Strive
SCHOOL

/ Validation Strategies

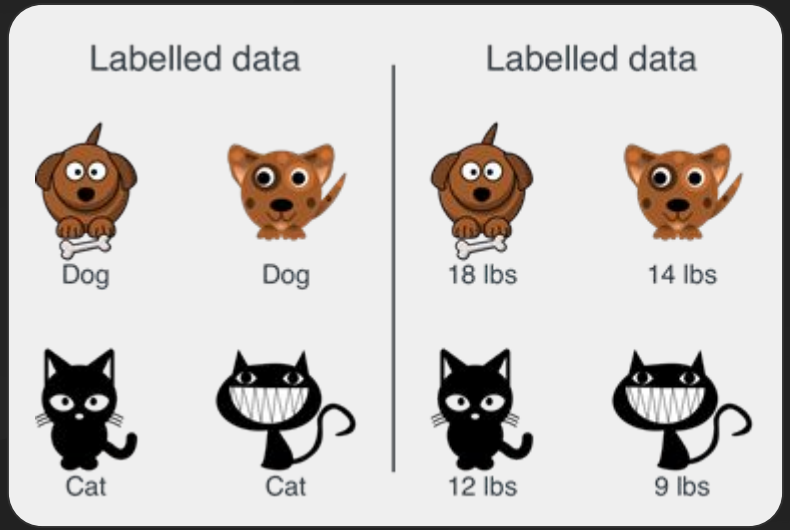
SimpleSplit, CrossValidation, Stratification,...

/ Validation

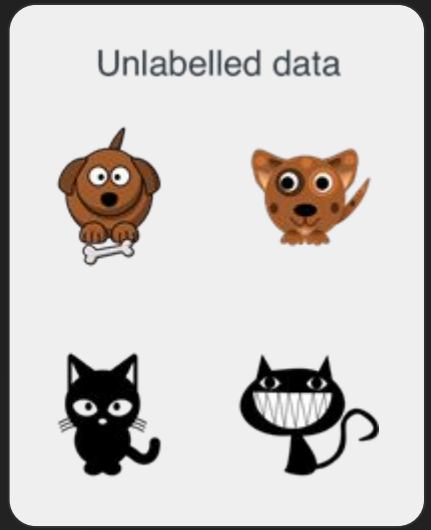
Is the process of testing that the model performs well on new unseen data. We should always validate our models.

