

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Classification I: Decision Trees

Wednesday 28 October 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 what classification is, and how it differs from clustering and association rules
2. Name the classifiers in this course and the one idea behind each
3. Read a decision tree and use it to classify a new farm
4. Explain 'how mixed is a group' with Gini and entropy, in plain words
5. Pick the best question for a tree using information gain, by hand

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

IIT Madras Zanzibar

minutes	part	what you get
0 – 10	The big picture	what classification is; every classifier we will study
10 – 15	What a decision tree is	read a tree; classify new farms
15 – 30	How a tree chooses questions	pure vs mixed groups; Gini; entropy; information gain
30 – 35	Trees on real data, in Python	8 lines of code on 190,000 farm-seasons
35 – 45	Your turn, then summary	a 5-minute exercise and its solution
45 – 50	Questions	

Where We Are: Three Big Data Mining Tasks

IIT Madras Zanzibar

Association rules

Week 5 (done)

What happens together?

e.g. fertiliser ⇒ seeds

no answer column

Classification

Weeks 6 – 7 (TODAY)

Which class will this case be?

e.g. farm → High or Low yield

answer column given (label)

Clustering

Weeks 8 – 9

Which cases are alike?

e.g. 4 natural farm groups

no answer column

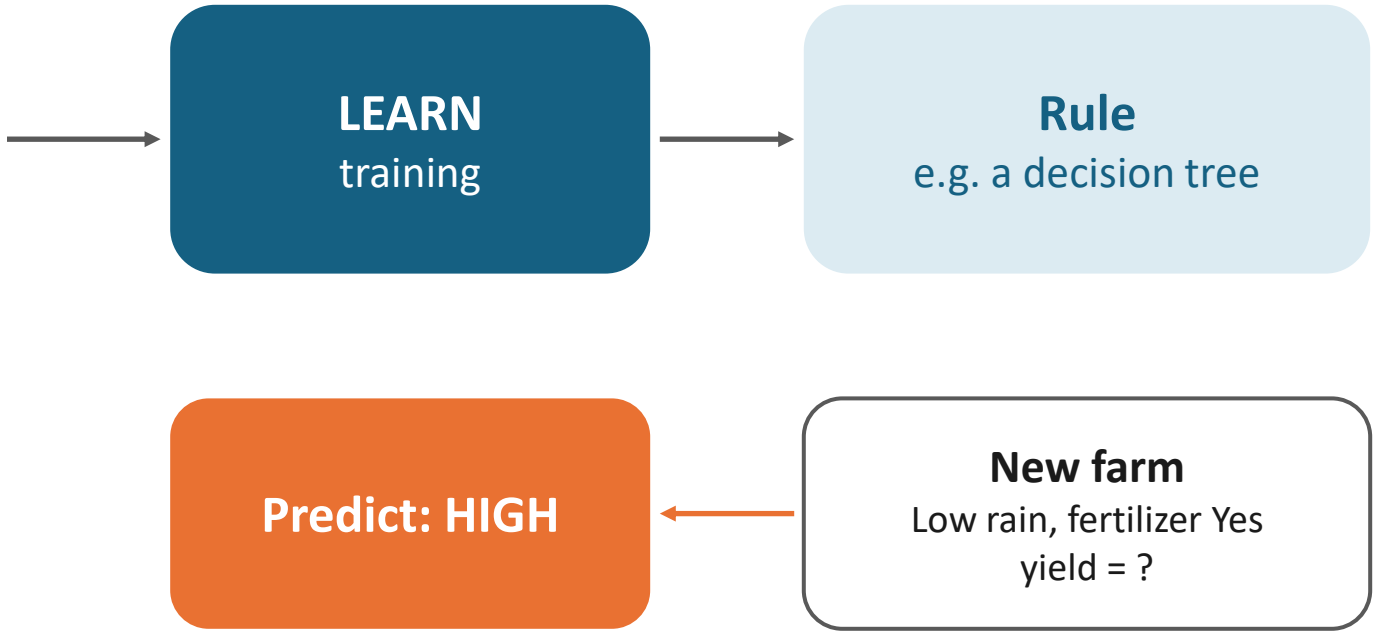
The key question: is there an answer column (a label) to learn from? Yes → classification. No → clustering.

