Week 1 Summary: KDD Process, Attribute Types, and Summary Statistics

Z3003 Data Mining: Scaling to Large Data · 16–18 September 2026 · Dr. Innocent Nyalala

Lecture 1: Course Introduction and the KDD Process

Data mining extracts valid, novel, useful, and understandable patterns from large datasets. Within the 7-step KDD pipeline (cleaning → integration → selection → transformation → mining → evaluation → presentation), preprocessing accounts for most engineering effort. Unlike testing one classical hypothesis, scanning thousands of candidate patterns creates false positives (testing $1,000$ random rules at $\alpha = 0.05$ yields roughly 50 spurious hits), and brute-force enumeration over n items faces $2^n - 1$ subsets.

Lecture 2: Task Families, Attributes, and Naïve Bayes

Every mining task belongs to one of five families: description, association, classification/regression, clustering, or outlier analysis. Attribute types govern permissible statistics:

Attribute Type	Valid Operators	Representative Location Measure
Nominal (crop, region, ID)	$=, \neq$	Mode
Ordinal (Poor < Fair < Good)	$=, \neq, <, >$	Mode, median
Interval (°C; no true zero)	$+, -$	Mode, median, mean
Ratio (yield, size in ha)	$\times, \div$	Mode, median, mean, ratios

For asymmetric binary attributes (malaria positive, fraud flag), rare 1s matter far more than 0s.

Lecture 3: Central Tendency, Dispersion, and Single-Pass Scaling

Mean $\bar{x} = \frac{1}{n}\sum x_i$ and sample variance $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ require only three running sums $(n, \sum x_i, \sum x_i^2)$ in $O(1)$ memory and merge additively across machines. Exact medians require storing or sorting all values ($O(n)$) and must be approximated via histograms at scale. On skewed farm sizes (500 Zanzibar farms: mean 1.46 ha vs. median 0.96 ha), report median and IQR $= Q_3 - Q_1$ with outlier fences $Q_1 - 1.5\,\text{IQR}$ and $Q_3 + 1.5\,\text{IQR}$.

Worked Check: Receipts & Fraud Baseline

Five agro-dealer receipts have {hoe} in 4, {gloves} in 3, and {hoe, gloves} in 3: support = 0.60, confidence(hoe → gloves) = 0.75, and lift = $0.75/0.60 = 1.25 > 1$. On $10,000$ payments with 50 frauds, an always-negative classifier scores 99.5% accuracy yet 0 recall.