

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Reliable Estimates: Validation, Cross-Validation, Bias and Variance

Thursday 5 November 2026 · 50 minutes

Dr. Innocent Nyalala | BS Semester 5 (DSAI) | IIT Madras Zanzibar

By the End of Today You Can

IIT Madras Zanzibar

1. Explain why one train/test split gives a lucky or unlucky score
2. Say what training, validation and test sets are each for
3. Run k-fold cross-validation; report mean and spread
4. Diagnose high bias and high variance from two numbers
5. Compare two models fairly on the same folds

Today's Plan (40 minutes + questions)

IIT Madras Zanzibar

minutes	part	what you get
0 – 5	One split is luck; three sets	20 splits, 20 different scores
5 – 13	k-fold cross-validation	picture, steps, our data, code
13 – 20	Bias and variance	dartboard; diagnose from two numbers
20 – 30	The full toolbox + Weka	LOOCV, bootstrap, nested CV, leakage, significance
30 – 40	Fair comparison, exercise, summary	then questions and further reading

Recap and Today's Question

IIT Madras Zanzibar

Yesterday

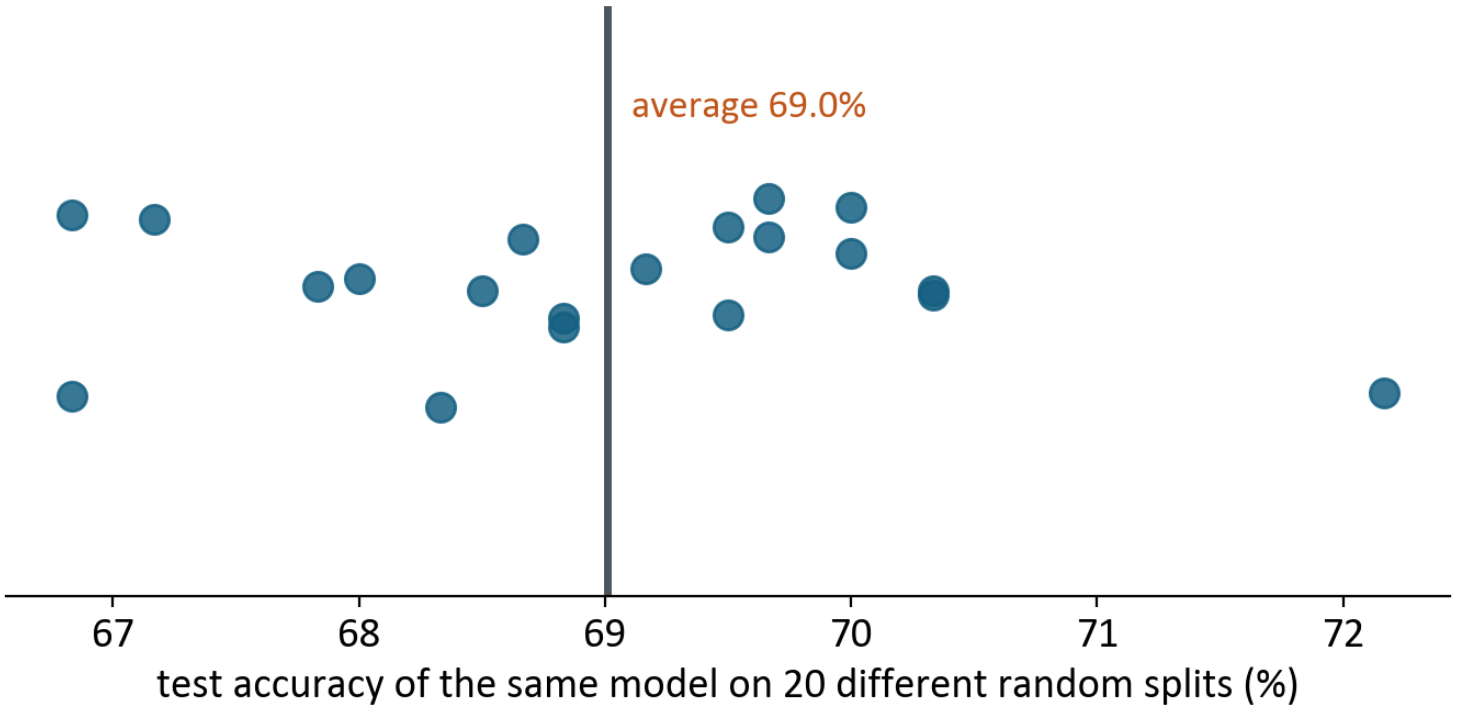
WHICH number?
accuracy, precision, recall, F1, AUC

Today

Can we TRUST the number?
Would another split give the same?