What Is Classification? Learn From the Past, Predict the New

IIT Madras Zanzibar

rainfall	fertilizer	yield
High	No	High
Low	No	Low
Low	Yes	High
Medium	No	Low

past farms: the answer is known



Three Words You Will Hear

IIT Madras Zanzibar

Class label

the answer column we predict

yield: High or Low

Attributes

the columns we predict FROM

rainfall, fertilizer, irrigated

Training / test data

learn from one part; check on a hidden part

like keeping exam questions back

The Classification Methods Ahead

IIT Madras Zanzibar

Today

Decision tree

asks the best questions, one after another

a doctor's checklist

Next

Naive Bayes

counts how often each class happened before

'70% chance of rain'

Then

Pruning

stops a tree from memorising

memorising past papers

Then

Ensembles

many trees vote

ask 100 farmers, not 1

Week 7

Evaluation

accuracy is not enough

a test that misses cases

Week 7

Cross-validation

test on several splits

several mock exams

Classification in Data Mining: Where It Is Used

IIT Madras Zanzibar

Banks

approve this loan?

Telecom

will this customer leave?

Mobile money

is this payment fraud?

Health

malaria likely?

Agriculture

High or Low yield?

Email

spam or not?

Large history of records + a known outcome → a rule that supports a decision.

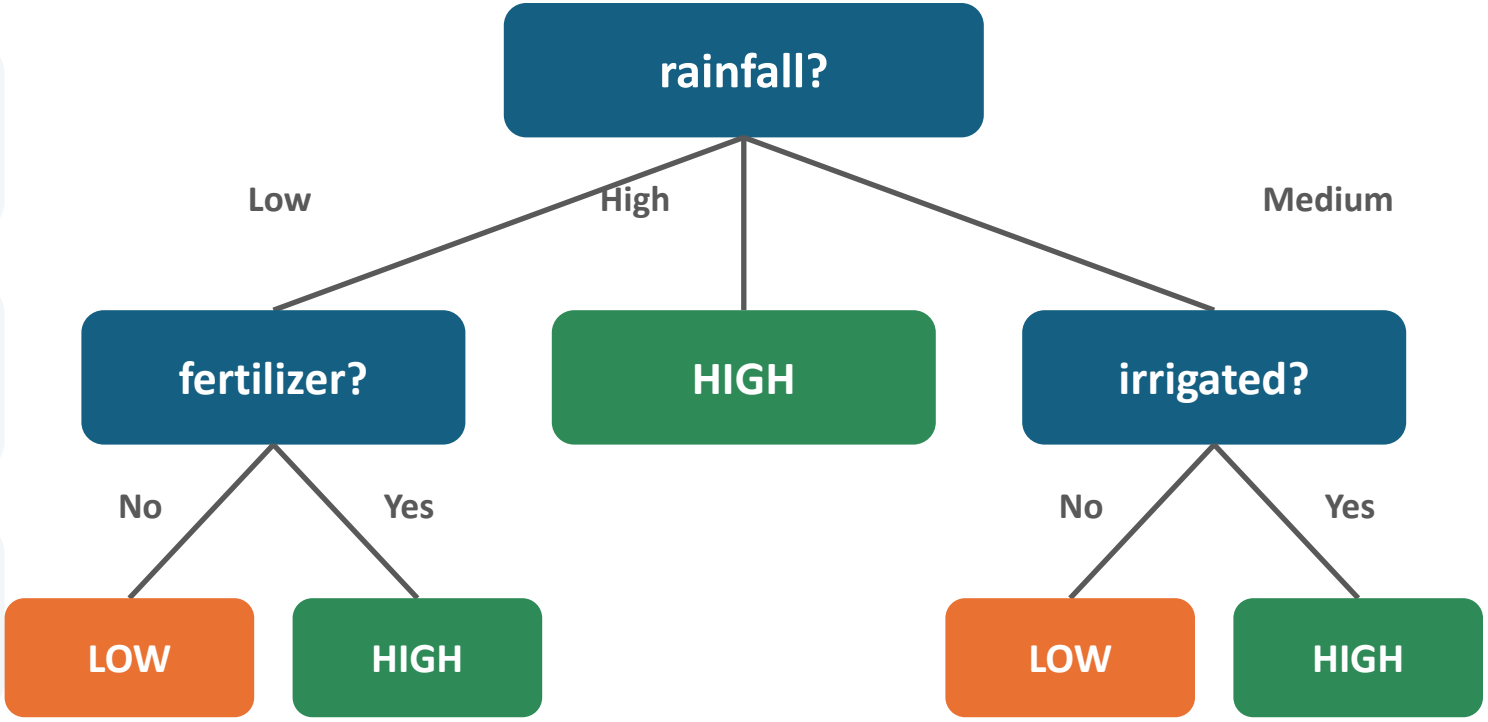
A Decision Tree Is a Flowchart of Questions

