

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Ensembles: Bagging, Random Forests and Boosting

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 combining many models reduces variance
2. Describe bootstrap samples, bagging and random forests
3. Describe how boosting learns from mistakes, step by step
4. Read a random forest's feature importance
5. Judge when an ensemble helps, and when the data sets the limit

Today's Plan (45 minutes + 5 minutes questions)

IIT Madras Zanzibar

minutes	part	what you get
0 – 7	Wisdom of crowds	why averaging cancels errors
7 – 22	Bagging and random forests	bootstrap, vote, out-of-bag, importance
22 – 32	Boosting	learn from mistakes; gradient boosting
32 – 38	All methods on our data	accuracy, time, an honest limit
38 – 45	Code, exercise, summary	
45 – 50	Questions	

Recap: The Variance Problem

IIT Madras Zanzibar

One full tree

62.8% on new farms
high variance

Fix 1 (Monday)

prune it: depth 8
72.1%

Fix 2 (today)

many trees vote:
an ENSEMBLE

Definition: ensemble = a group of models whose predictions are combined, by voting or averaging

The Wisdom of Crowds

IIT Madras Zanzibar

787 guesses

the weight of an ox, 1906 individual guesses widely scattered

The crowd's middle guess

within 1% of the true weight

Why it works

errors in different directions cancel
(only if guesses are independent)

For models: build many trees that are wrong in DIFFERENT ways, then combine them.