> **A score you cannot trust is worse than no score: it gives false confidence.**

One Split Is Luck



66.8% – 72.2%
same model, 20 random splits

average 69.0%
much more trustworthy

Same data, same model: a 5-point range from luck alone. Report an average over several splits.

Like Judging a Student on One Question

IIT Madras Zanzibar

One question

a good student can be unlucky;
a weak student can be lucky

one split

Many questions

luck averages out;
the score reflects real ability

cross-validation

Three Sets, Three Jobs

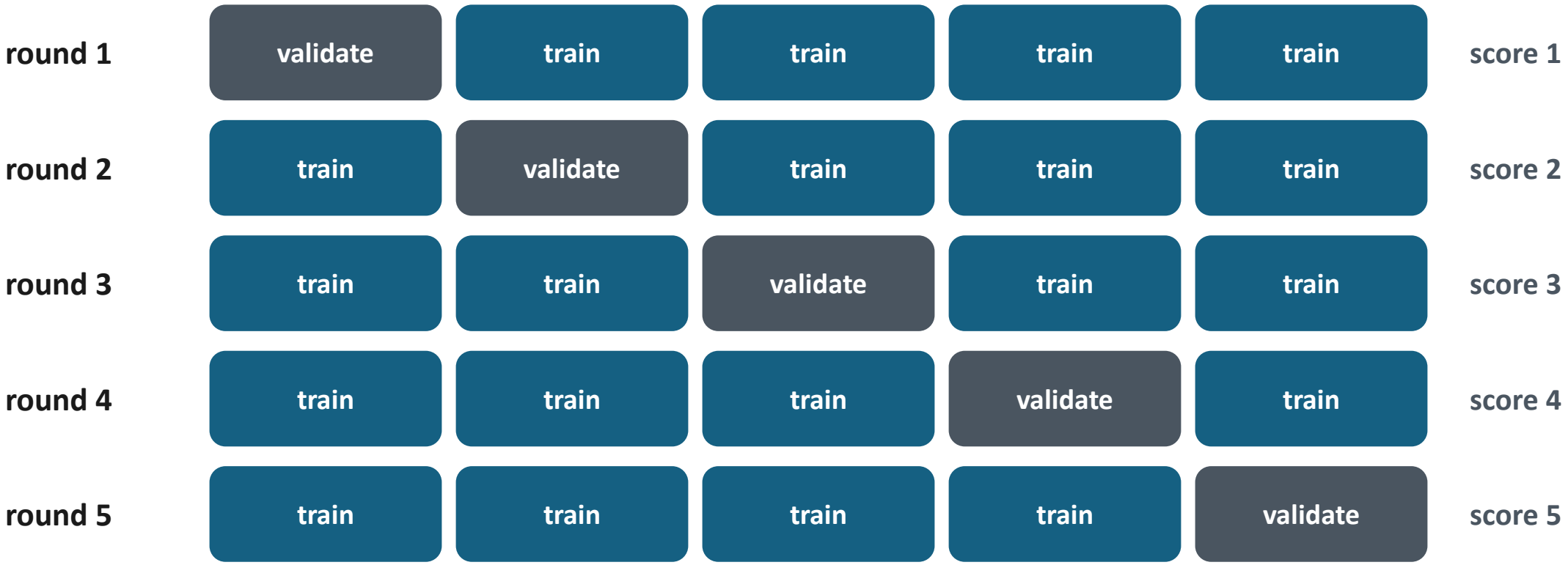
IIT Madras Zanzibar

TRAINING	VALIDATION	TEST
learn (revision notes)	choose (practice exams)	score once (real exam)

> **Peeking rule: every time the test set helps make a choice, it stops being a fair test.**

k-Fold Cross-Validation in One Picture

IIT Madras Zanzibar



Our data: 132,972 training seasons ÷ 5 ≈ 26,594 per fold; each round trains on 4 folds ≈ 106,378.
Every farm is validated exactly once. Report the MEAN of the 5 scores and their SPREAD.

k-Fold Cross-Validation, Step by Step

IIT Madras Zanzibar

1	**Cut**	split the training data into k equal folds (k = 5 or 10)
2	**Rotate**	each fold is the validation set once; train on the rest
3	**Average**	report the mean of the k scores and their spread
4	**Stratify**	every fold keeps the same share of High and Low

Definition: cross-validation = testing a model k times, each time on a different held-out part, and averaging

Cross-Validation Chooses the Tree Depth

IIT Madras Zanzibar

depth	5 fold scores (%)	mean
2	66.2 66.1 66.5 65.6 66.0	66.1%
6	70.9 70.8 72.0 70.8 71.1	71.1%
8	71.2 71.2 72.1 71.3 71.5	71.5% best
12	69.6 70.0 70.8 70.1 69.9	70.1%
full	62.4 62.5 62.6 62.4 62.4	62.5%

Choose depth 8 → retrain on all training data → test ONCE: 72.1%.