IIT Madras Zanzibar

Root the first question

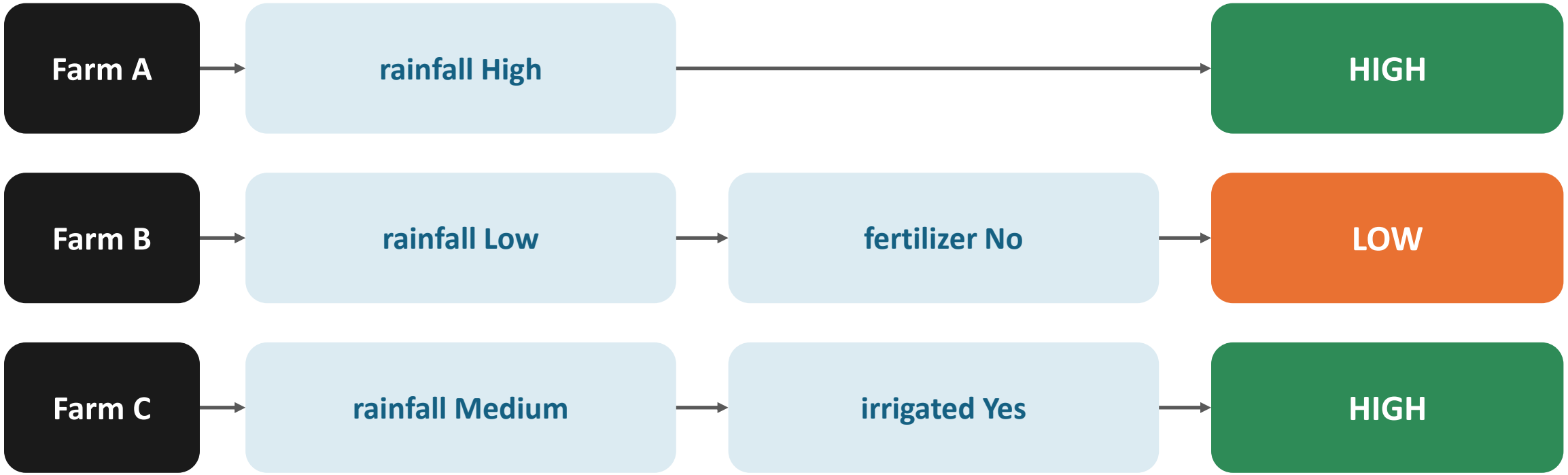
Branch one possible answer

Leaf the final prediction



Like a doctor's checklist. The computer learns which questions to ask, and in what order.

Use the Tree: Classify Three New Farms

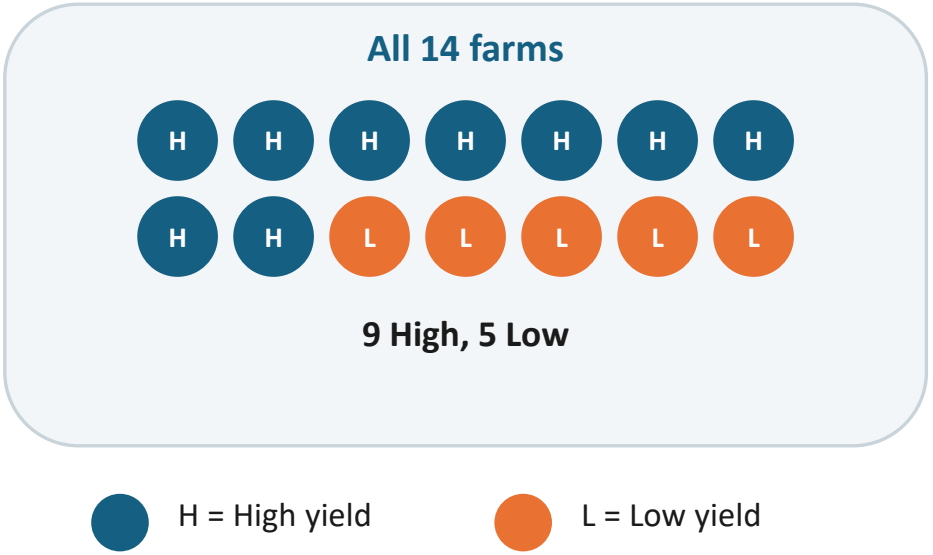


Farm B is never asked about irrigation: a tree asks only the questions on its own path.

How Does the Computer Choose? Our 14 Farms

IIT Madras Zanzibar

#	rainfall	size	fertilizer	irrigated	yield
1	Low	Large	No	Yes	Low
2	Low	Large	No	No	Low
3	High	Large	No	Yes	High
4	Medium	Medium	No	Yes	High
5	Medium	Small	Yes	Yes	High
6	Medium	Small	Yes	No	Low
7	High	Small	Yes	No	High
8	Low	Medium	No	Yes	Low
9	Low	Small	Yes	Yes	High
10	Medium	Medium	Yes	Yes	High
11	Low	Medium	Yes	No	High
12	High	Medium	No	No	High
13	High	Large	Yes	Yes	High
14	Medium	Medium	No	No	Low



Goal: find the question that sorts the dots into groups of one colour.