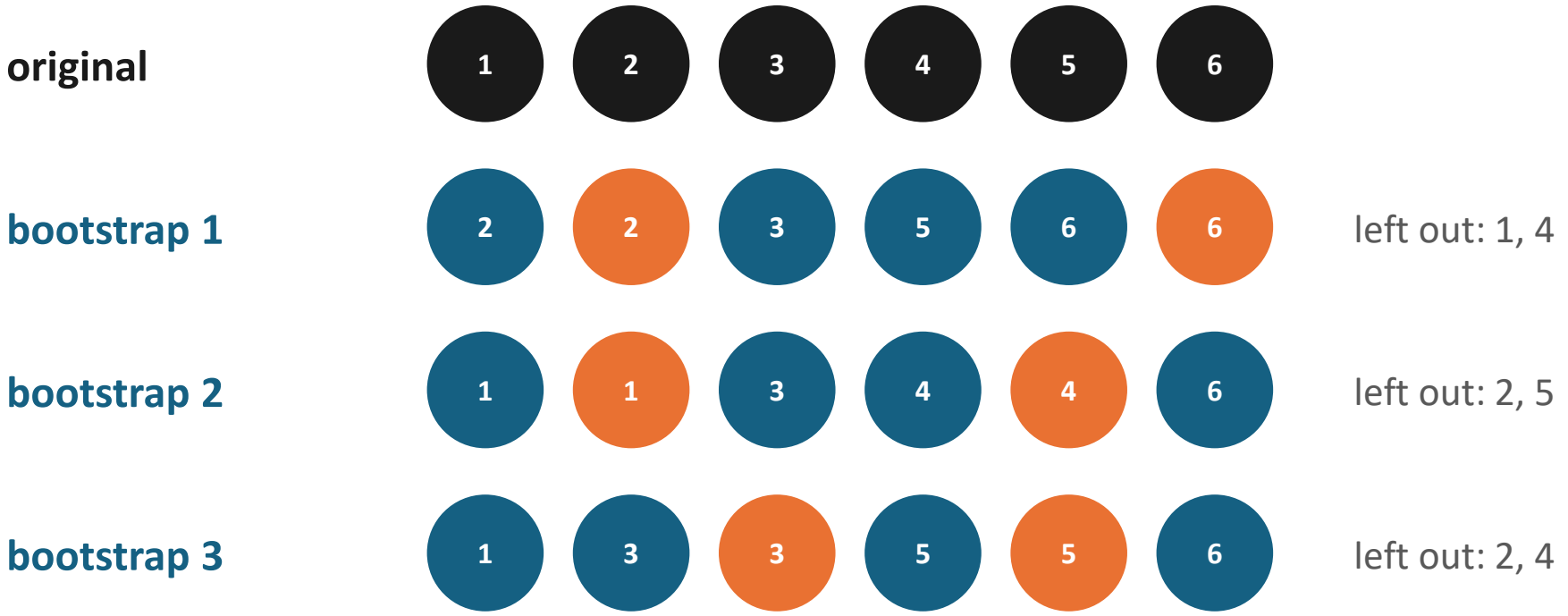
Why Averaging Shrinks Errors

IIT Madras Zanzibar

variance of an average of n independent errors = $\sigma^2 \div n$

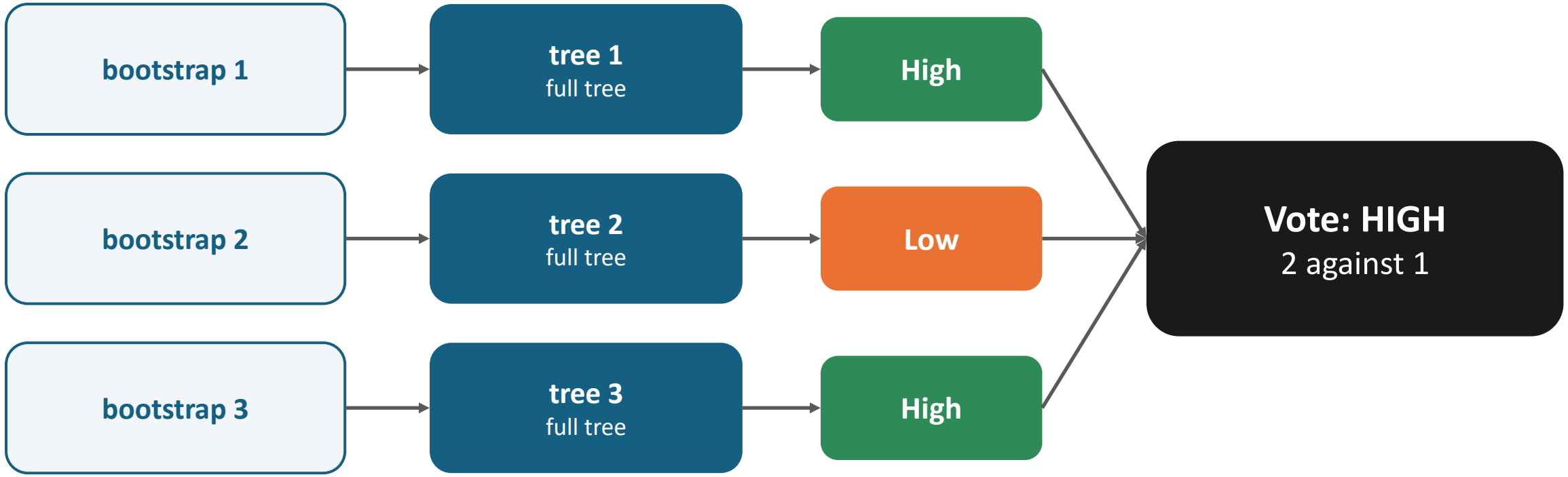
1	σ^2	how much ONE model wobbles from sample to sample
2	$\div n$	average n models: the wobble shrinks n times
3	**the catch**	only if errors are independent → make the trees DIFFERENT

Bootstrap Samples: Different Rows for Each Tree



Draw n rows WITH replacement: some rows twice (orange), some not at all.
Each tree sees a slightly different dataset.

Bagging: Bootstrap, Grow, Vote



Each tree overfits its own sample, but differently. In the vote their mistakes cancel.

A Free Validation Set: Out-of-Bag Farms

IIT Madras Zanzibar

$$P(\text{a farm is never drawn in n draws}) = (1 - 1/n)^n \approx 1/e \approx 0.37$$

	working	left out of each tree
n = 10	$(1 - 1/10)^{10} = 0.9^{10}$	0.349 → about 3 or 4 of 10 farms
n large	approaches 1/e	0.368 → about 37% of farms

