Sampling(Under, Over)

데이터 분석 전처리 판다스

Undersampling

In undersampling, we pull all the rare events while pulling a sample of the abundant events in order to equalize the datasets.

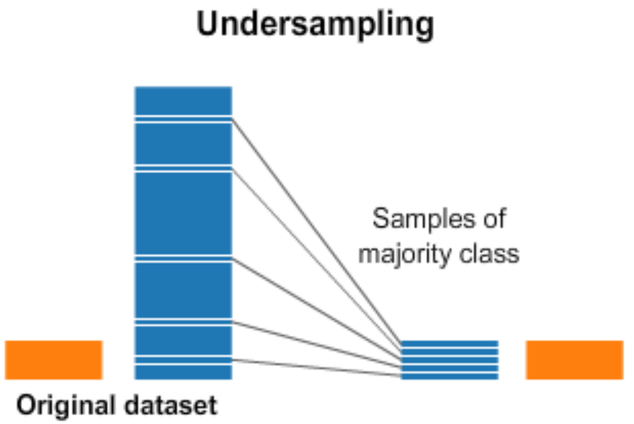
Original

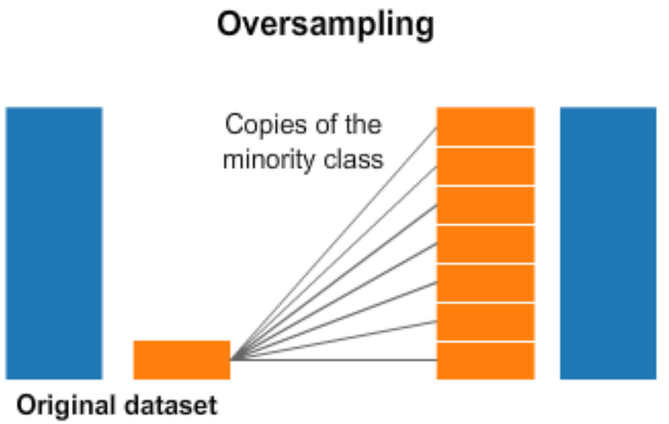
