TUTORIAL 3 · STUDENT HANDOUT

Normalization, Discretization & PCA on 500 Zanzibar Farms

Thursday 1 October 2026, 4:45–5:30 pm

Today: you will apply Lectures 7 and 8 to real data in Python, check every result against a number given in this handout, and receive the Mini-Project brief. Parts B–D of today's notebook are exactly the Preprocessing section of your project.

1. Learning outcomes

By the end of the tutorial you will be able to:

- Normalize attributes with **min-max** and **z-score** in scikit-learn — and verify one value **by hand**
- Fit a scaler on the **training set only**, and explain why that prevents data leakage
- Discretize an attribute with equal-frequency bins **and** with entropy / information gain, and say which is better and why
- Run PCA on standardized data, read a scree plot and loadings, and explain why raw-unit PCA misleads

2. What you need

Item	Details
Notebook	`W3_Tutorial3_STUDENT.ipynb` (course portal)
Dataset	`datasets/W1_zanzibar_crop_sample.csv` — 500 farms, same file as Tutorials 1–2
Packages	pandas, numpy, matplotlib, scikit-learn (install: `pip install scikit-learn`; Google Colab already has them)
From lectures	Lecture 7 (normalization, entropy) and Lecture 8 (PCA) — worked examples on the slides

3. Plan for the 45 minutes

Time	Part	What you do	Checkpoint
4:45–4:50	A. Load & clean	run the cell; 24 missing yields filled with crop median	(500, 9), 0 missing
4:50–5:00	B. Normalize	min-max, z-score, hand-check, fit on train only	0.8232; 1409.8
5:00–5:12	C. Discretize	equal-frequency bins; entropy(); information gain search	167/166/167; 0.39 at $1.20

5:12–5:22	D. PCA	scree plot, loadings, 2-D picture of all farms	42 / 20 / 20 / 16 / 1 %
5:22–5:30	E. Graded	4 questions; then read the Mini-Project brief	see rubric

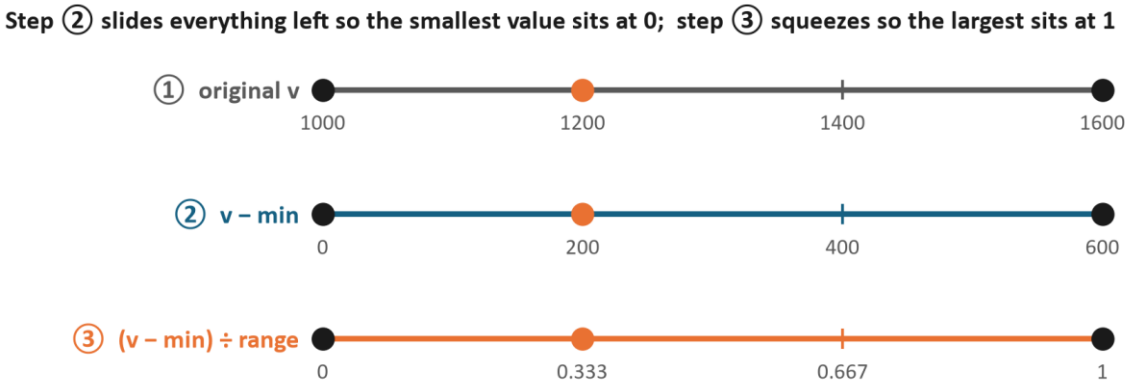
4. The ideas you need — explained from zero

4.1 Normalization: putting every attribute on the same ruler

Our farms record size in hectares (0.2 – 9), price in $/kg (0.2 – 9) and revenue in dollars (20 – 24,000). Any method that adds attributes together — a distance, PCA — would be controlled by revenue simply because its numbers are bigger. Normalization rescales every attribute to a common range. It changes the ruler, **not** the order of values or the shape of the distribution.

$$\text{min-max: } v' = (v - \min) / (\max - \min) \qquad \text{z-score: } z = (v - \mu) / \sigma$$

v = one farm's value · v′, z = new value · min, max = smallest/largest in the column · μ = column mean · σ = column standard deviation



Min-max in two moves: subtract the min (smallest → 0), then divide by the range (largest → 1).