Supervised and Unsupervised Data

Supervised data



Unsupervised data



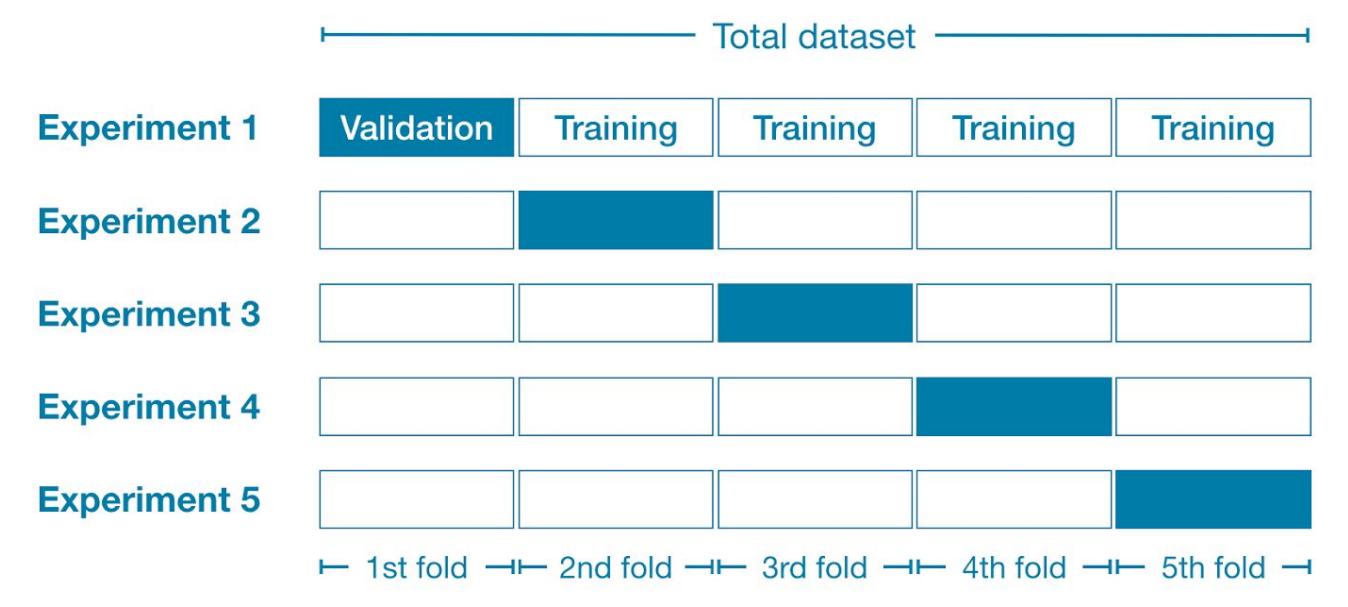
Simple Split Validation

- Dataset (usually supervised)
  - Train — The data that the model learns
  - Validation — Data to check that the model performs well
- New data (usually unsupervised)
  - Test — The new data to do inference

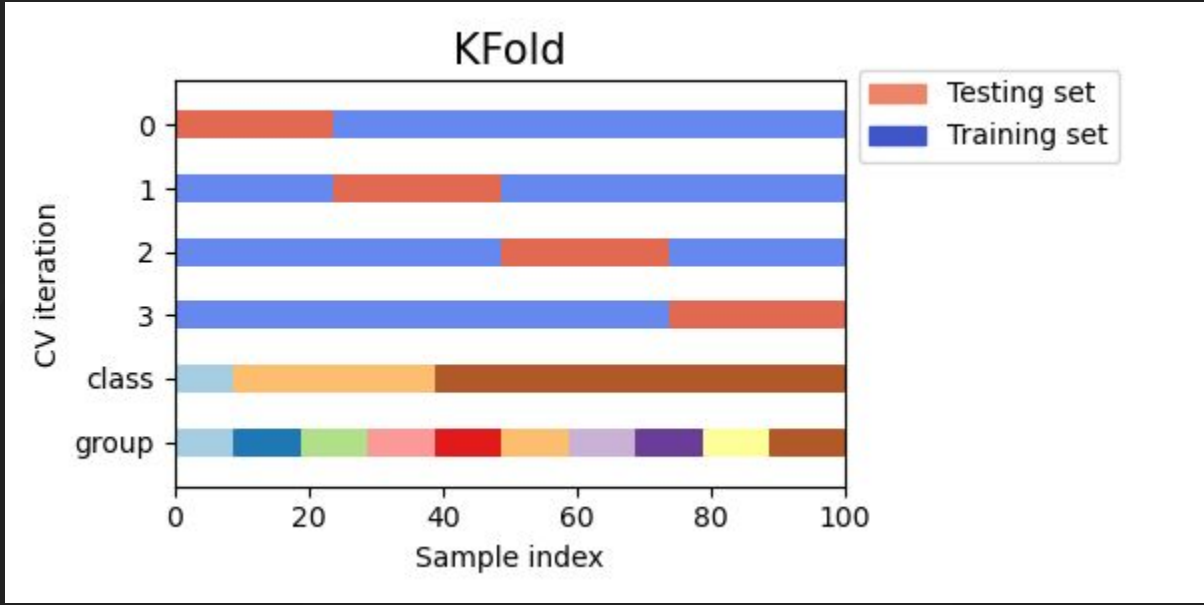
Cross Validation (aka K-Fold)

/ Cross Validation with 5 Folds (K=5). This validates more! (5 times)