Abundant dataset

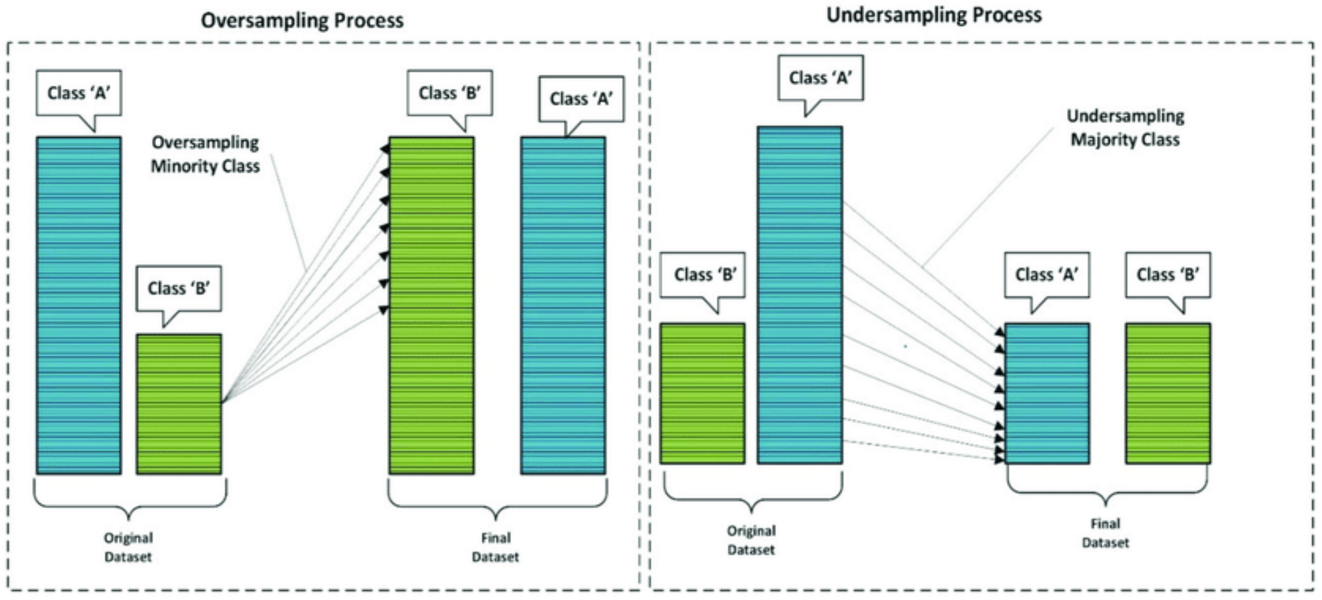
Rare dataset

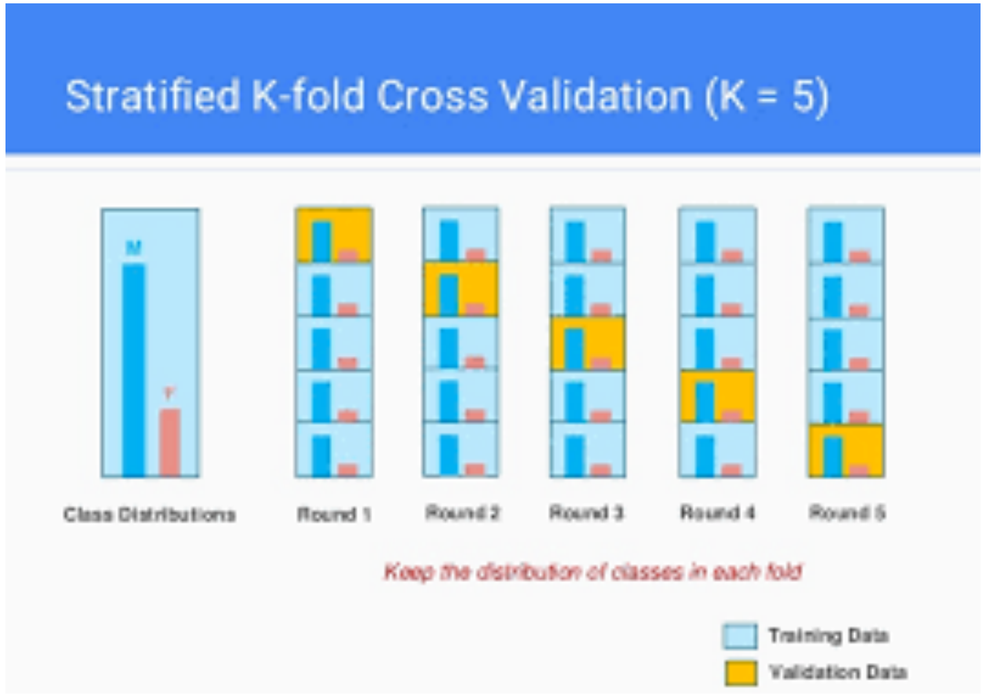
Oversampling

These methods can be used separately or together;one is not better than the other. Which method a data scientist uses depends on the dataset and analysis.