Worked check (you will reproduce this in B2)	Value
Farm 1 rainfall v	1,780.8 mm
Column min and max	682.2 mm and 2,016.8 mm
v − min	1,780.8 − 682.2 = 1,098.6
max − min	2,016.8 − 682.2 = 1,334.6
v′	1,098.6 / 1,334.6 = **0.8232**

Data leakage: when you split data into training and test sets, learn min/max or μ/σ from the TRAINING rows only and reuse them on the test rows. Learning them from all rows lets the test set 'leak' into preprocessing, and your reported accuracy becomes too optimistic.

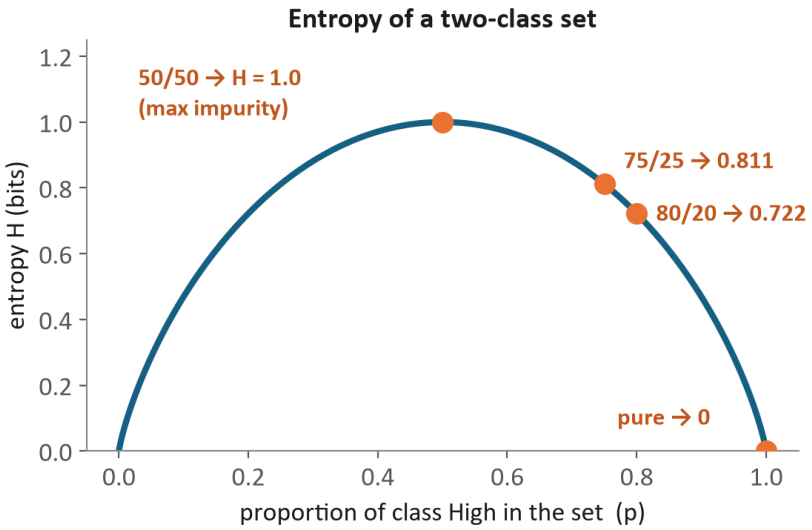
4.2 Entropy and information gain: choosing a boundary with the class label

Our class label is whether a farm's yield is **High** or **Low** (split at the median, 2.19 t). **Entropy** measures how mixed a group is: 0 for a pure group, 1 for a perfect 50/50 mix. **Information gain** measures how much a split point reduces that mixing.

$$H(S) = -\Sigma\, p_i \log_2 p_i \qquad IG = H(\text{before}) - [\,(n_L/n)\cdot H(L) + (n_R/n)\cdot H(R)\,]$$

p_i = share of class i · n = records in the group · n_L, n_R = records left/right of the split · $\log_2 x = \ln x \div \ln 2$

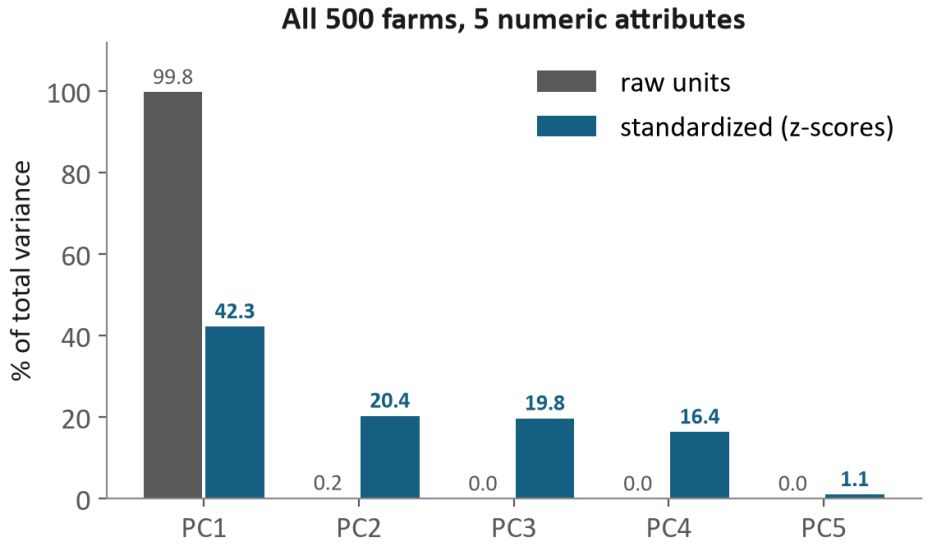
Group	p(High)	Entropy H	Reading
8 High, 0 Low	1.00	0	pure — no surprise
6 High, 2 Low	0.75	0.811	some mixing
4 High, 4 Low	0.50	1.000	maximum mixing



Entropy of a two-class group as the share of 'High' changes.

