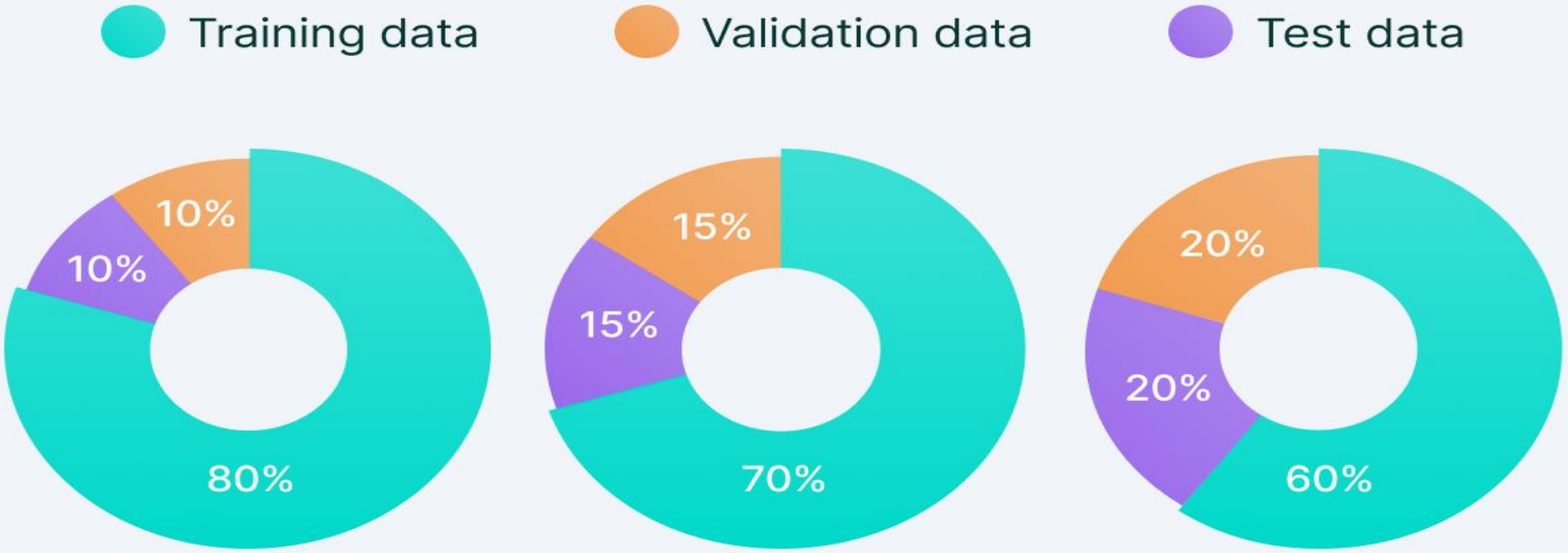
Normal train_test_split from Scikit Learn



Problem With train_test_split

1. What if the split we make isn't random?
2. What if one subset of our data has only people from a certain state, employees with a certain income level but not other income levels, only women or only people at a certain age?

This will result in overfitting, even though we're trying to avoid it!

❑ This is where cross validation comes in.

" The above is most of the blogs mentioned about which I don't understand that.

I think the disadvantages is not overfitting but underfitting. When we split the data , assume State A and B become the training dataset and try to predict the State C which is completely different than the training data that will lead to underfitting. Can someone fill me in why most of the blogs state 'test-split' lead to overfitting.