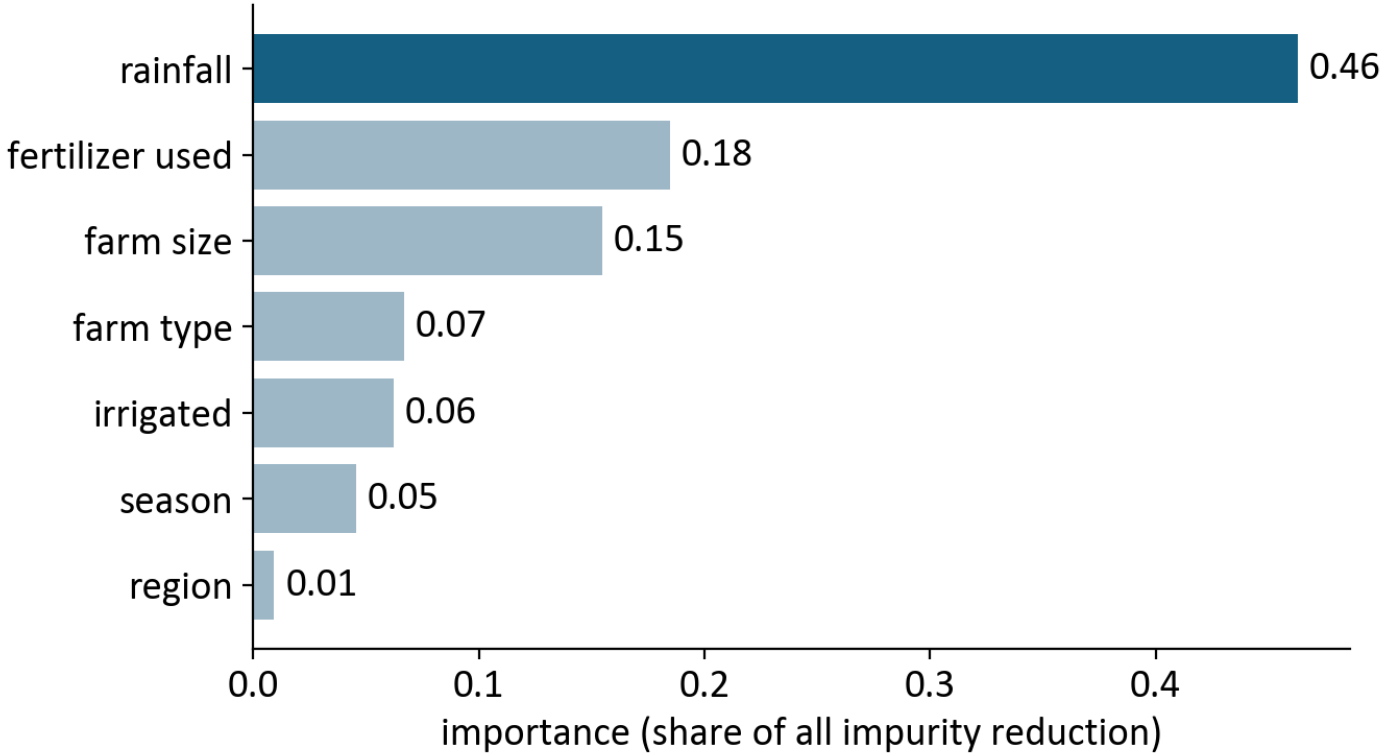
About 37% of farms are 'out-of-bag' for each tree: a free validation set (oob_score=True).

Random Forests: Also Randomise the Questions

IIT Madras Zanzibar

1	**Random rows**	each tree gets a bootstrap sample (like bagging)
2	**Random questions**	each split chooses among a random few attributes (≈ √d)
3	**More different trees**	errors cancel better in the vote
4	**The price**	200 trees cannot be read: use feature importance

Feature Importance: What Did the Forest Use?



Rainfall 46%
of all purity gained

Fertilizer 19%
farm size 15%

Crop, region ≈ 1%
barely matter once others known

Importance = share of all impurity reduction an attribute provides, over all trees.

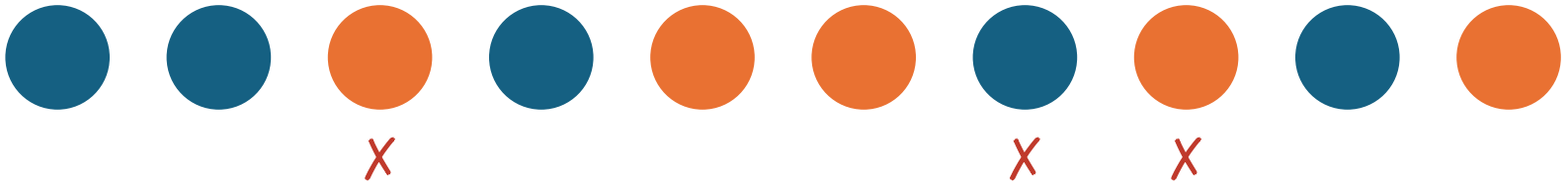
How Many Trees?

trees	test accuracy
1	70.9%
5	71.8%
10	72.0%
50	72.1%
100	72.2%
200	72.2%

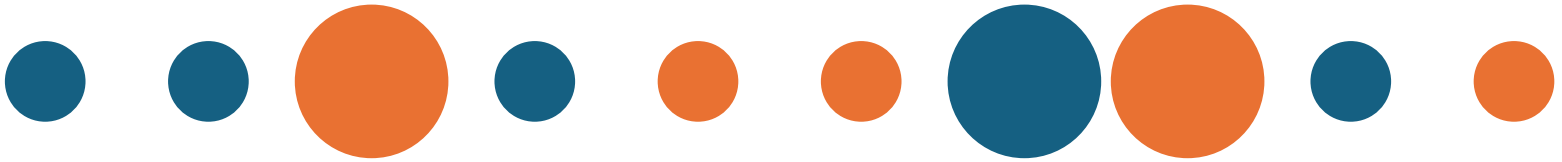
More trees never hurt
accuracy rises, then flattens;
only time grows, and trees train
in parallel (n_jobs = -1)

Boosting: Learn From Your Mistakes

Round 1
equal weights



Round 2
mistakes weigh more



Repeat

hundreds of small trees, each fixing the last

Weighted vote

more accurate stumps get a bigger say

Like a student who re-studies only the questions they got wrong after each practice test.