A Good Question Makes Pure Groups

IIT Madras Zanzibar

Ask about rainfall

rainfall = High

H H H H

PURE

rainfall = Low

H H L L L

mixed

rainfall = Medium

H H H L L

mixed

Good: one group is pure

Ask about farm size

size = Large

H H L L

mixed

size = Medium

H H H H L L

mixed

size = Small

H H H L

mixed

Poor: all groups still mixed

Putting a Number on 'How Mixed'

IIT Madras Zanzibar

4 H, 0 L: pure



measure	value
error	0
Gini	0
entropy	0

3 H, 1 L: a little mixed



measure	value
error	0.25
Gini	0.375
entropy	0.811

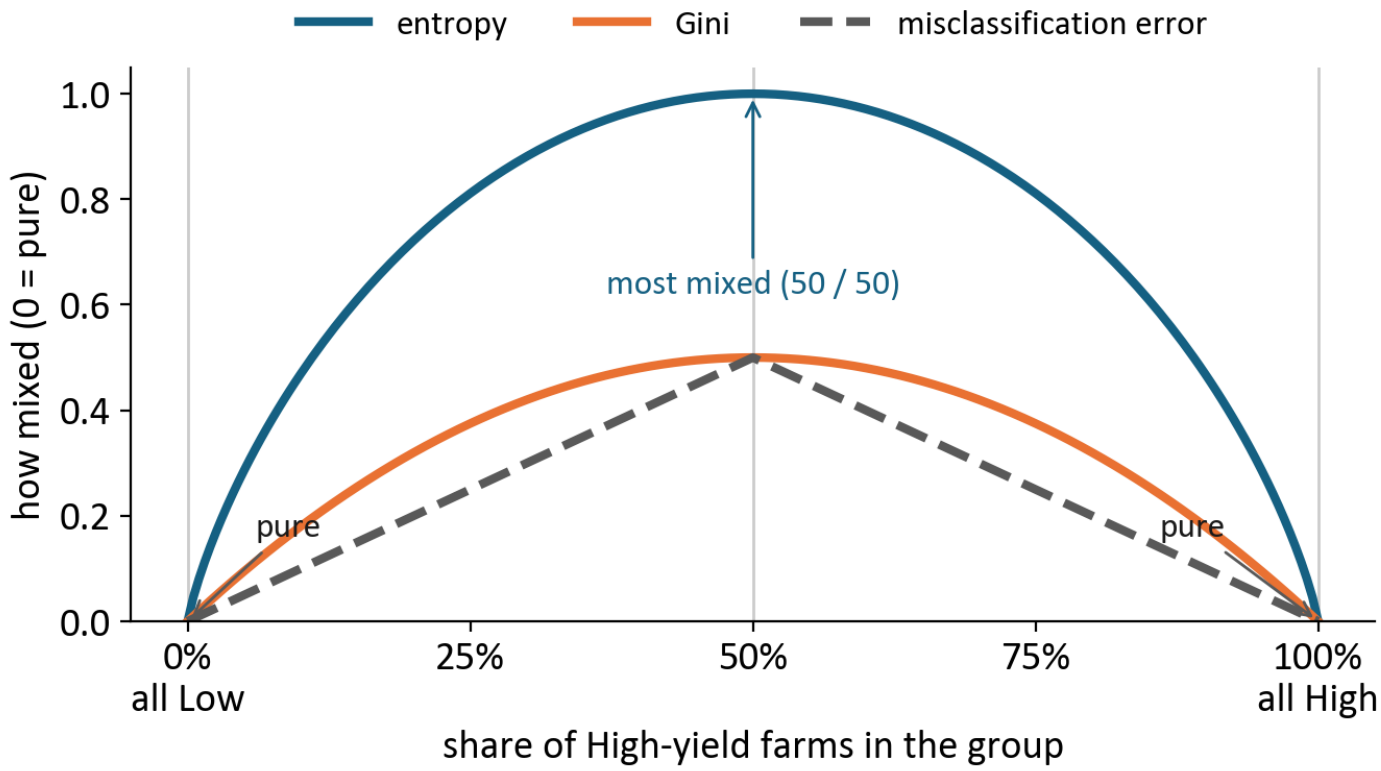
2 H, 2 L: most mixed



measure	value
error	0.5
Gini	0.5
entropy	1.0

Impurity = a number for how mixed a group is. 0 = pure. Biggest at 50 / 50.

The Impurity Curve, Explained



Bottom	share of High farms in the group
Side	how mixed the group is
Ends	0: pure (one class only)
Middle	peak: 50 / 50, most mixed

The closer to half-and-half, the more mixed the group.