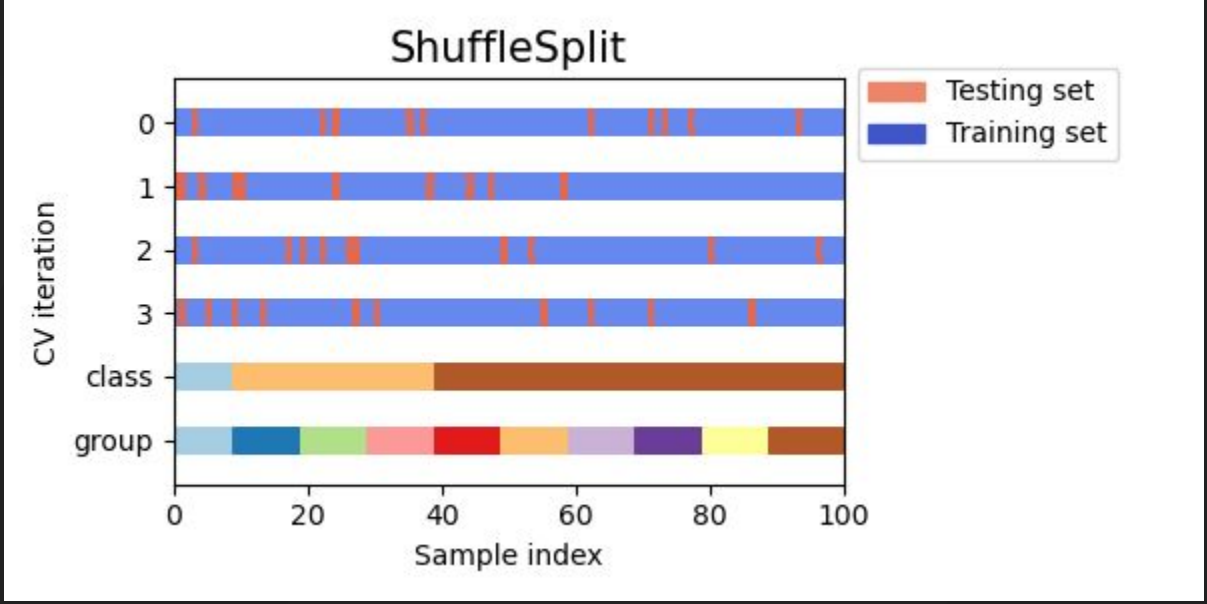
/ Remember: On each experiment a new model is trained from scratch.



