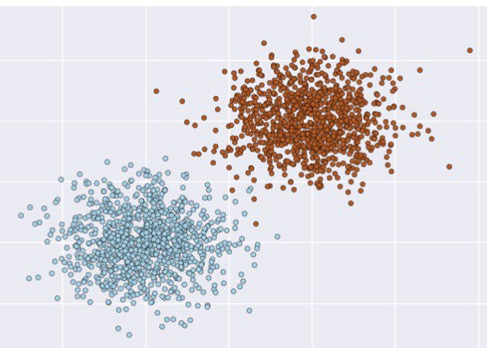
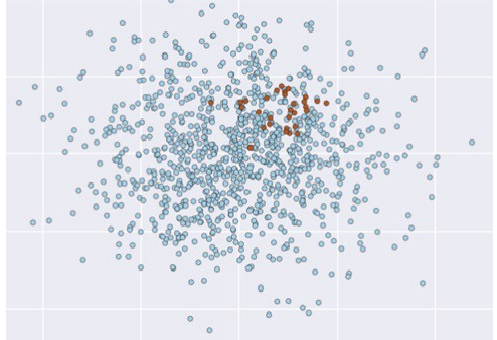
**Imbalanced Dataset**

**UnderSampling and OverSampling**

A dataset is imbalanced if at least one of the classes constitutes only a very small minority.The issue of class imbalance can result in a serious bias towards the majority class, reducing the classification performance and increasing the number of false negatives.

Balanced Dataset Imbalanced Dataset

**Ways to Handle Imbalance data :**

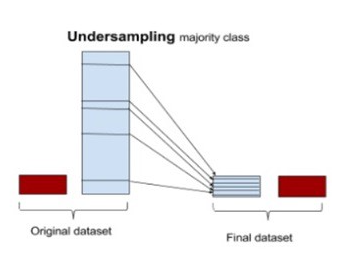
1. Using more appropriate error matrix
2. Resampling strategies (Data Level )
3. Algorithmic Techniques

**Resampling strategies (Data Level ):**

In re-sampling of data either we reduce the proportion of dominant class which is **under-sampling** or we increase proportion of minority class which is called as **oversampling.**

However most successful approaches uses both oversampling and under-sampling together.

**UNDER-SAMPLING :**



There are many ways of under-sampling. Out of which some are mentioned below:

1. Random under-sampling
2. Cluster centroid Under-sampling
3. Tomek Link

**Random under-sampling** :Method works by randomly choosing the samples from dominant class.

There are two major drawbacks of these technique:

Major drawback of this technique is that we eliminate samples randomly, which may lead to loss of potential information .

**Cluster centroid under-sampling :**clusters of majority class and replace that cluster with centroid of that cluster. So we undersample majority class by forming clusters and replacing it with cluster centroids.

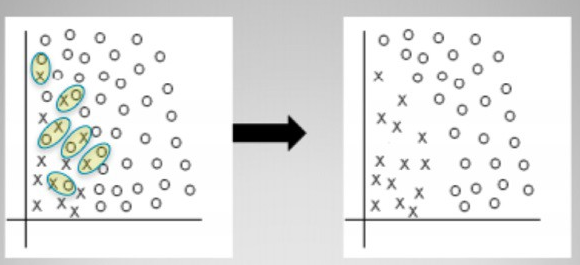
For example:

Majority class : 100 samples

Minority class : 10 samples

Here , in this case we can form 10 clusters of majority class and replace 100 points with 10 data points i.e by cluster centroid.

## Tomek links:Tomek link pair has two opposite class data points who are their own nearest neighbors. Main idea is to separate minority and majority class .



Suppose ,

d(A,B) : distance between two data points A & B

Then, a(A,B) is Tomek link if and only if

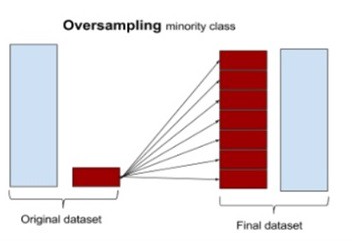
There is no such point ‘C’ , such that,

d(A,C) < d(A<B) or d(B,C) < d(A,B)

If pair of samples form tomek link then either one of the sample is noise or both are placed at border.

As under-sampling technique we eliminate majority class point , while as part of data mining we eliminate both points.

**Over-sampling:**



**Some of over- sampling techniques are mentioned below:**

1.Random Oversampling:

2 .SMOTE

**Random Oversampling:**

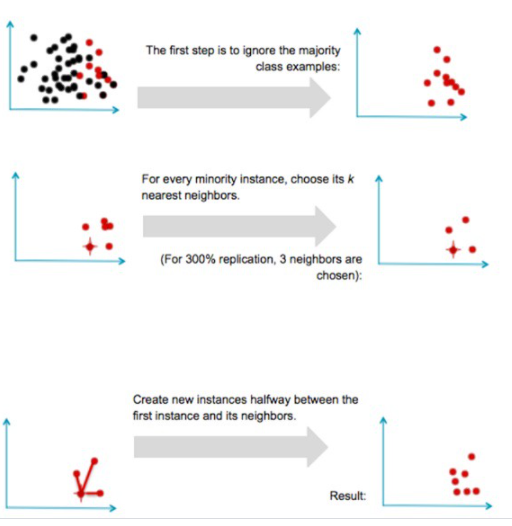
In random oversampling technique we replace the samples with existing minority samples.

In simple terms we can say that we just do multiplication of existing Minority class.

Drawback of this technique us that , this technique is highly prone to over fitting.

**SMOTE(Synthetic Minority Oversampling Technique):**

* In this technique we calculate difference between sample under consideration and its nearest neighbors. Once the distance is calculated we multiply that with the number between 0 and 1.
* We add it to sample under consideration.
* Which gives us new sample point for minority class .
* Depending upon the amount of oversampling required , neighbours from k-NN are randomly chosen.



**3. Algorithmic approach:**

1. Cost sensitive approach
2. Choice of algorithm

****Cost sensitive training:****

* **It is called as penalised training. It can be done by penalizing for wrong prediction in minority class.**
* **It can be done by customizing your error matrix.**

**Example:**

**(10\*False –ve + 1\* false +ve)/6**

****Choice of algorithm:****

* **Ensemble methods are found to good for handling imbalance data. **Random forest** is found to be good at handling imbalance data.**
* **sklearn’s implementations of these algorithms provides option to handle imbalanced data-set by setting the **class\_weight** parameter.**