Gini Index in Plain Words

IIT Madras Zanzibar

Pick 2 farms at random.
Gini = the chance they have DIFFERENT answers.

Pure group → 0
Half-and-half → 0.5

both High	$0.643^2 = 0.413$
both Low	$0.357^2 = 0.128$
they agree	$0.413 + 0.128 = 0.541$
Gini (disagree)	$1 - 0.541 = 0.459$

Formula: Gini = 1 − (share High)² − (share Low)²

Entropy: Decoding H(S) Symbol by Symbol

IIT Madras Zanzibar

$$H(S) = -\Sigma\ p_i \cdot \log_2 p_i$$

$H(S)$	mixedness of group S
p_i	share of class i
$\log_2 p_i$	how many halvings reach p_i
Σ and $-$	add over classes; make it positive

Plain words: the average number of yes/no questions needed to guess a random farm's answer.
Pure → 0 questions. 50 / 50 → 1 question.

Entropy of Our 14 Farms, Step by Step

IIT Madras Zanzibar

class	share p_i	$\log_2 p_i$	$p_i \times \log_2 p_i$
High	9/14 = 0.643	−0.637	−0.410
Low	5/14 = 0.357	−1.485	−0.530
add up			−0.940
H(S)	**flip the sign**		**0.940**

Entropy = 0.940: close to the maximum of 1, so the 14 farms are very mixed.

Information Gain = How Much Less Mixed After the Question

IIT Madras Zanzibar

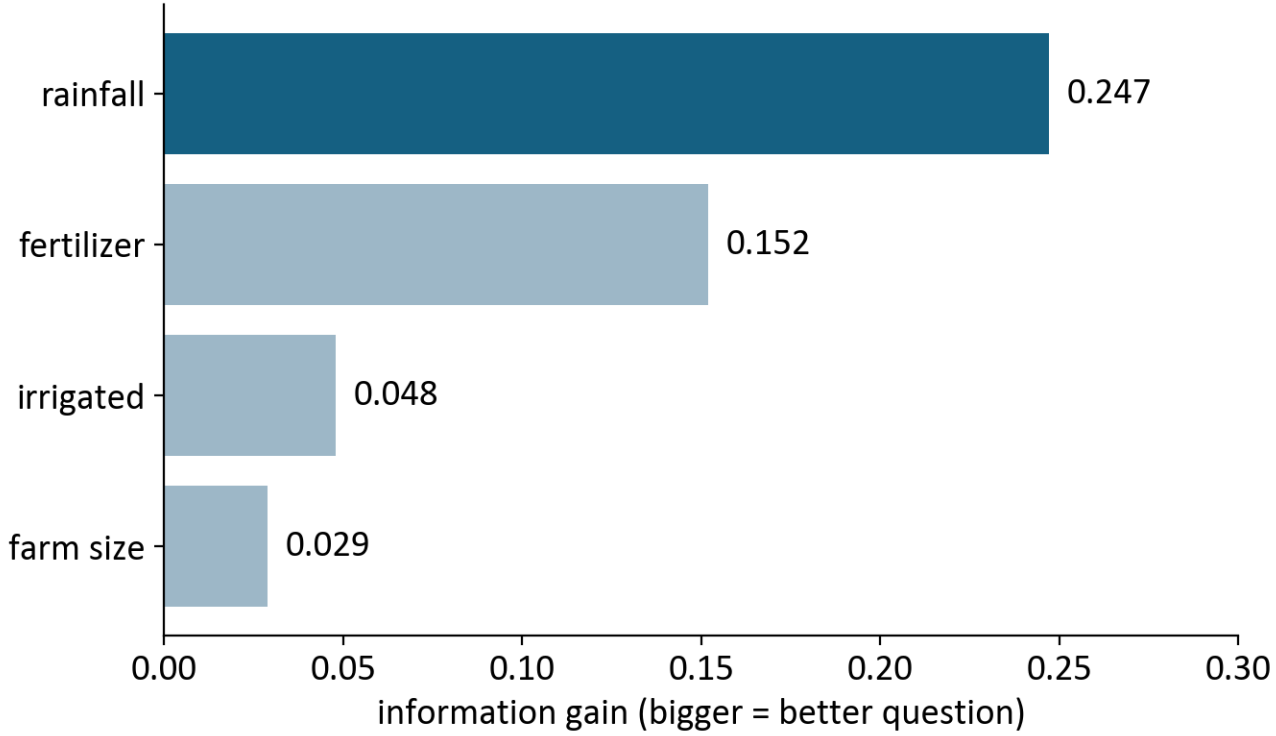
BEFORE
14 farms
9 H, 5 L
entropy 0.940

- **High: 4 farms** — 4 H, 0 L → entropy 0
- **Low: 5 farms** — 2 H, 3 L → entropy 0.971
- **Medium: 5 farms** — 3 H, 2 L → entropy 0.971