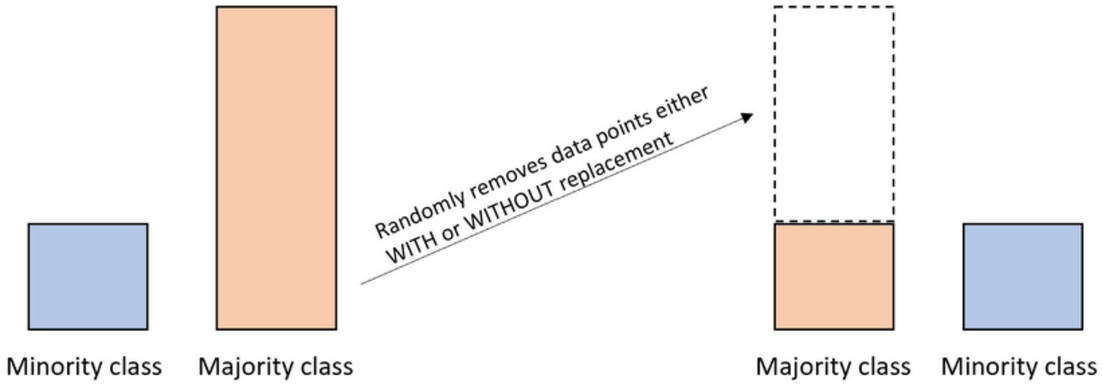

Undersampling
Samples of
majority class
Original dataset


Oversampling
Copies of the
minority class
Original dataset

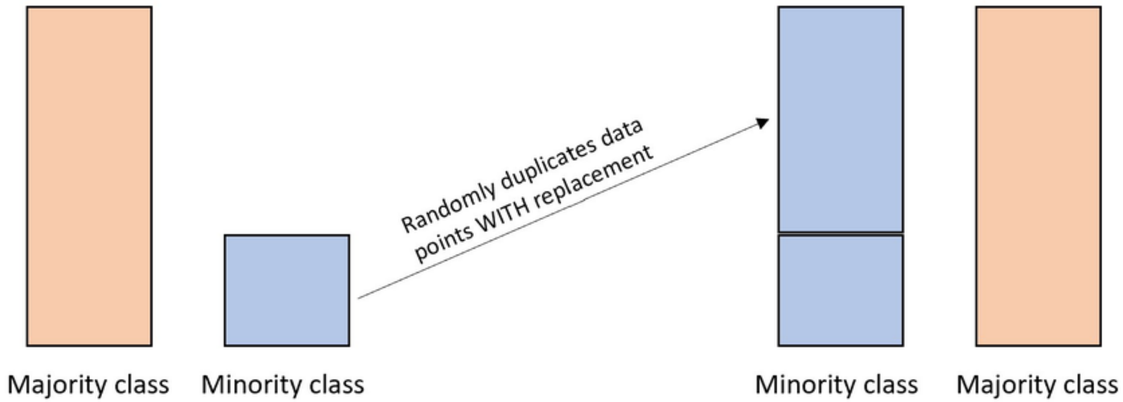

Oversampling Process
Class 'A'
Class 'B'
Oversampling
Minority Class
Original
Dataset
Class 'B'
Class 'A'
Final
Dataset
Undersampling Process
Class 'A'
Undersampling
Majority Class
Class 'B'
Original
Dataset
Class 'A'
Class 'B'
Final
Dataset

