

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Course Wrap-Up and Handover to Part 2

Lecture 26 · Friday 13 November 2026

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

Nine Weeks in One Picture

IIT Madras Zanzibar

Week	You learned to …	Signature result on our data
1	describe data; the KDD process	500 Zanzibar farms, 9 attributes
2	measure similarity; clean and integrate	24 missing yields filled by crop median
3	transform, reduce, warehouse	raw PCA 99.8% (scale artefact) → honest 42%
4	build cubes; OLAP	2017 drought found by drill-down
5	mine and judge patterns	rice seed → urea lift 6.8; hoe 0.91
6	classify with trees and Naive Bayes	full tree 99.9% train vs 62.9% test
7	evaluate honestly; ensembles	90% accurate lazy model helps no one
8–9	cluster, at scale	4 farm types; 31 unusual farms

What You Can Now Do

IIT Madras Zanzibar

1. **Prepare data**
   clean, integrate, transform, reduce — and justify every choice
2. **Mine it**
   patterns, classifiers, clusters — choosing the right tool
3. **Judge it**
   the right measure, honest validation, fairness by group
4. **Scale it**
   sample, summarise, prune, fewer passes, parallelise
5. **Explain it**
   plain-language reports a decision-maker can act on

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	normalization (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

Mini-Project: Final Checklist (Due Tonight)

IIT Madras Zanzibar

- ☐ Notebook runs top to bottom; outputs visible
- ☐ Every preprocessing choice justified in one sentence
- ☐ Scalers fitted on training data only
- ☐ Mining result properly evaluated (test set, thresholds, silhouette)
- ☐ Scaling and ethics paragraphs specific to YOUR data
- ☐ One-page plain-language summary; sources cited

Thank You

IIT Madras Zanzibar

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.

Stay in touch: office hours continue until 17 November; Part 2 starts with Prof. Mahindran.