GAIN = BEFORE − AFTER
0.940 − 0.694 = 0.247

group	share of farms	entropy	share × entropy
rainfall = High	4 / 14 = 0.286	0.000	0.286 × 0.000 = 0.000
rainfall = Low	5 / 14 = 0.357	0.971	0.357 × 0.971 = 0.347
rainfall = Medium	5 / 14 = 0.357	0.971	0.357 × 0.971 = 0.347
AFTER (add up)			**0 + 0.347 + 0.347 = 0.694**

Score Every Question, Pick the Biggest Gain



question	entropy gain	Gini gain
rainfall	0.247	0.116
fertilizer	0.152	0.092
irrigated	0.048	0.031
farm size	0.029	0.019

Winner: rainfall, the biggest gain. It becomes the root. Gini gives the same order.

Entropy and Information Gain in 11 Lines of Python

```
 1  import numpy as np, pandas as pd          # farms = our 14-farm table
 2  def entropy(labels):
 3      p = labels.value_counts(normalize=True)   # share of each class
 4      return -(p * np.log2(p)).sum()            # -sum p*log2(p)
 5  def gain(df, col, target="yield"):
 6      before = entropy(df[target])
 7      after = sum(len(g) / len(df) * entropy(g[target])
 8                  for _, g in df.groupby(col))  # weighted average
 9      return before - after
10  for col in ["rain", "size", "fert", "irr"]:
11      print(col, round(gain(farms, col), 3))
```

output	gain
rain	0.247
size	0.029
fert	0.152
irr	0.048

Same numbers as by hand.

Gini version: change line 4 to return 1 - (p ** 2).sum()

Why Not Just Count the Errors?

IIT Madras Zanzibar

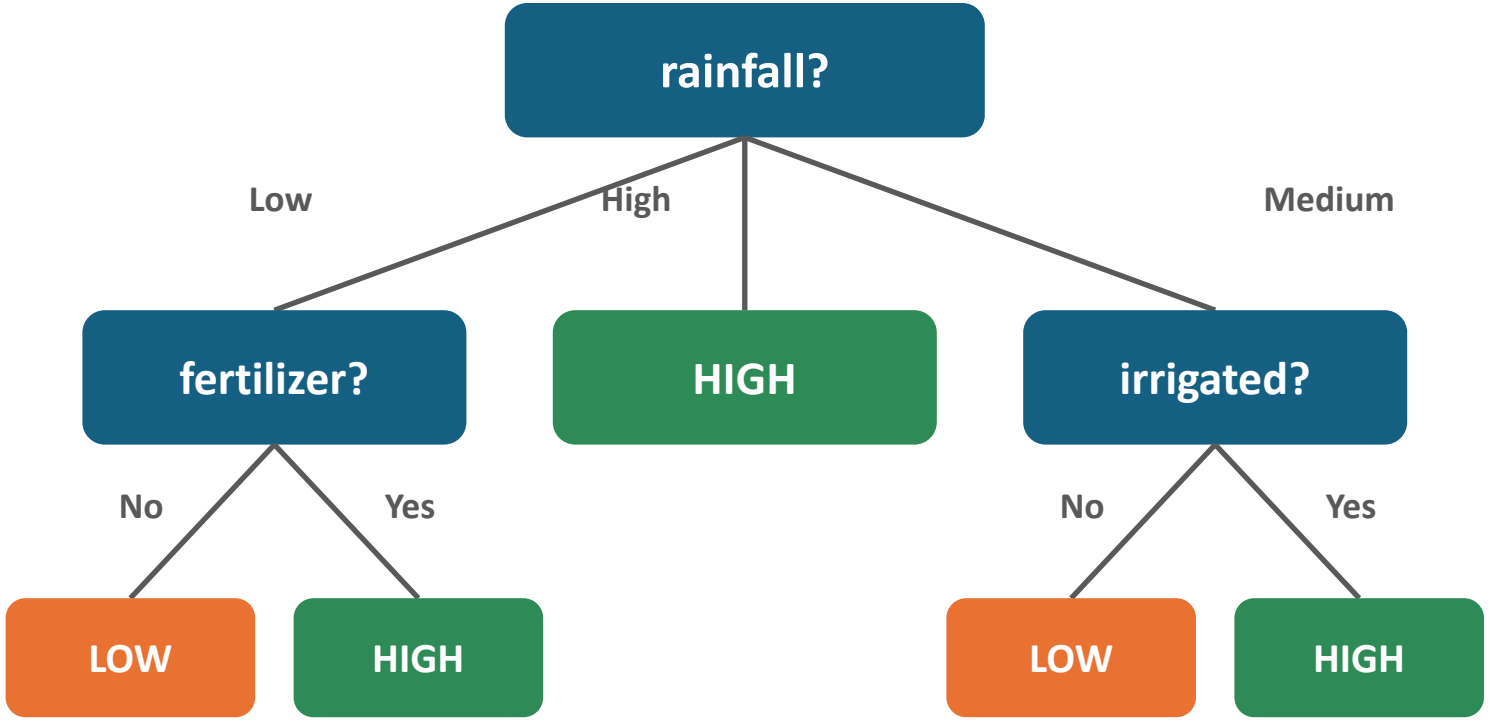
question	error gain	entropy gain
rainfall	0.071	0.247
fertilizer	0.071	0.152
irrigated	0.000	0.048
farm size	0.000	0.029