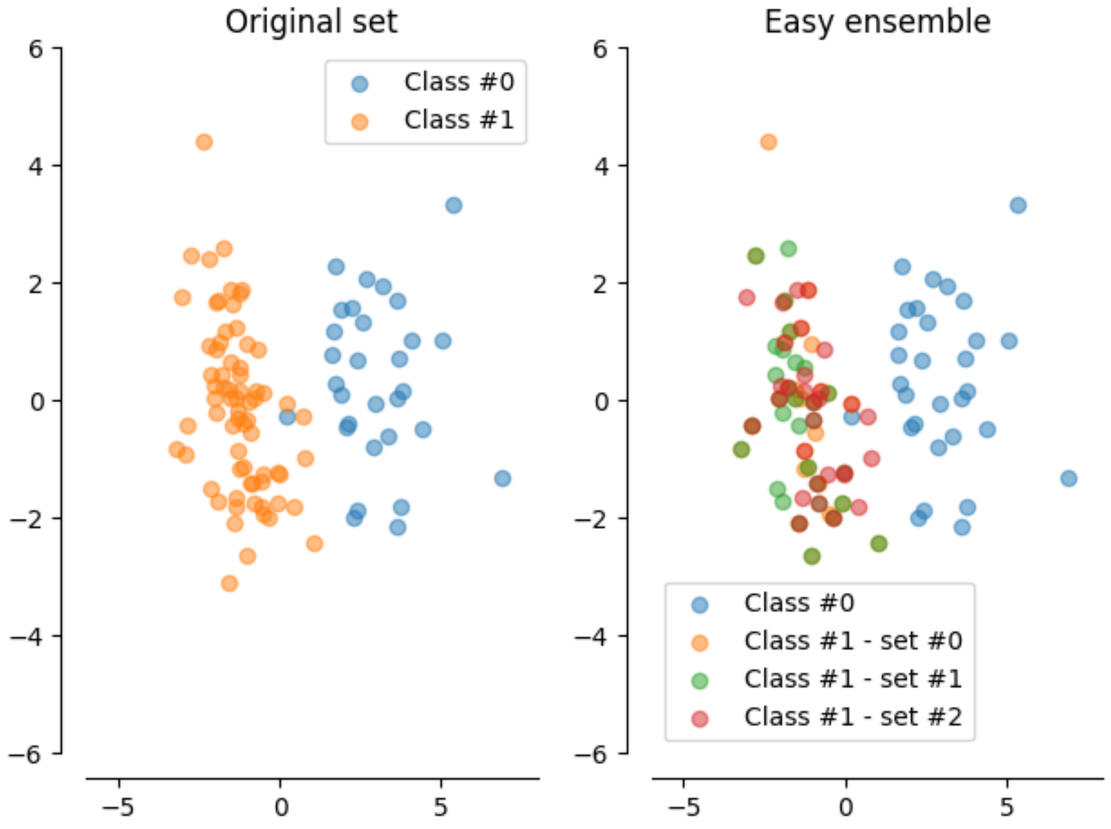

Stratified K-fold Cross Validation (K = 5)
M
F
Class Distributions
Round 1
Round 2
Round 3
Round 4
Round 5
Keep the distribution of classes in each fold
Training Data
Validation Data

