

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Synthesis: The Whole Pipeline on Our Farms

Thursday 12 November 2026 · 50 minutes

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

By the End of Today You Can

IIT Madras Zanzibar

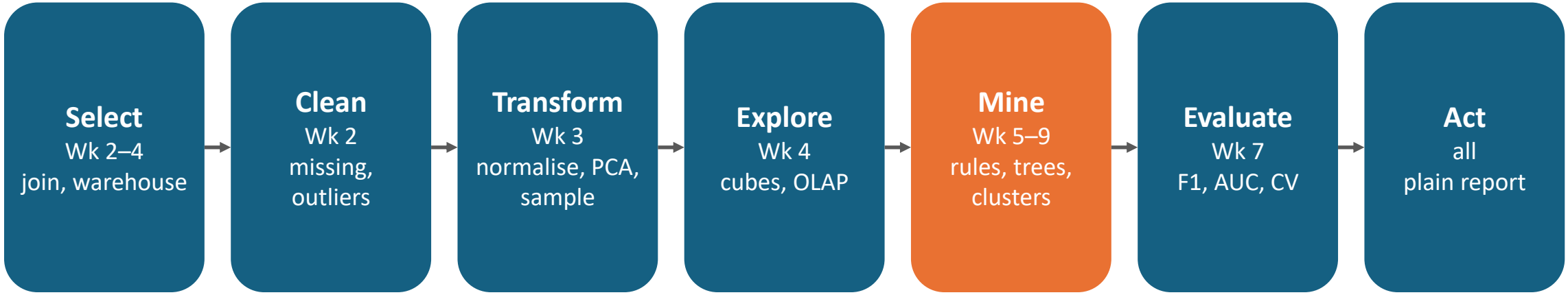
1. Place every technique on a step of the data mining (KDD) process
2. Plan a complete analysis for a real question
3. Choose a technique for each step, and justify it
4. Name the scaling idea behind each method
5. Report results honestly: limits, fairness, privacy

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

IIT Madras Zanzibar

minutes	part	what you get
0 – 6	The process, filled in	every week on one pipeline
6 – 22	One question, end to end	which farms need support next season?
22 – 30	Choosing and scaling	decision guide; five scaling ideas
30 – 36	Ethics and honest reporting	fairness, limits, privacy
36 – 45	Wrap-up and handover	nine weeks, Part 2, checklist, exercise
45 – 50	Questions	

The Data Mining Process, Now Filled In



Lecture 1's warning, now proven: preparing the data (the first four boxes) took more effort than mining, and decided whether the mining could be trusted.

One Real Question: Which Farms Need Support?

IIT Madras Zanzibar

Ministry: 'Find farms at risk of a failed season, in time to send help. About 9,000 visits per season.'

Who is at risk?

rare-class classification
(Weeks 6 – 7)

What kind of farm?

clustering for targeting
(Weeks 8 – 9)

What to bring?

association rules on dealer baskets (Week 5)

Translate first: a vague request → a precise task with a label, a measure and a constraint.

Steps 1 – 3: Select, Clean, Transform

IIT Madras Zanzibar

1	**Select**	farm seasons + farm details via the star schema: 200,000 rows
2	**Clean**	25 gauge errors (9,999 mm) → missing; 8,111 seasons without yield removed
3	**Label**	failed = bottom 10% of yield within its crop → 10.2% positives
4	**Encode**	categories → 0/1 columns; z-score only for clustering

Report every choice: each cleaning and labelling decision changes the result.

The Whole Pipeline in Python

```
 1  d = pd.read_csv("zanzibar_farm_seasons.csv")
 2  d.loc[d.rainfall_mm >= 5000, "rainfall_mm"] = np.nan          # gauge errors
 3  d = d.dropna(subset=["yield_tonnes", "rainfall_mm"])           # report the count!
 4  cut = d.groupby("crop").yield_tonnes.transform(lambda s: s.quantile(0.10))
 5  d["failed"] = (d.yield_tonnes <= cut).astype(int)              # 10.2% positives
 6  X = pd.get_dummies(d[attributes], drop_first=True)
 7  Xtr, Xte, ytr, yte = train_test_split(X, d.failed, test_size=0.3, stratify=d.failed)
 8  tree = DecisionTreeClassifier(max_depth=8).fit(Xtr, ytr)       # pruned
 9  risk = tree.predict_proba(Xte)[:, 1]
10  visit = risk >= 0.2                                            # fits ≈ 9,000 visits
11  print(recall_score(yte, visit), precision_score(yte, visit), roc_auc_score(yte, risk))
```

lines	step
1	select
2 – 3	clean
4 – 6	transform
7 – 9	mine
10 – 11	evaluate

Every line is a technique from this course.

Steps 4 – 5: Mine and Evaluate

IIT Madras Zanzibar

measure	value	in plain words
farms flagged (test set)	9,261	≈ the 9,000-visit budget
recall	53%	more than half of failing farms found
precision	33%	1 visit in 3 is to a failing farm
F1	0.41	balanced score
ROC AUC	0.83	ranking quality over all thresholds

Threshold 0.2 fits the visit budget and reaches over half of the failing farms;
the 90%-accurate lazy model reaches none.