Cross Validation

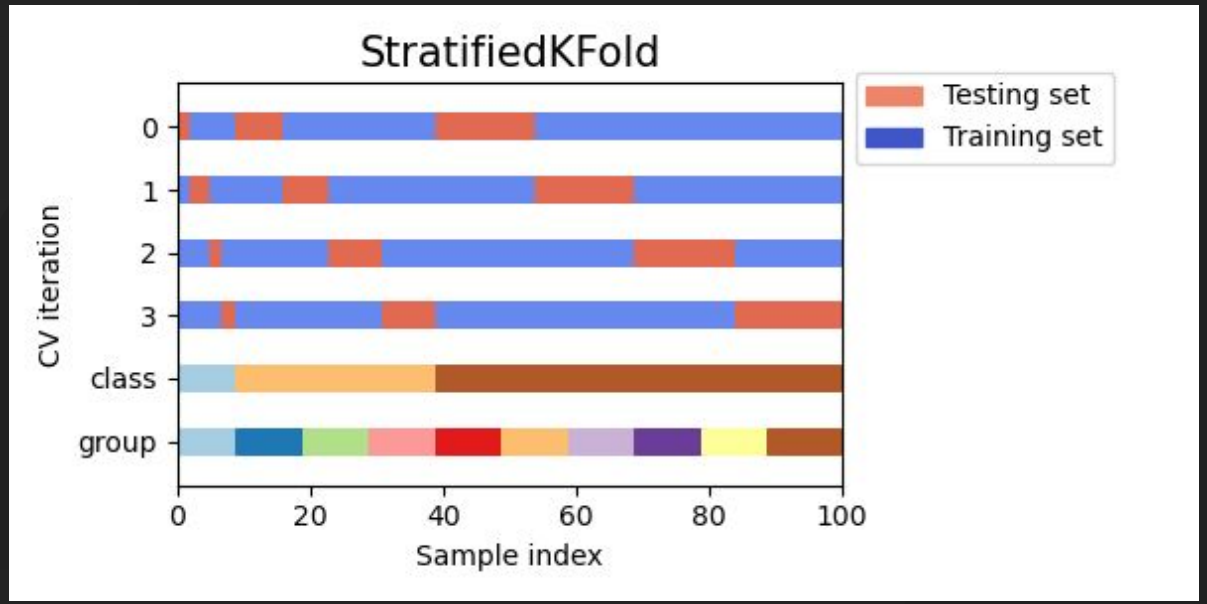


Cross Validation randomized



Cross Validation with Stratification

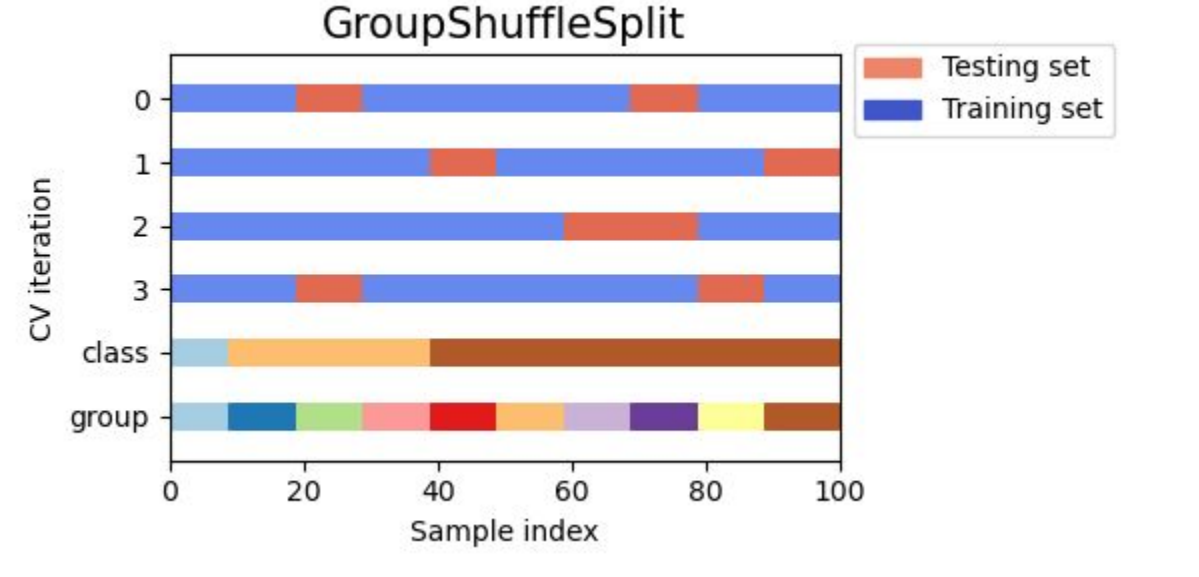
/ Each fold have the same class distribution. Very useful for classification!



