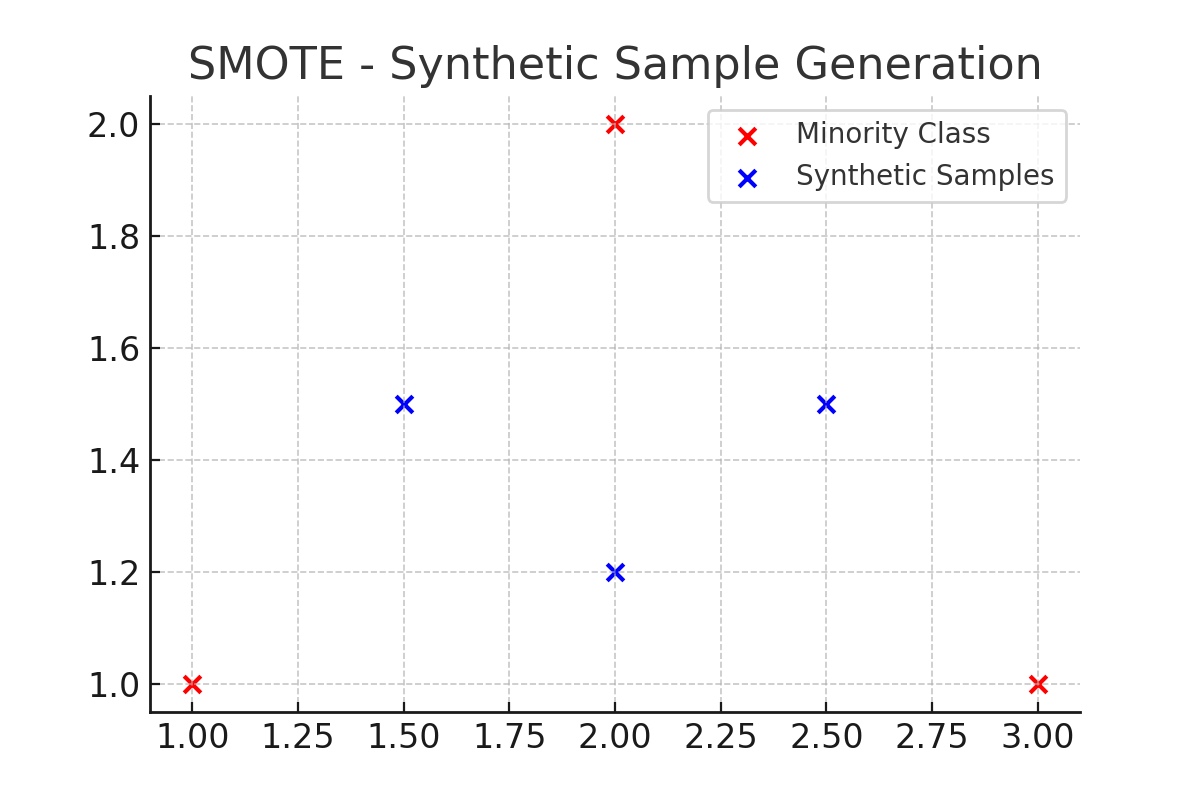
Class Imbalance (CI) Techniques: Analysis and Performance Impact

# 1. Explanation of Selected CI Techniques

## SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE is an oversampling technique that generates synthetic samples for the minority class by interpolating between existing minority class examples. It helps to balance the dataset by increasing the number of samples in the minority class. The synthetic samples are generated by taking each minority class sample and introducing synthetic examples along the line segments joining the k-nearest neighbors of the minority class.



## Random Under-Sampling

Random Under-Sampling reduces the number of majority class samples by randomly selecting a subset of the majority class. This technique aims to balance the class distribution by decreasing the size of the majority class. However, this method can lead to the loss of important information from the majority class, which may impact model performance.

## SMOTEENN (Combination of SMOTE and Edited Nearest Neighbors)

SMOTEENN is a hybrid approach that combines SMOTE and Edited Nearest Neighbors (ENN). After generating synthetic samples using SMOTE, the ENN algorithm is applied to remove noisy samples from the dataset. This technique helps to improve the quality of the synthetic samples by removing samples that are likely to be incorrectly classified.

# 2. Impact of CI Solutions on Classification Performance

The impact of each CI solution on classification performance is assessed by comparing the ROC-AUC scores of various models trained on the original imbalanced dataset and the balanced dataset created using each CI technique.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Baseline ROC-AUC | SMOTE ROC-AUC | Under-sampling ROC-AUC | SMOTEENN ROC-AUC | Comment |

# 3. Impact of Algorithms on CI Solutions

The performance of CI solutions is compared across different algorithms to determine if certain models are more sensitive to changes in the class distribution. The results are compared for both baseline and CI-based models.