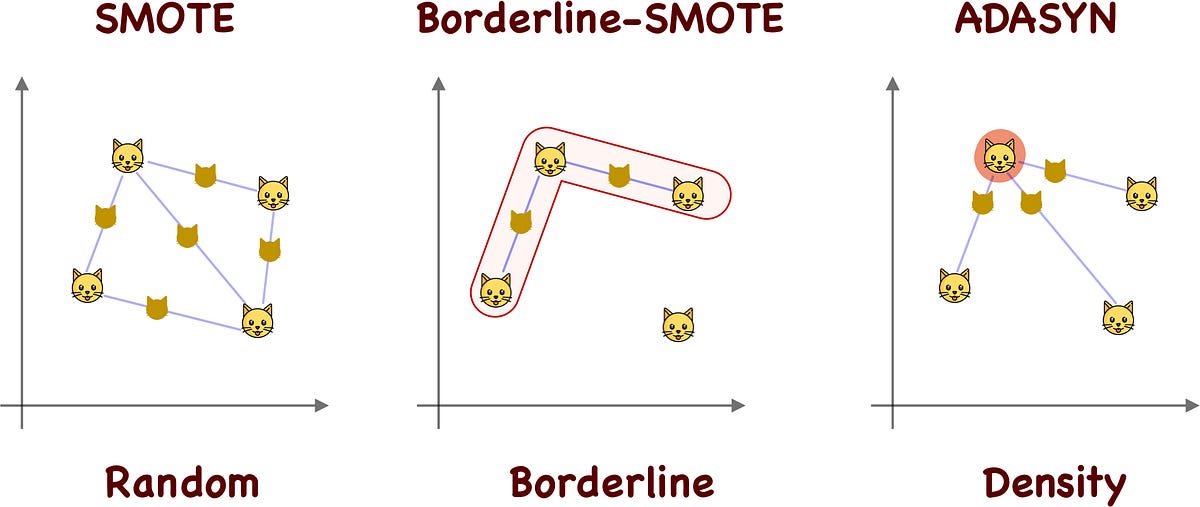
# **Questions/Answers**

# **Explain each selected CI technique with diagram, if necessary, in your own words**

I Select three different CI techniques mentioned in project description:

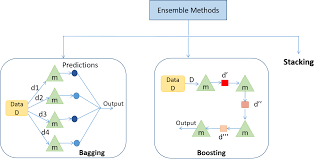
## **Oversampling with SMOTE (Synthetic Minority Over-sampling Technique)**

SMOTE is a oversampling technique designed to address class imbalance. It works by generating synthetic samples for the minority class by interpolating between existing minority class samples. Unlike traditional oversampling which simply duplicates existing instances SMOTE creates new synthetic samples by interpolating between existing minority class instances. This involves selecting two or more similar instances and generating a new sample that lies somewhere between them in the feature space. By doing like that SMOTE balances the dataset more effectively and helps the model learn better from the minority class and by reducing the risk of overfitting.



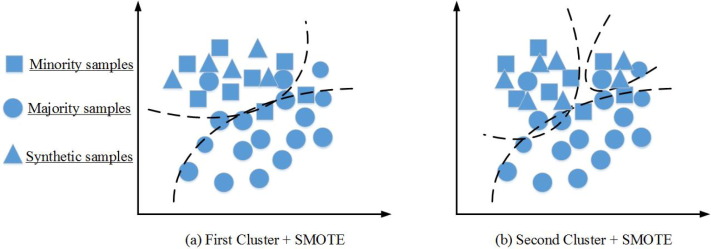
## **Ensemble Methods and Weights**

Ensemble methods combine multiple weak learners to create a strong learner which can be effective in handling class imbalance. While weights are also provided so that the less represented data also gets more focus. Techniques such as bagging and boosting are commonly used. In bagging multiple models are trained on different subsets of the data and their predictions are averaged which reduces overfitting and increases accuracy. In Random Forest is a well-known example of this method. In boosting models are trained sequentially with each model correcting the errors of the previous one thereby improving overall performance. Assigning different weights to models or samples can emphasize harder to predict instances further enhancing model robustness and accuracy.



## **Clustering-based Oversampling**

This approach combines the SMOTE technique with KMeans clustering to generate synthetic samples. It generates synthetic samples for the minority class based on clusters within the data. The process begins by clustering the minority class data into several groups. Synthetic samples are then generated within these clusters ensuring that the new samples reflect the distribution and structure of the original minority class. This method maintains the underlying structure of the minority class data and prevents overfitting by ensuring that synthetic samples are similar to real samples within each cluster. It provides a more nuanced approach to balancing datasets preserving the natural characteristics of the minority class while improving model performance.



**What is the impact of each CI solution on the classification performance (compare with baseline)?**

The impact of each class imbalance solution on classification performance is notable. Initially, the baseline models achieved decent accuracy, exceeding 80 percent. However, they exhibited a bias towards the majority class and lacked generalization. Upon implementing the cIass imbalance techniques the models demonstrated improved generalization and performance compared to the baseline models as evidenced by the discernible enhancements observed in both the confusion matrix and classification report. As per results of different dataset i can conclude like that each class imbalance technique resulting in more reliable and accurate classification models than before applying technique.

**Does the performance of a CI solution get impacted significantly by the choice of different algorithms (compare for both baseline and CI-based)?**

Indeed, the choice of different algorithms significantly impacts the performance of class imbalance (CI) solutions. In particular techniques like SMOTE and clustering-based approaches demonstrated substantial improvements in performance compared to both the baseline and other CI methods. Conversely ensemble methods and weight-based approaches showed only marginal enhancements over the baseline models. This suggests that the effectiveness of CI solutions can vary depending on the algorithm used with some algorithms being more conducive to addressing class imbalance than others. Overall i have seen the performance improvement of classification algorithms with the help of these CI techniques.

***NOTE:*** *I have mentioned the conclusion of each CI technique on different datasets at the end of each notebook*.

***Thanks***