

Train / Validation / Test Split - 101

Train / Validation / Test Split - 101

Training Data



The diagram illustrates a data split into three categories: Training Data, Validation Data, and Test Data. The Training Data is represented by a blue rectangle, the Validation Data by a green rectangle, and the Test Data by a red rectangle. A bracket is drawn under the Training Data rectangle, pointing to a text box below.

Training Data

Validation Data

Test Data

- Data sample used to fit the model

Train / Validation / Test Split - 101

Validation Data



- Data sample used to evaluate the model
- Used to fine-tune model hyperparameters
- Model occasionally „sees“ the data, but does not learn from it
- Affects the model indirectly

Train / Validation / Test Split - 101

Test Data



- Data sample used to provide an unbiased evaluation of final model
- Provides gold-standard
- Model “sees” data only once
- used to evaluate competing models
- Same distribution as validation data
- Often only training and validation data is used, and no test data.

Train / Validation / Test Split - 101

Split Ratio

- Depends on two things: total number of samples, actual model
- Some models require more training data
- Validation data big enough to detect differences between models
- Models with few hyperparameters will be easy to validate → smaller validation dataset
- Models with many hyperparameters will be harder to validate → larger validation dataset

Train / Validation / Test Split - 101

Interactive

Live Shiny-app!

Train / Test / Validation Split

Data Count

Train Ratio:

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

0.8

Test / Holdout Ratio:

0 0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16 0.18 0.2

0.1

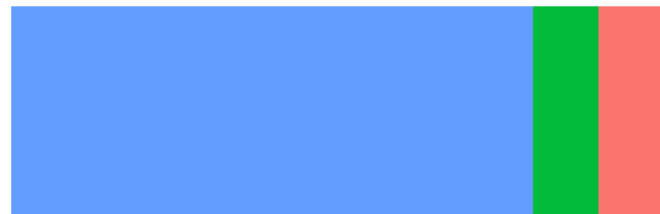
Seed

Selection Type

☒ Linear

☐ Random

Train / Test / Holdout Split



0 250 500 750 1000

Type	Ratio [%]	Data Count [-]
train	80.00	800.00
test	10.00	100.00
holdout	10.00	100.00