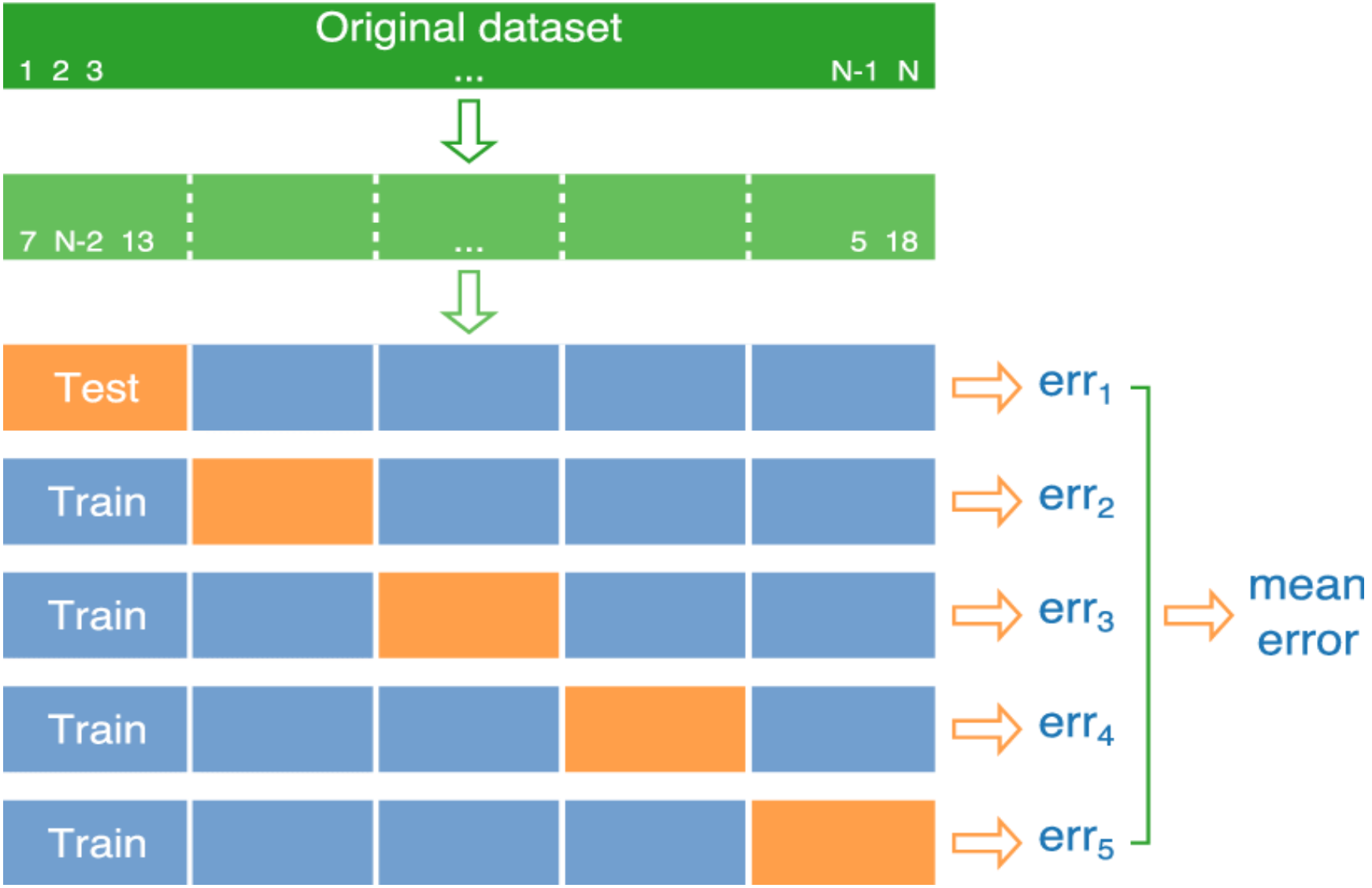
KFold Cross-Validation

- ✓ K-fold cross-validation is a superior technique to validate the performance of our model.
- ✓ It evaluates the model using different chunks of the data set as the validation set.
- ✓ We divide our data set into K-folds.
- ✓ K represents the number of folds into which you want to split your data.
- ✓ If we use 5-folds, the data set divides into five sections.
- ✓ In different iterations, one part becomes the validation set.

How it could be done?

1. Randomly divide a dataset into *k* groups, or “folds”, of roughly equal size.
2. Choose one of the folds to be the holdout set. Fit the model on the remaining k-1 folds. Calculate the test MSE on the observations in the fold that was held out.
3. Repeat this process *k* times, using a different set each time as the holdout set.
4. Calculate the overall test MSE to be the average of the *k* test MSE’s.

KFold Cross-Validation



Advantages

- We end up using all the data for training and testing and this is very useful in case of small datasets.
- It covers the variation of input data by validating the performance of the model on multiple folds.
- Multiple folds also helps in case of unbalanced data.
- Model performance analysis for every fold gives us more insights to fine tune the model.
- Used for hyper parameter tuning.