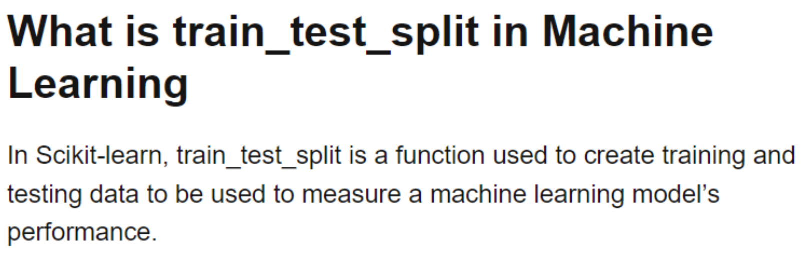
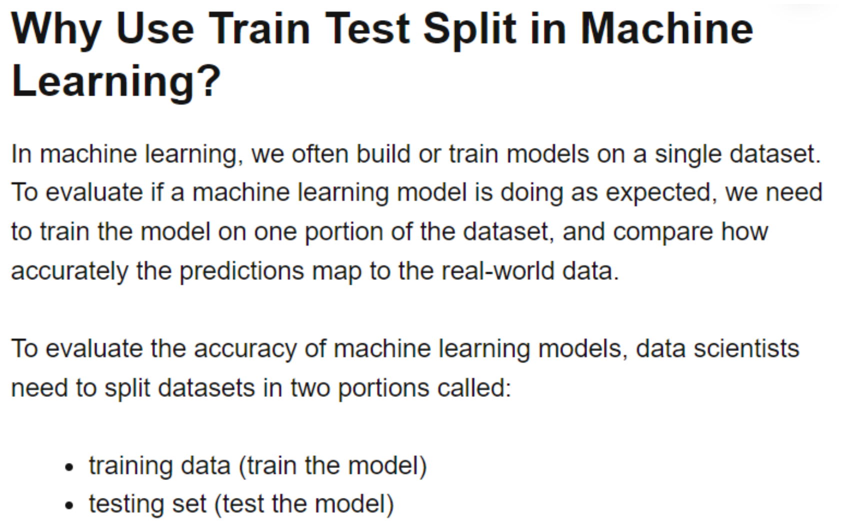
# Train Test Split in Python



## What is train_test_split in Machine Learning

In Scikit-learn, train_test_split is a function used to create training and testing data to be used to measure a machine learning model's performance.



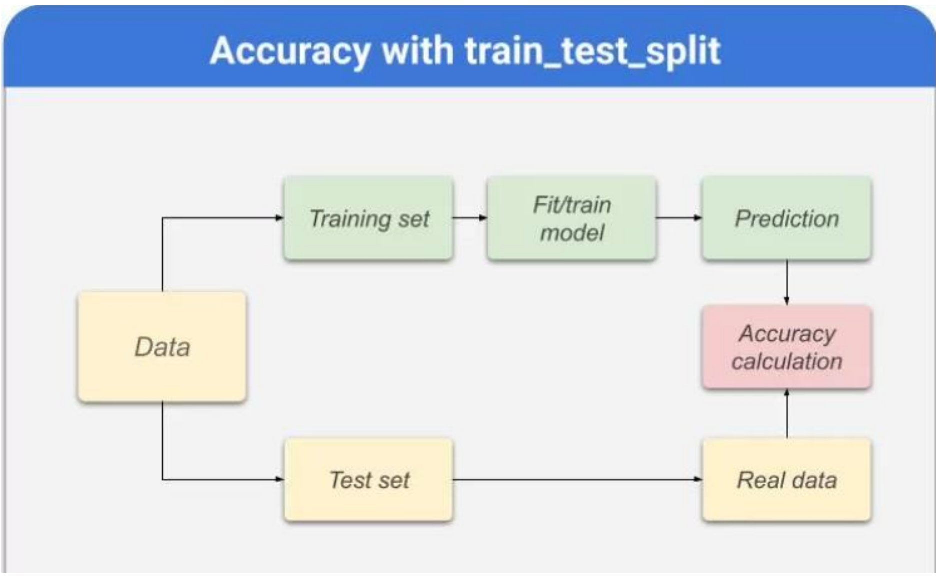
## Why Use Train Test Split in Machine Learning?

In machine learning, we often build or train models on a single dataset. To evaluate if a machine learning model is doing as expected, we need to train the model on one portion of the dataset, and compare how accurately the predictions map to the real-world data.

To evaluate the accuracy of machine learning models, data scientists need to split datasets in two portions called:

- training data (train the model)
- testing set (test the model)

## How Does Train Test Split Work?



## How to Use Train Test Split

1. **Split a dataset** into a training and testing set
2. **Provide the testing size** with the test_size parameter
3. **Train a model** on the training set
4. **Make predictions** on the training set
5. **Compute the accuracy** with a metrics such as the accuracy or accuracy_score

| **train_test_split Parameters** | **Description** | **Options/Values** | **Default** |
|---|---|---|---|
| test_size | Size of the testing subset | Float (0.0 to 1.0) or int | 0.25 |
| train_size | Size of the training subset | Float (0.0 to 1.0) or int | None |
| random_state | Random seed for reproducibility | int or RandomState instance | None |
| shuffle | Whether to shuffle the data before splitting | bool | True |
| stratify | Array-like or None. If not None, split data in a stratified fashion | array-like or None | None |