Cross Validation with Groups

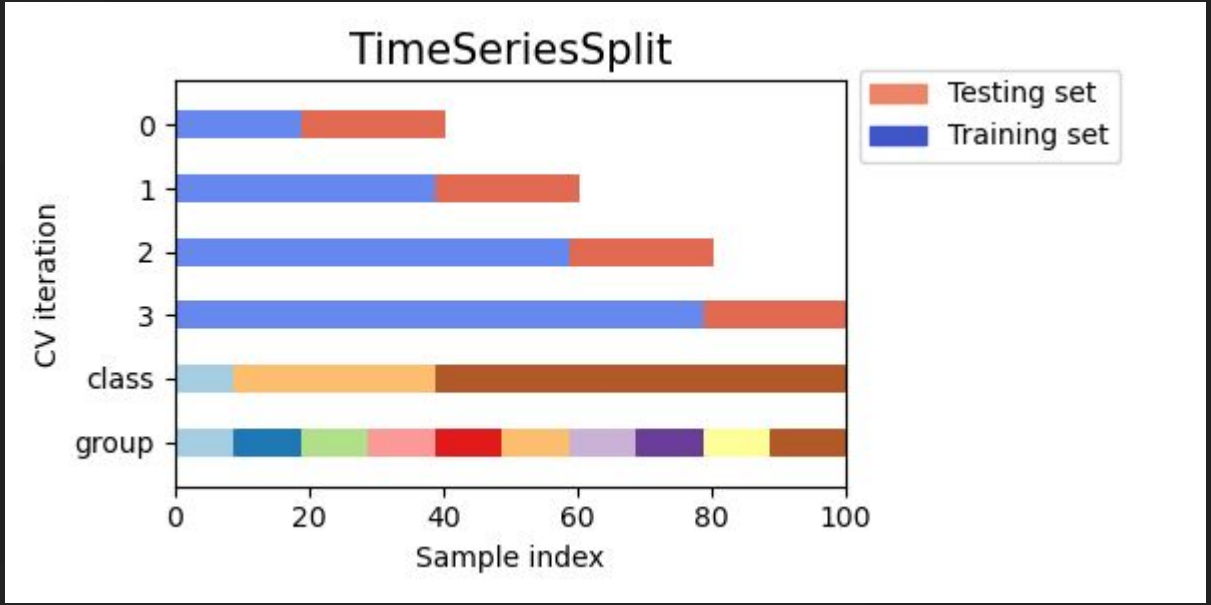
/ Useful when same samples are from the same individual.

E.g. patient have 3 records, and we want to validate on new patients.



Cross Validation for Time Series

/ Validate always on the future data



Sklearn validation

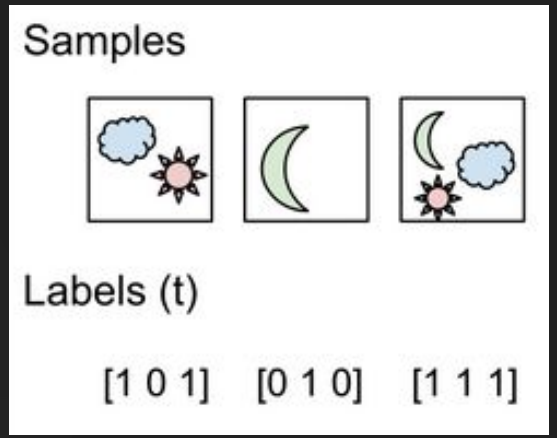
/ Many options to choose

`model_selection.GroupKFold`([n_splits])	K-fold iterator variant with non-overlapping groups.
`model_selection.GroupShuffleSplit`([...])	Shuffle-Group(s)-Out cross-validation iterator
`model_selection.KFold`([n_splits, shuffle, ...])	K-Folds cross-validator
`model_selection.LeaveOneGroupOut`()	Leave One Group Out cross-validator
`model_selection.LeavePGroupsOut`(n_groups)	Leave P Group(s) Out cross-validator
`model_selection.LeaveOneOut`()	Leave-One-Out cross-validator
`model_selection.LeavePOut`(p)	Leave-P-Out cross-validator
`model_selection.PredefinedSplit`(test_fold)	Predefined split cross-validator
`model_selection.RepeatedKFold`(*[, n_splits, ...])	Repeated K-Fold cross validator.
`model_selection.RepeatedStratifiedKFold`(*[, ...])	Repeated Stratified K-Fold cross validator.
`model_selection.ShuffleSplit`([n_splits, ...])	Random permutation cross-validator
`model_selection.StratifiedKFold`([n_splits, ...])	**Stratified K-Folds cross-validator**.
`model_selection.StratifiedShuffleSplit`([...])	Stratified ShuffleSplit cross-validator
`model_selection.TimeSeriesSplit`([n_splits, ...])	Time Series cross-validator

Validation for multilabel classification

/ iterative-stratification is a project that provides scikit-learn compatible cross validators with stratification for multilabel data.

- MultilabelStratifiedKFold
- MultilabelRepeatedStratifiedKFold
- MultilabelStratifiedShuffleSplit



MultiLabel dataset

/ Q&A

What are your doubts?

Strive
SCHOOL