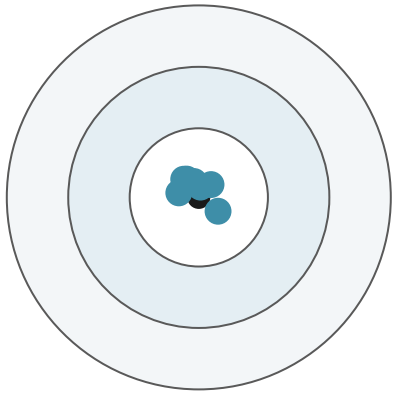
Cross-Validation in Python

```
1  from sklearn.model_selection import cross_val_score, StratifiedKFold
2  folds = StratifiedKFold(n_splits=5, shuffle=True, random_state=0)
3  for depth in [2, 6, 8, 12, None]:
4      s = cross_val_score(DecisionTreeClassifier(max_depth=depth), Xtv, ytv, cv=folds)
5      print(depth, s.round(3), "mean", s.mean().round(3), "sd", s.std().round(3))
6  final = DecisionTreeClassifier(max_depth=8).fit(Xtv, ytv)
7  print("test:", final.score(Xte, yte))           # once, at the end: 0.721
```

line	what it does
2	5 stratified folds
4	5 rounds in one call
5	mean and spread
6 – 7	retrain, test once

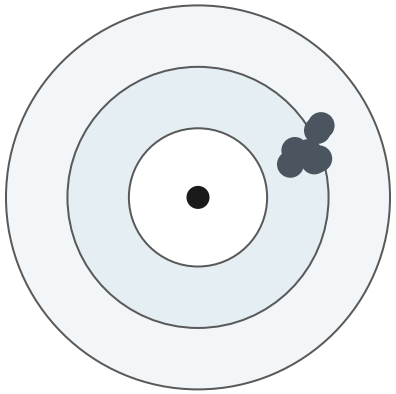
cross_val_score does all k rounds for you and returns the k scores.

Bias and Variance: Two Ways to Be Wrong



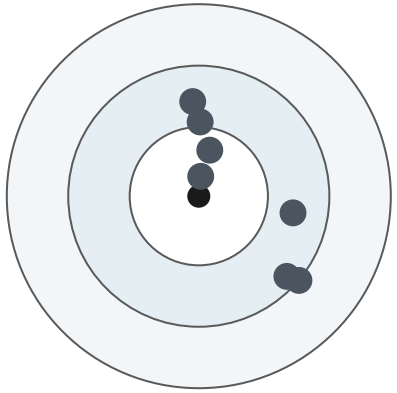
low bias, low variance

what we want



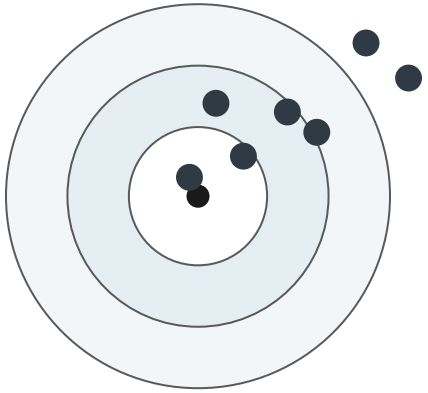
HIGH bias, low variance

too simple: underfitting



low bias, HIGH variance

too flexible: overfitting



high bias, high variance

worst of both

Bias and Variance on Our Farms

IIT Madras Zanzibar

depth	training	5-fold CV	gap	diagnosis
1	67.8%	66.4%	67.8 − 66.4 = 1.4	high BIAS (too simple)
4	73.7%	70.6%	73.7 − 70.6 = 3.1	balanced: best
full	100%	62.3%	100 − 62.3 = 37.7	high VARIANCE (memorized)

2,000 farms: best depth 4. 133,000 farms: best depth 8.
More data lets a model be more flexible before it overfits.