Error: rainfall = fertilizer (tie).
It cannot see that rainfall made a PURE group.

Entropy and Gini see every step towards purity.

That is why decision trees use Gini or entropy, not the plain error.

Grow the Tree: Repeat Inside Each Branch



Pick the best question → split → repeat inside every branch → stop when groups are pure.

When Does the Tree Stop Growing?

IIT Madras Zanzibar

1. Pure group	everyone has the same answer → leaf
2. No questions left	leaf = the majority answer
3. No farms in a branch	leaf = the parent's majority
4. A limit is reached	max depth or too few farms (used in practice)

ID3 uses information gain · C4.5 uses gain ratio · CART (scikit-learn) uses Gini.

Decision Trees in Python: 190,000 Farm-Seasons

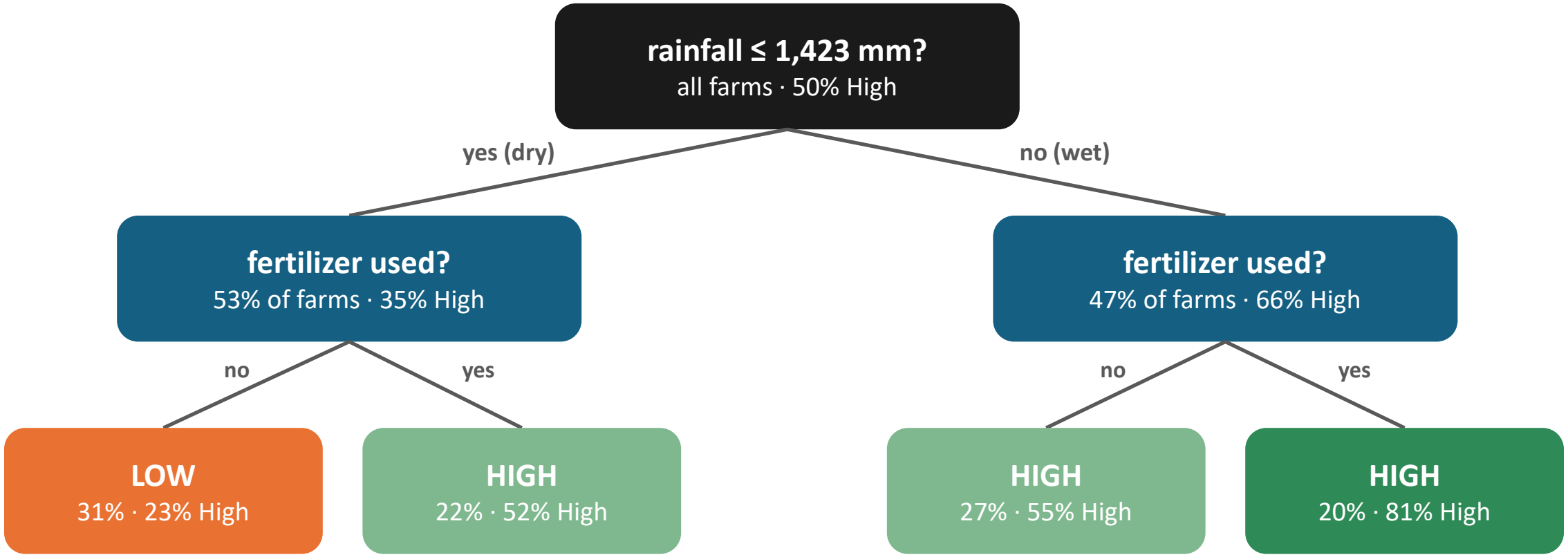
IIT Madras Zanzibar

```
 1  import pandas as pd
 2  from sklearn.tree import DecisionTreeClassifier
 3  from sklearn.model_selection import train_test_split
 4  d = pd.read_csv("zanzibar_farm_seasons.csv").dropna()
 5  X = pd.get_dummies(d[["rainfall_mm", "fertilizer_used", "irrigated"]])
 6  y = d["yield_class"]
 7  Xtr, Xte, ytr, yte = train_test_split(X, y, test_size=0.3)
 8  tree = DecisionTreeClassifier(criterion="gini", max_depth=2)
 9  tree.fit(Xtr, ytr)
10  print(tree.score(Xte, yte))      # about 0.66
```

line	what it does
1 – 3	load the tools
4	read data, drop missing
5 – 6	attributes and label
7	hide 30% as test data
8 – 9	make and grow a Gini tree
10	accuracy on hidden farms

Depth 2: 66% correct on unseen farms. Depth 8: 70%. Guessing: 50%.

