**🎓 Lecture: Strategies to Handle Imbalanced Data in Classification**

**📌 Problem Recap:**

In classification problems, **imbalanced data** occurs when one class (e.g., “Yes”) has significantly more samples than another (e.g., “No”). This can:

* Lead to **biased models**.
* Mislead accuracy metrics (predicting the majority class could still give high accuracy).
* Harm the **reliability and fairness** of predictions.

**🛠️ Solutions for Imbalanced Data:**

**1. Resampling Techniques**

Modify the dataset to balance class distribution.

**a. Undersampling**

* Reduce the number of majority class samples.
* Pros: Simplifies data.
* Cons: Risk of **losing important information**.

**b. Oversampling**

* Duplicate or replicate minority class instances.
* Pros: Balances classes.
* Cons: Risk of **overfitting** due to repeated data.

**c. SMOTE (Synthetic Minority Oversampling Technique)**

* Creates **synthetic (new) minority samples** by interpolating between existing ones.
* Pros: Better than basic duplication; adds variety.
* Cons: Still synthetic – not as diverse as real data.

**2. Weighted Classes**

Assign **more penalty** to mistakes on minority class.

* Adjust **class weights** in the model.
* Most algorithms (e.g., logistic regression, SVM, decision trees) support this.
* Goal: Tell the model **minority class errors matter more**.

**3. Data Augmentation**

Used especially in **image and text** datasets.

* For images: Rotate, zoom, flip, etc.
* For text: Synonym replacement, rephrasing.
* Purpose: Increase diversity in the **minority class** data without simple repetition.

**4. Algorithm-Specific Tweaks**

**a. Weighted SVM**

* Different weights for different classes during training.

**b. Adjusted Decision Trees**

* Modify how trees **split nodes**, considering class imbalance.

**🎯 Goal**

Move beyond just “accuracy.”  
Build **fair, generalizable models** that treat all classes equally and perform well across the board.

**🧪 Next Lesson Preview:**

We’ll dive into **evaluation metrics**:

* **Precision**
* **Recall**
* And others designed to measure performance on **imbalanced datasets** more meaningfully.