Diagnose From Two Numbers, Then Fix

IIT Madras Zanzibar

1	**train LOW, validation LOW**	high bias → more flexible model, better attributes
2	**train HIGH, validation much lower**	high variance → prune, more data, ensembles (next Monday)
3	**both high and close**	good fit → keep it

Comparing Two Models Fairly

IIT Madras Zanzibar

model	same 5 folds (%)	mean
tree (depth 8)	71.2 71.2 72.1 71.3 71.5	71.5%
Naive Bayes	68.9 69.2 70.0 69.2 69.1	69.3%
baseline (most common)	50.5 50.5 50.5 50.5 50.5	50.5%

Wins every fold

tree beats NB on all 5: a real difference

Baseline

50.5%: shows how much was learned

Fair Comparison in Python

```
1  folds = StratifiedKFold(n_splits=5, shuffle=True, random_state=0)   # made ONCE
2  models = {"tree": DecisionTreeClassifier(max_depth=8),
3            "naive bayes": CategoricalNB(),
4            "baseline": DummyClassifier(strategy="most_frequent")}
5  for name, model in models.items():
6      s = cross_val_score(model, X[name], y, cv=folds)      # same folds for all
7      print(name, s.round(3), s.mean().round(3))
8  # rare class? add scoring="f1" or scoring="roc_auc"
```

line	what it does
1	folds made once
2 – 4	models + baseline
6	same folds for all
8	rare class: F1 / AUC

X[name]: Naive Bayes uses banded categories, the tree uses the raw columns.

Honest Validation in the Real World

IIT Madras Zanzibar

Kaggle

hidden test set; peeking is punished

Clinical trials

final analysis fixed in advance

Credit models

validated on a later time period

Weather

checked on unseen past seasons

Spam filters

re-validated as spammers change

Your mini-project

graded on proper train/test separation

> **The final score is honest only if the test data never helped make a choice.**

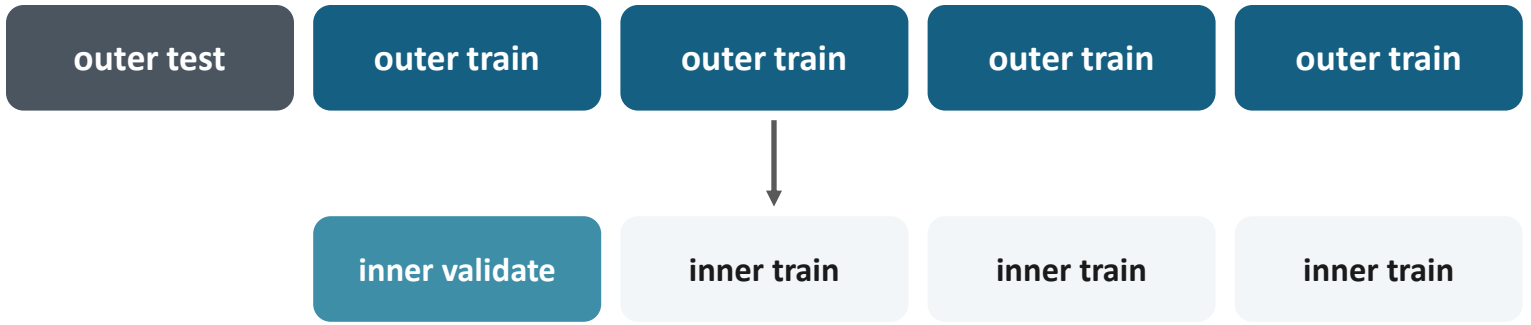
Four Ways to Estimate How Good a Model Is

IIT Madras Zanzibar

method	how	strength	weakness	use when
holdout	one train / test split	fast	one split is luck	huge data
k-fold (k = 5, 10)	each fold tested once	reliable default	k times the work	most problems
leave-one-out	k = number of records	uses all data	very slow; high variance	tiny datasets
bootstrap (.632)	draw n records with replacement; the ≈36.8% never drawn are the test set	good for small data	can be optimistic	small data, confidence intervals

Kohavi (1995): stratified 10-fold is a good default.
Our 190,000 seasons: holdout or 5-fold. 50 patients: leave-one-out or repeated 10-fold.

Repeated and Nested Cross-Validation



Repeated CV

re-shuffle and repeat k-fold several times; average all scores

Nested CV

inner loop tunes, outer loop scores: no optimism from tuning

Varma and Simon (2006): reporting the tuning score itself is optimistic.
scikit-learn: cross_val_score(GridSearchCV(model, grid), X, y).

