

# Split it wrong, and every score after that is fiction.

D8A.ACADEMY/RESOURCES/TRAIN-TEST-SPLIT

## 01 Three sets, not two

Tuning on the test set turns it into a second training set.



→ Every time you change a hyperparameter after reading the test score, that score stops being an estimate of anything.

## 02 Split first. Always.

Fit the scaler on train only, then apply it. A Pipeline makes the leak impossible.

```
from sklearn.pipeline import Pipeline

# 1. split BEFORE anything touches the data
X_tr, X_te, y_tr, y_te = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42)

# 2. the scaler learns from train rows only
pipe = Pipeline([("scale", StandardScaler()),
                  ("model", LogisticRegression())])
pipe.fit(X_tr, y_tr)
```

→ Scale before the split and the test mean leaks into training. The score goes up and the model gets worse.

## 03 Pick the right split

Random is the default. Your data is usually not.

|            |                                                             |
|------------|-------------------------------------------------------------|
| random     | rows are independent and order does not matter              |
| stratified | imbalanced classes: keeps the ratio in both sets            |
| group      | repeated subjects: all rows of one user stay together       |
| time-based | anything with a date: train on the past, test on the future |

→ If the same customer appears in train and test, the model recognises them instead of generalising. That is a group split, not a random one.

### COMMON TRAPS

## 04 Trap: the set you peeked at

A test set you have read ten times is a validation set with a misleading name.

|             |                                               |
|-------------|-----------------------------------------------|
| the leak    | you tuned until the test number looked good   |
| the symptom | offline gains that never show up live         |
| the fix     | tune on validation, open test once, then stop |

→ If you have already burned the test set, the honest move is to say so and collect a fresh holdout.