4.3 PCA: fewer columns, most of the information

PCA finds new axes — **principal components** — along which the data is most spread out. Each component is a weighted mix of the original attributes (the weights are called **loadings**). The **eigenvalue** of a component is the variance along it; its share of the total is the % of variance explained. Always **standardize first**: in raw units revenue (variance ≈ 33 million) swamps farm size (variance ≈ 2), and PC1 falsely claims 99.8% of the variance.



Our 500 farms: raw-unit PCA (grey) vs standardized PCA (teal).

How many components to keep?	Rule
Elbow	keep components before the scree curve flattens
Cumulative target	keep enough for 80–90% of the variance
Kaiser (standardized data)	keep components with eigenvalue > 1

5. Working through the notebook

Read each explanation cell first. Replace every `# TODO` with code. Cells marked ✍ need a written answer in your own words. **After every part, compare your output with the checkpoint below — if it differs, fix it before moving on.**

Part	Task	☑ You should see
A	Load, impute missing yields	24 missing → 0; shape (500, 9)
B1	Min-max and z-score all 5 columns	min-max 0 … 1; z means ≈ 0, std = 1
B2	Hand-check farm 1 rainfall	0.8232 by hand and in scikit-learn
B3	Fit scaler on train only	train mean ≈ 1409.8 vs full mean 1397.9
C1	3 equal-frequency rainfall bins	167 / 166 / 167; cuts ≈ 1279 and 1503 mm
C2	Your entropy() function	≈ 1.0 for all 500 farms; 0 for a pure group
C3	Best split by information gain	rainfall ≈ 0.005 (useless); price ≈ 0.39 at ≈ $1.20/kg
D1	PCA on standardized data	≈ 0.423, 0.204, 0.198, 0.164, 0.011
D4	Loadings of PC1 and PC2	PC1: price, revenue (+), yield (−); PC2: farm size, rainfall (+)

D5	2-D scatter of all farms	clove and cashew farms separate along PC1

Discovery in C3: rainfall tells us almost nothing about whether yield is high or low in this dataset, but price per kg does. Cheap bulk crops (cassava, rice, coconut; below ≈ $1.20/kg) have high tonnage. Supervised discretization finds this automatically — equal-frequency binning never could.

6. Graded exercise (Part E) and rubric

#	Task	What 'complete' looks like	Marks
E1	Min-max `price_per_kg`; hand-check farm 1 (0.26; min 0.20; max 9.06)	working shown; 0.0068 both ways	2
E2	`pd.cut` farm size into 4 equal-width bins	counts 403 / 78 / 13 / 6 and a one-line reason	2
E3	PCA on raw (unstandardized) columns	ratios ≈ [0.998, 0.002, 0, 0, 0] printed and compared with D1	2
E4	Why standardize before PCA?	2–3 sentences quoting YOUR E3 numbers and naming revenue's scale	4

7. Common errors and fixes

Symptom	Cause	Fix
ModuleNotFoundError: sklearn	scikit-learn not installed	`pip install scikit-learn`, then restart the kernel
FileNotFoundError for the CSV	notebook not next to the datasets folder	put the notebook in the folder that contains datasets/
Entropy prints nan	$\log_2$ of 0 for an empty class	use value_counts(): classes with 0 records are not listed
Loadings have opposite signs to a friend's	PCA signs are arbitrary	fine — the meaning is the same; compare sizes
qcut: 'Bin edges must be unique'	many repeated values	use fewer bins or duplicates='drop'

8. Submission

- Run all cells so every output is visible
- Save as **YourName_W3_Tutorial3.ipynb**
- Upload to the course portal before **5:30 pm today**
- Then read **Mini_Project_Brief.pdf** — your project's Preprocessing section uses exactly Parts B–D