Handling Class Imbalance with Oversampling and Explainable AI

This is a group project, and each team may have up to 5 students. For fair and unbiased evaluation, all teams must work on the *same* problem (different from what we discussed in class earlier). You are required to complete the assignment within 2 days. The evaluation will take place on Tuesday. Please provide the details of all team members in the form given below.

Link to the Google Form to mention your team details:
https://forms.gle/Kp4pfuVVtgnywxwTA

Problem Overview

You will work with a binary classification problem where the dataset has a clear class imbalance (minority class ≤ 20% of total samples). The dataset may be tabular, medical, image-based, or any other domain, as long as it is feasible to process within the given time.

Your objectives are to:

- Build a baseline model on the imbalanced dataset.
- Apply an advanced oversampling technique based on a research paper.
- Train and compare the improved model.
- Use Explainable AI (XAI) methods to demonstrate how the synthetic samples influence the model's decision-making.

Tasks to be Completed

1. Dataset Selection & Problem Framing (5 Marks)

- Select a freely available binary classification dataset with significant class imbalance.
- Clearly describe:
  - Dataset source
  - Class distribution
  - Features and target
- Justify why this dataset is appropriate for imbalance handling.

2. Oversampling Technique Implementation (10 Marks)

- Build a baseline model **without oversampling** and evaluate it using appropriate metrics
- Choose an **advanced oversampling technique** (example: SMOTE variant, GAN-based oversampling, ADASYN, recent algorithm from research literature).
- Implement or adapt the method and **cite the original research paper**.
- Train a new model on the oversampled dataset.
- Compare performance before and after oversampling using the right metrics.
- Present the improvement (or lack of improvement) clearly.

3. Explainable AI Analysis (5 Marks)

Use an Explainable AI technique to analyze the model trained on oversampled data.

- How the model treats synthetic minority samples compared to real minority samples :
    - Do they have similar feature distributions?
    - Do they occupy similar latent representations?
    - Do they fall within the same decision boundary?
- Why does the classifier predict the synthetic samples as a minority class?