The Tree the Computer Grew From Real Data



Same logic as by hand: rainfall first, then fertilizer. Darker green = more sure of HIGH.

Your Turn (5 minutes, pairs)

IIT Madras Zanzibar

10 students: 7 passed (P), 3 failed (F). Two possible first questions:

question	answer Yes	answer No
attended tutorials?	6 P, 1 F	1 P, 2 F
studied in a group?	4 P, 1 F	3 P, 2 F

1. Which question makes purer groups? (look at the counts)
2. Gini of all 10 students
3. Entropy of all 10 (if fast)

Recipe: Gini = 1 − (share P)² − (share F)². Entropy = −(pP·log₂pP + pF·log₂pF).

Solution 1: Which Question Makes Purer Groups?

IIT Madras Zanzibar

Question: attended tutorials?

Yes: 7 students

P P P P P P F

6 P, 1 F: almost pure

No: 3 students

P F F

1 P, 2 F

Purer

Question: studied in a group?

Yes: 5 students

P P P P F

4 P, 1 F

No: 5 students

P P P F F

3 P, 2 F

Still mixed

Answer 1: attended tutorials makes purer groups, so it is the better first question.

Solution 2: Gini of All 10 Students, Step by Step

IIT Madras Zanzibar

step	what we compute	working
1	shares	P = 7/10 = 0.7 F = 3/10 = 0.3
2	both passed	$0.7^2 = 0.7 \times 0.7 = 0.49$
3	both failed	$0.3^2 = 0.3 \times 0.3 = 0.09$
4	they agree	0.49 + 0.09 = 0.58
5	**Gini (they disagree)**	**1 − 0.58 = 0.42**

Gini = 1 − (share P)² − (share F)² = 1 − 0.49 − 0.09 = 0.42

Solution 3: Entropy of All 10 Students, Step by Step

IIT Madras Zanzibar

class	share p_i	$\log_2 p_i$	$p_i \times \log_2 p_i$
Passed (P)	7/10 = 0.7	−0.515	0.7 × −0.515 = −0.360
Failed (F)	3/10 = 0.3	−1.737	0.3 × −1.737 = −0.521
add up			−0.360 + (−0.521) = −0.881
H(S): flip sign			**0.881**

Entropy = $-(0.7 \times \log_2 0.7 + 0.3 \times \log_2 0.3) = 0.881$
Calculator tip: $\log_2 x = \log x \div \log 2$

Solution 4: Confirm With Information Gain

IIT Madras Zanzibar

group	share	entropy	share × entropy
tutorials = Yes	7/10 = 0.7	0.592	0.7 × 0.592 = 0.414
tutorials = No	3/10 = 0.3	0.918	0.3 × 0.918 = 0.275
gain			0.881 − (0.414 + 0.275) = 0.191

group	share	entropy	share × entropy
group study = Yes	5/10 = 0.5	0.722	0.5 × 0.722 = 0.361
group study = No	5/10 = 0.5	0.971	0.5 × 0.971 = 0.485
gain			0.881 − (0.361 + 0.485) = 0.035

Before = 0.881. Tutorials gain 0.191 > group study gain 0.035 → root: attended tutorials.

Summary

IIT Madras Zanzibar

Classification	learn from labelled past cases; predict the class of new ones
Decision tree	a flowchart of questions; follow one path from root to leaf
Choosing questions	pick the split with the purest groups: Gini or entropy; gain = before − after
Growing	repeat in every branch; stop when pure or at a limit; 8 lines in scikit-learn

Next: Naive Bayes, classifying by counting probabilities instead of asking questions.

Questions (5 minutes)

IIT Madras Zanzibar

Q: Does it find the best possible tree?

A: No: it is greedy, it picks the best question now and never looks ahead.

Q: Do I need to normalise the data?

A: No: each question compares one attribute to a cut-off, so scale does not matter.

Q: Can a tree predict a number?

A: Yes: a regression tree predicts the average value in each leaf.

Q: Why was farm size never used?

A: It barely made the groups purer, so the tree skipped it.