Leakage: When the Test Set Secretly Helps

IIT Madras Zanzibar

1. Preparing before splitting

scaling or filling with all data → Pipeline

2. Same farm in both sets

farm 17 in train 2019 and test 2020 → GroupKFold

3. Future predicts the past

train on earlier seasons only → TimeSeriesSplit

4. Duplicate records

same row on both sides → remove first

```
pipe = make_pipeline(StandardScaler(), LogisticRegression())      # preparation inside each fold
cross_val_score(pipe, X, y, cv=GroupKFold(5), groups=d.farm_id)    # all seasons of a farm together
```

Is the Difference Real? Testing on the Same Folds

IIT Madras Zanzibar

model	fold 1	fold 2	fold 3	fold 4	fold 5
tree	71.2	71.2	72.1	71.3	71.5
Naive Bayes	68.9	69.2	70.0	69.2	69.1
difference	**2.3**	**2.0**	**2.1**	**2.1**	**2.4**

step	working	result
mean of differences	(2.3 + 2.0 + 2.1 + 2.1 + 2.4) ÷ 5	2.18
std of differences	typical distance of the 5 differences from 2.18	0.164
standard error	0.164 ÷ √5 = 0.164 ÷ 2.236	0.073
paired t	2.18 ÷ 0.073	≈ 30
corrected t (Nadeau–Bengio)	**$2.18 \div \sqrt{(1/5 + 1/4) \times 0.164^2} = 2.18 \div 0.110$**	**≈ 20, p < 0.001**

1/5: five folds; 1/4: each validation fold is 1/4 the size of its training part. t far above 2 → the difference is real.

Cross-Validation in Weka: No Code (5-Minute Demo)

IIT Madras Zanzibar

1	**Explorer**	every classifier: Cross-validation, Folds = 10 (stratified) → Start
2	**Experimenter**	New → add farm_seasons_5000.csv → 10-fold CV × 10 repetitions
3	**Algorithms**	add J48 and NaiveBayes → Run
4	**Analyse**	Experiment → Perform test: corrected paired t-test; v / * = significant

Repeated cross-validation, the same folds for every model, and a significance test: three clicks.

Your Turn (4 minutes, pairs)

IIT Madras Zanzibar

model	fold 1	fold 2	fold 3	fold 4	fold 5	training
A	70	72	69	71	73	73%
B	75	60	78	62	77	99%

1. Compute the mean of each model
2. Which has more variance? Give two clues
3. Which would you deploy, and why?
4. Suggest one fix for model B

Solution, Step by Step

IIT Madras Zanzibar

question	working	answer
mean A	70 + 72 + 69 + 71 + 73 = 355; ÷ 5	71.0
mean B	75 + 60 + 78 + 62 + 77 = 352; ÷ 5	70.4
variance	B swings 60 – 78 (A: 69 – 73); B train 99 vs ≈ 70	B
deploy	same mean, far more consistent	A
fix B	prune, more data, or an ensemble	prune

Summary

IIT Madras Zanzibar

1	**One split is luck**	20 splits of the same data: 66.8% – 72.2%
2	**Three sets**	train to learn, validate to choose, test once
3	**k-fold CV**	each record validated once; mean and spread; chose depth 8
4	**Bias vs variance**	low/low = bias; high/much lower = variance

Next: ensembles, many trees voting to cure high variance.

Questions

IIT Madras Zanzibar

Q: Why 5 or 10 folds?

A: Balance: more folds = more training data per round, but slower.

Q: Does CV replace the test set?

A: No. It replaces the single validation set; still test once.

Q: Why stratify?

A: Every fold keeps the class shares; vital for rare classes.

Q: Is lower variance always better?

A: Only if bias stays low: grouped AND centered.

Further Reading: Validation, Bias and Variance

IIT Madras Zanzibar

resource	why read it	access
James, Witten, Hastie & Tibshirani, An Introduction to Statistical Learning (Python ed., 2023)	cross-validation and bootstrap, with labs	free: statlearning.com
Hastie, Tibshirani & Friedman, The Elements of Statistical Learning (2009)	bias, variance, model assessment	free: hastie.su.domains
Kohavi (1995), IJCAI	cross-validation vs bootstrap	proceedings
Dietterich (1998), Neural Computation 10(7): 1895–1923	tests for comparing classifiers	journal
Nadeau & Bengio (2003), Machine Learning 52: 239–281	corrected resampled t-test	journal
Varma & Simon (2006), BMC Bioinformatics 7: 91	why nested cross-validation	free, open access
Kaufman, Rosset, Perlich & Stitelman (2012), ACM TKDD 6(4): 15	leakage in data mining	journal
scikit-learn User Guide: Cross-validation	every splitter, with code	free: scikit-learn.org