Handling imbalanced dataset using SMOTE.:

Colab:

https://colab.research.google.com/drive/1iG9RzzVhIn8MIfUlVHLDCB3Zvs-P_Zji#scrollTo=T_efwb-25zLR

SMOTE (Synthetic Minority Over-sampling Technique)

Purpose:

- To address **class imbalance** in datasets, where the minority class has far fewer samples than the majority class.
- Helps ML models perform better by **reducing bias toward the majority class**.

How it works:

1. Select a sample from the **minority class**.
2. Identify its **k nearest neighbors** (usually k=5).
3. Randomly choose one of these neighbors and generate a **synthetic sample** along the line segment connecting the original sample and the neighbor.
4. Repeat until the minority class is sufficiently oversampled.

Key Points:

- **Synthetic samples** are created, not just duplicated (avoids overfitting like simple oversampling).
- Works well with numeric features; categorical features need special handling.
- Often combined with undersampling of the majority class for better balance.

Advantages:

- Reduces bias toward majority class.
- Increases minority class representation without duplication.

Limitations:

- Can create **noisy or overlapping samples** if minority class is sparse.
- Not ideal for all types of data (categorical-heavy datasets).