Random Under Sampling



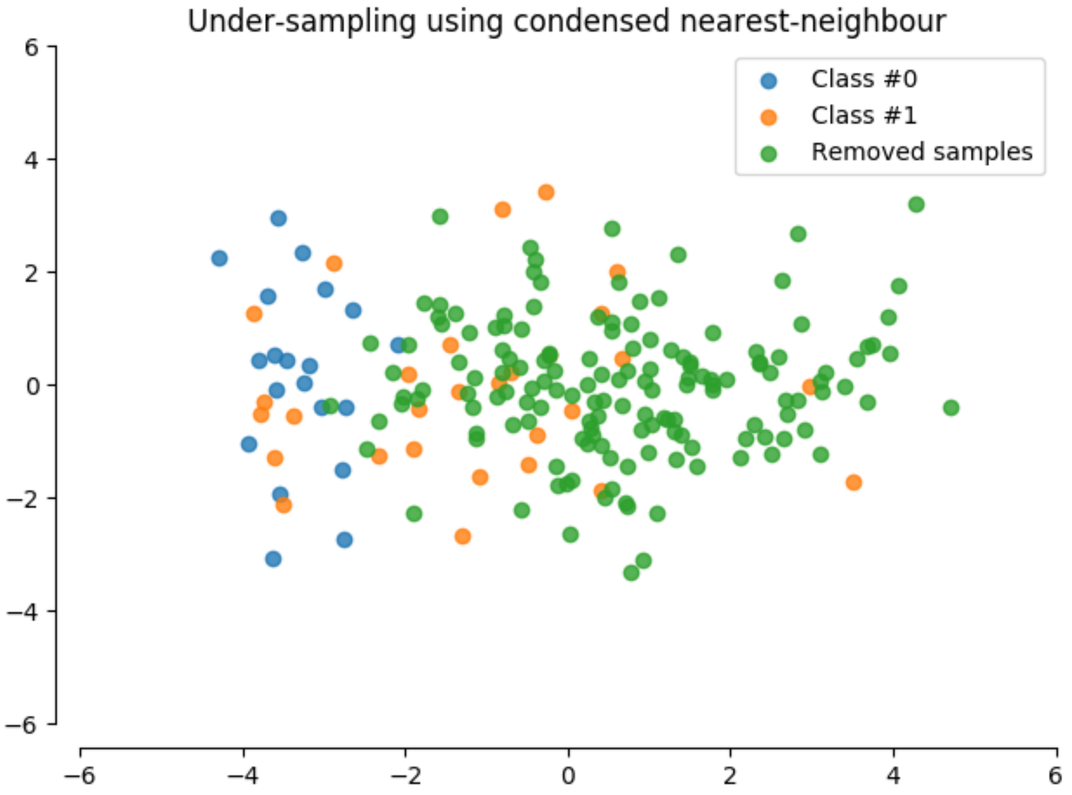
Random Over Sampling



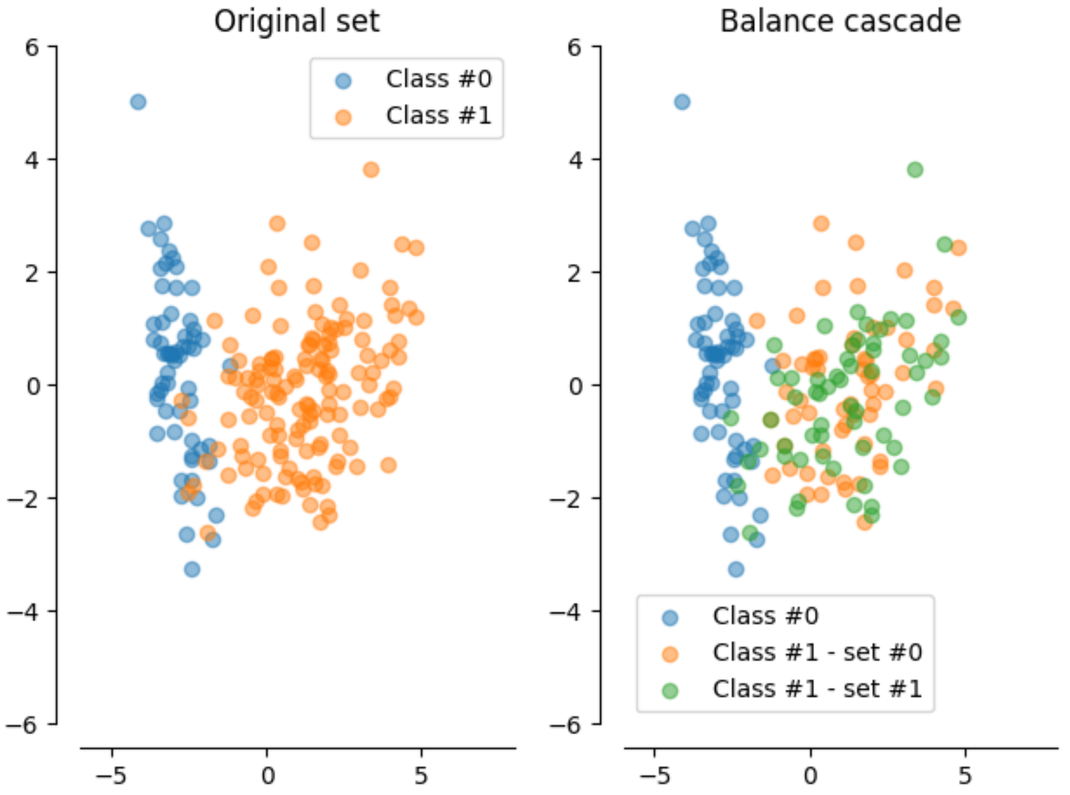


큰 비중 클래스 N개의 작은 비중의 클래스와 동일한 크기 데이터셋 분리
9 : 1 비중 -> 1번 부터 9번까지 데이셋 분리

Condensed Nearest-Neighbor