Gradient Boosting: Fit What Is Still Wrong

IIT Madras Zanzibar

1	**Start**	a simple first guess (e.g. overall share of High)
2	**Fit the errors**	a small tree predicts what is still wrong
3	**Add a little**	prediction += learning rate × new tree
4	**Repeat**	hundreds of small corrections

bagging / forest	boosting
independent trees, in parallel; reduces VARIANCE	trees in sequence, each fixes the last; reduces BIAS

Every Method on Our 190,000 Farm-Seasons

IIT Madras Zanzibar

method	test accuracy	training time
pruned tree (depth 8)	72.1%	0.6 s
full tree	62.8%	1.3 s
bagging (50 full trees)	68.8%	10.2 s
random forest (200 trees)	72.2%	2.1 s
AdaBoost (200 stumps)	72.2%	71.9 s
gradient boosting (histogram)	72.3%	5.4 s

The Honest Lesson: The Data Sets the Limit

IIT Madras Zanzibar

Ensembles fixed

the full tree's variance:
62.8% → 72%

They cannot fix

missing information: pests, soil, planting dates, skill

To do better

collect better attributes, not a fancier algorithm

Every good model stops near 72%: the limit is in the data, not the algorithm.

Ensembles in Python

IIT Madras Zanzibar

```
1  from sklearn.ensemble import RandomForestClassifier, HistGradientBoostingClassifier
2  rf = RandomForestClassifier(n_estimators=200, min_samples_leaf=50, n_jobs=-1, oob_score=True)
3  rf.fit(Xtr, ytr)
4  print("out-of-bag", rf.oob_score_, " test", rf.score(Xte, yte))    # ≈ 0.72
5  imp = pd.Series(rf.feature_importances_, index=Xtr.columns)
6  print(imp.sort_values(ascending=False).head())             # rainfall first
7  gb = HistGradientBoostingClassifier().fit(Xtr, ytr)
8  print("boosting test", gb.score(Xte, yte))                 # ≈ 0.72
```

line	what it does
2	200-tree forest
3	train
4	OOB and test
5 – 6	importance
7 – 8	boosting

Same fit / score pattern as a single tree: swapping models is one line.

Ensembles in the Real World

IIT Madras Zanzibar

Credit scoring

gradient boosting is the workhorse

Mobile money

forests flag unusual payments

Satellite maps

forests map crop types

Medicine

importance ranks risk factors

Competitions

boosted trees win on tables

Weather

forecast = average of many runs

Many different, imperfect models, combined, are steadier than any one.

Your Turn (5 minutes, individually)

IIT Madras Zanzibar

1. Five trees vote: High, Low, High, High, Low. What does the bagged ensemble predict?

2. Ten farms are bootstrapped. About how many are left out of one sample?

3. In boosting, a farm is misclassified by the first stump. What happens to its weight?

4. A forest and a single tree both score 72%. Give one reason to prefer each.

Solution, Step by Step

IIT Madras Zanzibar

question	working	answer
1 vote	High 3, Low 2	High
2 left out	$0.9^{10} = 0.349 \rightarrow 0.349 \times 10 \approx 3.5$	3 or 4 farms
3 boosting	misclassified → weight goes up	next stump focuses on it
4 choose	tree: readable, fast · forest: steadier, importance	depends on the need

Summary

IIT Madras Zanzibar

1	**Wisdom of crowds**	independent errors cancel: $\sigma^2 \div n$
2	**Bagging, forests**	bootstrap + vote; full tree 62.8% → 68.8%; forest 72.2%
3	**Boosting**	sequential: each tree fixes the last; 72.3%
4	**Honest limit**	≈ 72% for all: better data beats fancier algorithms

Next week: clustering, learning WITHOUT labels, starting with k-means.

Questions (5 minutes)

IIT Madras Zanzibar

Q: Is a forest always better than one tree?

A: Not always more accurate (as here), but steadier on new data.

Q: Can a forest overfit?

A: More trees do not; each tree can, so limit the leaf size.

Q: Why was AdaBoost slow?

A: Its trees must be built one after another, never in parallel.

Q: What does the learning rate do?

A: Makes each correction small: slower to learn, overfits less.