**Training Set:** The part of the dataset that is used for training the learning algorithm (model).

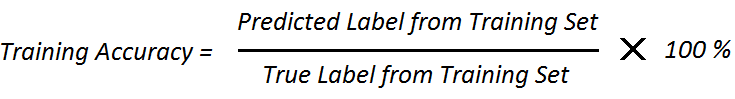
e.g.

If we have 70 samples in the Training set, we train the model by showing it the true labels for each sample. The model tries to learn the features for each sample and learn the output for the set of features.

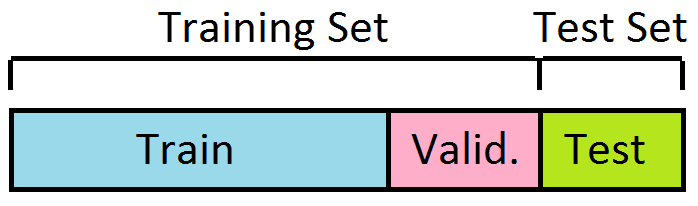
**Training Accuracy:** It is performance metric which measures how well the machine did when predicting samples from the training set. This should be easy for the machine since it is predicting from the same pool of data it was trained on. Therefore, it is usually a high number.

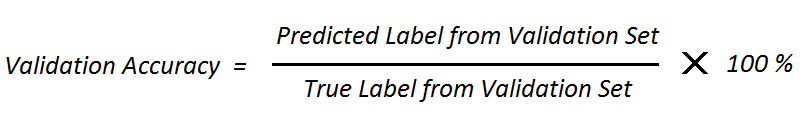
e.g.

For the 70 training samples, the machine is first trained as described above. Then, the machine is evaluated by letting it predict the output labels by using the same training samples.

**

**Validation Set:** The part of the Training Set that is held out from the model. The Validation Set is used to make an unbiased estimate of model performance by letting the model predict the output labels of previously unseen data.





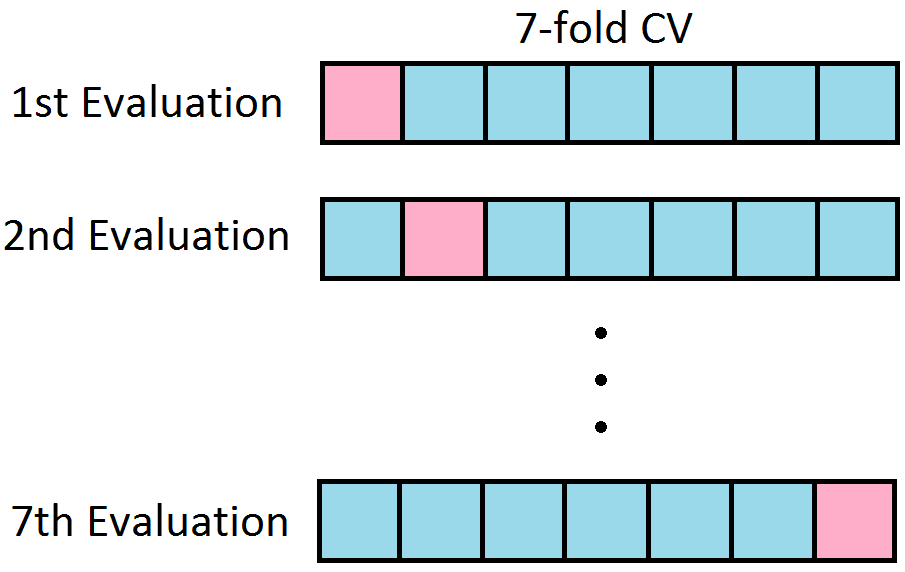
Once we get the Validation Accuracy, we can change the hyperparameter values and check again in hopes of increasing this accuracy value.

**K-fold Cross-Validation:**A validation technique where the Training Set is divided into K equal parts. Then, the model is evaluated K times, by selecting K-1 portion of the data for training and one part for validating. In each model evaluation, a different portion of the Training Set is used for validation.

The final Validation Accuracy is the average of all K evaluations.

e.g.

For the 70 training samples, if we perform a 7-fold Cross-Validation, we divide the Training Set into 7 equal parts. We evaluate the model 7 times, each time using a different portion for the Training Set (60 samples) and the Validation Set (10 samples), to get 7 Validation Accuracies. Our final Validation Accuracy will be an average of all 7 accuracies.



**Note:** In K-fold CV, for each model evaluation, the ratio Training Set : Validation Set = K-1 : 1.

**Test Set:** The part of the dataset that is used to get an unbiased evaluation of the model performance. This part is held out from model training, just like the Validation Set. Therefore, it is similar to the Validation Set, and is often used interchangeably.

However, the main difference between testing and validation is, after validation we try to improve the Validation Accuracy by changing the hyperparameter values. During testing, we assume the model is “ready” and is not tuned any further. So, the Test Accuracy reflects how the model will perform in the “real world” situation, to unseen data.

