# Week 6 Summary: Decision Trees, Naive Bayes, and Pruning at Scale

Z3003 Data Mining: Scaling to Large Data · 28 Oct – 3 Nov 2026 · Dr. Innocent Nyalala

## Lecture 16: Classification I: Decision Trees

Decision trees partition labeled examples greedily using node impurity scores:

$$\text{Gini}(S) = 1 - \sum_i p_i^2, \qquad H(S) = -\sum_i p_i \log_2 p_i.$$

Information Gain $\text{IG}(S, A) = H(S) - \sum_v \frac{|S_v|}{|S|} H(S_v)$ selects the split maximizing purity (ID3 uses gain, C4.5 uses gain ratio, CART uses Gini). Plain error rate fails to reward splits that improve class balance without flipping majority predictions.

## Lecture 17: Classification II: Naïve Bayes

Assuming conditional independence given class $C$, the MAP decision rule picks the class maximizing $\log P(C) + \sum_{j=1}^{d} \log P(x_j \mid C)$. Laplace smoothing $\hat{P}(x_j = v \mid C) = \frac{\text{count}(v,C)+1}{|C|+k_j}$ prevents zero probabilities from wiping out the product, and log sums prevent floating-point underflow. Naïve Bayes trains in one additive pass across distributed chunks.

## Lecture 18: Overfitting, Pruning, and Histogram Tree Scaling

On $189,960$ Zanzibar farm-seasons ($132,972$ train / $56,988$ test), an unpruned tree reaches 100% training accuracy with $38,513$ leaves yet drops to 62.9% on test farms, whereas a depth-8 pre-pruned tree achieves 72.1% test accuracy. Post-pruning uses validation reduced-error pruning or cost-complexity pruning $\text{error}(T) + \alpha|T|$. Always reserve the test set for a single final evaluation. To scale tree learning, bin continuous features into $b$ histogram quantiles ($b - 1$ split candidates) and aggregate counts across partitions.

**Worked Check: 14-Farm Entropy Split & Reduced-Error Pruning**

For 14 farms (9 High, 5 Low, $H = 0.940$), splitting on rainfall yields High ($4H, 0L$, $H = 0$), Medium ($3H, 2L$, 0.971), and Low ($2H, 3L$, 0.971), giving IG $= 0.940 - 0.694 = 0.247$ (beating fertilizer 0.152 and irrigation 0.048). In validation pruning, if a subtree predicts $18 + 33 = 51$ of 75 farms correctly while collapsing to a single Low leaf predicts 55, collapse the branch.