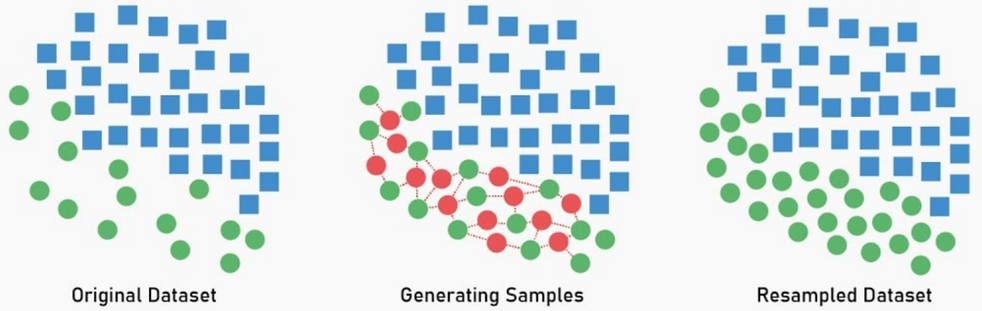
Balance Cascade

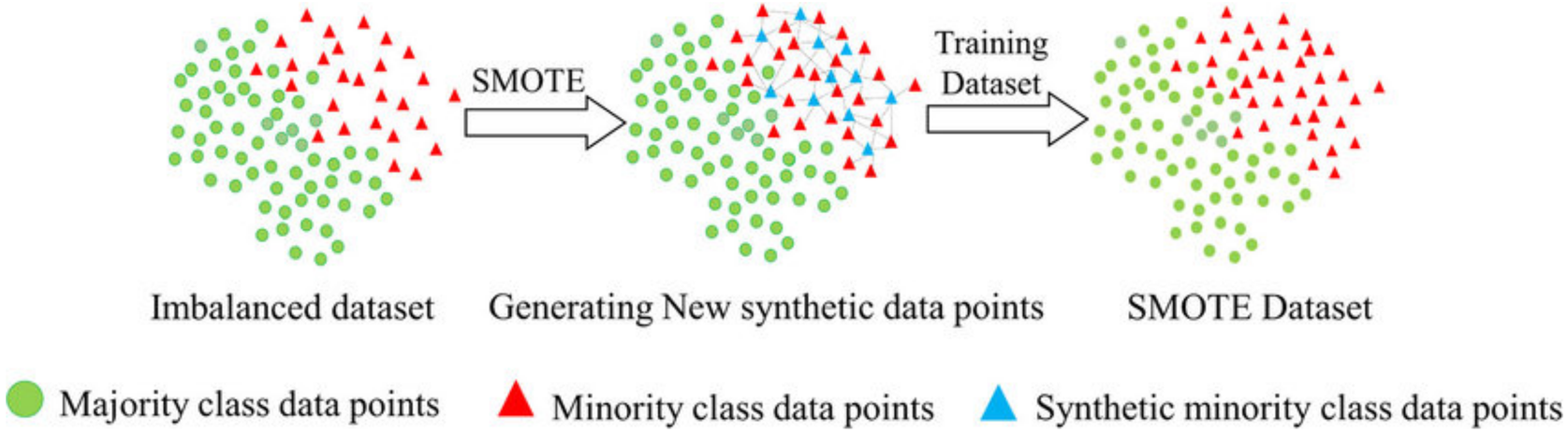


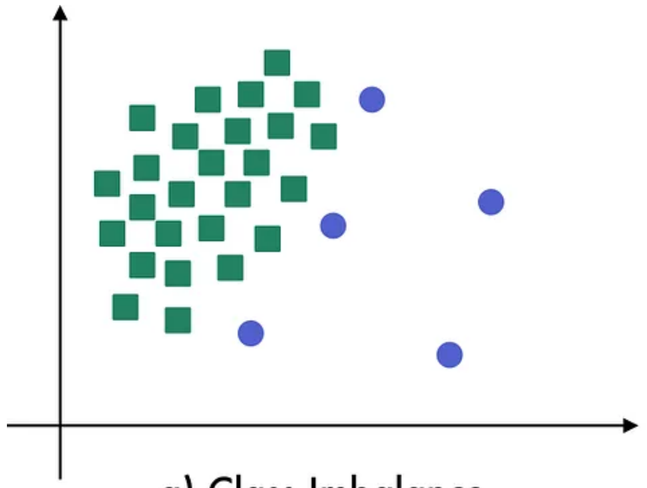
SMOTE

HANDLE IMBALANCED DATASET

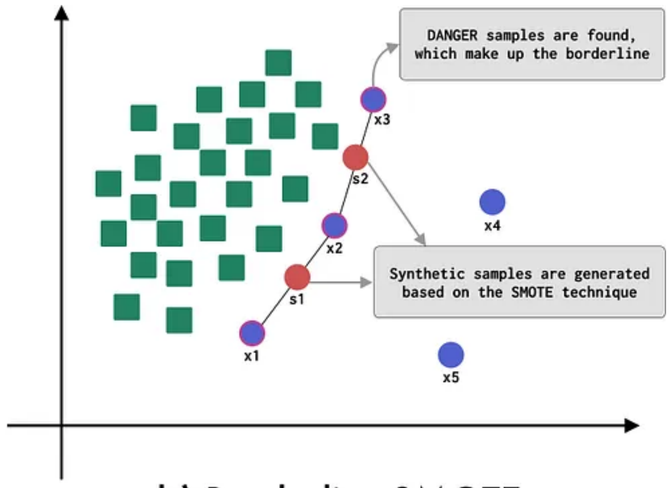
