**Day 19 — 15 July 2025 (Tuesday)**

**Topic: Handling Imbalanced Datasets & Model Reliability**

Spam datasets often have imbalance (more ham emails than spam). Today, we learned techniques to address class imbalance such as:

* Oversampling spam emails
* Undersampling ham emails
* TF-IDF tuning
* Adjusting classification threshold
* Stratified train-test split

We also discussed how imbalanced datasets may cause model bias (predicting most emails as ham). Learning these strategies is crucial for real-world ML systems where class imbalance is common.

**Key concept:** Balanced dataset = higher fairness + better reliability.