Handling Imbalanced Dataset:

Colab:

https://colab.research.google.com/drive/1oMNk9HE4Bzuw9jEPtlNw3sECuYLCgIYO#scrollTo=ZGnM1nTiYRjz

What is imbalanced dataset?

Imbalanced datasets have unequal class distributions, causing models to favor the majority class and ignore minority classes.

Upsampling (Oversampling)

Definition: Increase the number of samples in the minority class to balance the dataset.

Methods:

- **Random Oversampling:** Randomly duplicate minority class samples.
- **SMOTE (Synthetic Minority Oversampling Technique):** Generate synthetic samples based on feature space similarities.

Pros:

- Prevents information loss (no majority class data is removed).
- Helps the model learn minority class patterns.

Cons:

- Can lead to overfitting (duplicates or synthetic samples may not add new information).

Downsampling (Undersampling)

Definition: Reduce the number of samples in the majority class to balance the dataset.

Methods:

- **Random Undersampling:** Randomly select a subset of majority class samples.
- **Tomek Links / Edited Nearest Neighbors:** Remove majority class samples that are ambiguous or near minority samples.

Pros:

- Reduces training time.
- Prevents majority class bias.

Cons:

- May lose useful information from discarded samples.

Summary:

- **Upsampling:** Add minority samples → prevents information loss but can overfit.
- **Downsampling:** Reduce majority samples → faster and simpler but can lose data.
- Goal: Achieve a **balanced dataset** to improve model performance on minority class.