Synthetic Minority Oversampling Technique



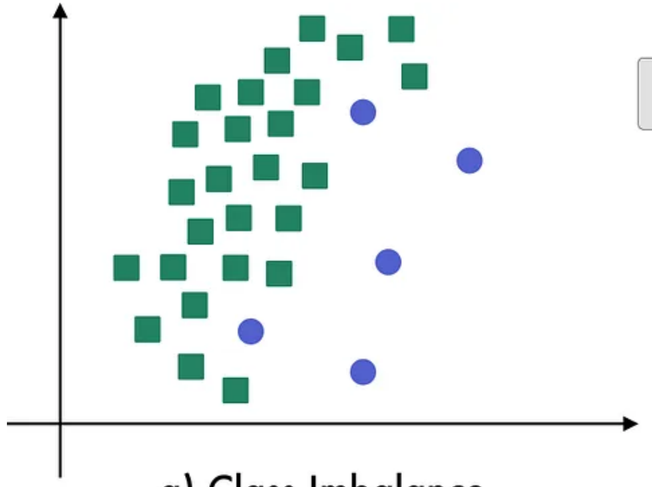




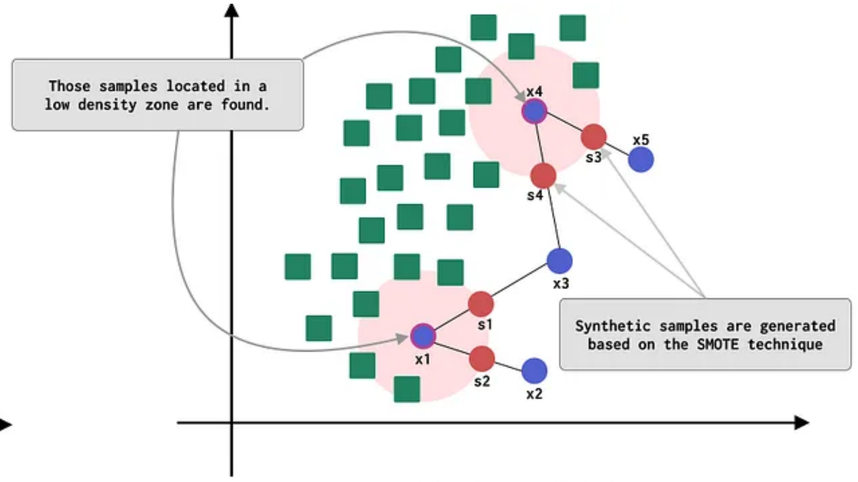
a) Class Imbalance



b) Borderline-SMOTE



a) Class Imbalance



b) ADASYN