Steps 6 – 7: Target, Package, Act

1	**Who**	classifier at threshold 0.2 → list of farms to visit
2	**What kind**	clusters: 'small, no inputs' → subsidy; 'mid-size, no inputs' → advice
3	**What to bring**	rules: rice seed + urea (lift 6.8); NPK + cassava cuttings (lift 11.5)
4	**Check first**	DBSCAN's unusual farms: verify the records
5	**Report**	one page, plain words, numbers and limits

Choosing a Technique: a Decision Guide

IIT Madras Zanzibar

your question	task	start with	then try
What goes together?	pattern mining	Apriori / FP-Growth	lift, Kulczynski
Which group does this belong to?	classification	tree, Naive Bayes	forest, boosting
What natural groups exist?	clustering	k-means + silhouette	hierarchical, DBSCAN
How do totals vary?	OLAP	cube, roll-up, drill-down	iceberg cube
Too many columns?	reduction	PCA (standardise first)	feature importance
Too many rows?	reduction	stratified sample	mini-batch, SON, BIRCH

Start from the QUESTION, not the algorithm.

The Course's Throughline: Exact vs Scalable

IIT Madras Zanzibar

weeks	exact method	scalable idea
2 – 3	full statistics, full PCA	sampling, one-pass (Welford), randomised PCA
4	full data cube	iceberg cube, partial materialisation
5	all itemsets	Apriori pruning, FP-Growth, PCY, SON
6 – 7	exact split search	sampling, histogram splits
8 – 9	k-means, hierarchical	mini-batch k-means, BIRCH, DBSCAN

sample **summarise** **prune** **fewer passes** **parallelise**

Ethics and Honest Reporting

IIT Madras Zanzibar

Fairness

compare recall across regions and farm types:
who gets missed?

Honest limits

≈ 72% ceiling: pests, soil, timing
are not in the data

Privacy

farm records are personal data:
aggregate, anonymise, restrict

Scope

state what it must NOT be used for
(e.g. denying credit)

Always break results down by group: equal accuracy can hide unequal errors (COMPAS, Lecture 1).

Nine Weeks in One Picture

IIT Madras Zanzibar

week	you learned to …	signature result on our data
1	describe data; the KDD process	Zanzibar farms, 9 attributes
2	similarity; clean and integrate	missing yields filled, recorded
3	transform, reduce, warehouse	raw PCA 99.8% (scale artefact) → honest 42%
4	cubes and OLAP	2017 drought found by drill-down
5	mine and judge patterns	rice seed → urea, lift 6.8
6	trees, Naive Bayes, pruning	full tree: 100% train, 63% test
7	evaluate honestly; ensembles	90% accurate model helps no one
8 – 9	cluster, at scale	4 farm types; unusual farms flagged

Part 2 With Prof. Mahindran, From 17 November

IIT Madras Zanzibar

Part 2 topic	builds on (Part 1)
advanced classification	trees, Naive Bayes, evaluation, ensembles (Weeks 6 – 7)
advanced clustering	k-means, hierarchical, DBSCAN (Weeks 8 – 9)
deep learning	normalisation (Week 3); overfitting and validation (Weeks 6 – 7)
outlier detection	IQR, robust z (Weeks 2 – 3); DBSCAN noise (Week 9)
Quiz 2 and final exam	everything above

Revise Weeks 6 – 9 before Part 2 starts on 17 November.

Mini-Project: Final Checklist

IIT Madras Zanzibar

1	**Runs**	notebook runs top to bottom; outputs visible
2	**Justified**	every preprocessing choice explained in one sentence
3	**No leakage**	scalers fitted on training data only
4	**Evaluated**	test set, right measure, threshold or silhouette
5	**Specific**	scaling and ethics paragraphs about YOUR data
6	**Summary**	one page, plain words, sources cited

Your Turn (5 minutes, pairs)

IIT Madras Zanzibar

Mobile money company: 'Which of our agents are likely to stop trading next month?'

1. Which data mining task is this?
2. What is the label, and how would you build it?
3. Which measure would you report, and why not accuracy?
4. Which method would you start with, and how would you validate it?
5. Which scaling idea, and one ethics check?

Solution

IIT Madras Zanzibar

question	a good answer
task	classification with a rare positive class
label	no transactions next month, built from past months
measure	recall, precision, F1, AUC: not accuracy (rare class)
method	pruned tree → forest or boosting; cross-validate, test once
scaling, ethics	summarise per agent or sample; histogram boosting; recall by region

Summary and Thank You

IIT Madras Zanzibar

1	**Pipeline**	select → clean → transform → explore → mine → evaluate → act
2	**Question first**	failed seasons: classifier + clusters + rules
3	**Five scaling ideas**	sample, summarise, prune, fewer passes, parallelise
4	**Honesty**	fairness by group, limits, privacy, scope

Data mining is not about algorithms. It is about asking a good question, preparing data honestly, choosing the right tool, and telling the truth about what you found. Thank you.

Questions (5 minutes)

IIT Madras Zanzibar

Q: Which topic matters most for Part 2?

A: Evaluation and overfitting (Weeks 6 – 7): every advanced method is judged that way.

Q: What should go in my portfolio?

A: Your mini-project and its one-page plain-language summary.

Q: Where can I learn more?

A: Han, Pei & Tong (textbook); Leskovec, Rajaraman & Ullman, Mining of Massive Datasets.

Q: When are office hours?

A: They continue until 17 November, when Part 2 begins.