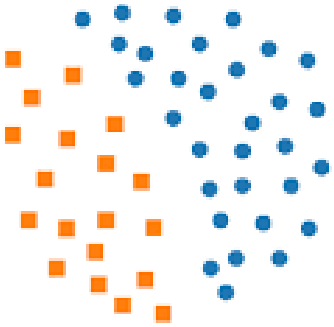
## What are Balanced and Imbalanced Datasets?

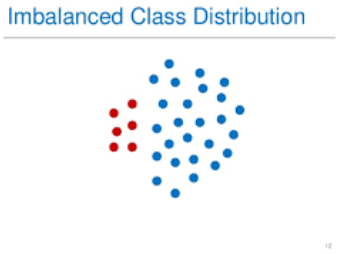
**Balanced Dataset: —** Let's take a simple example if in our data set we have positive values which are approximately same as negative values. Then we can say our dataset in balance.

Consider Orange color as a positive values and Blue color as a Negative value. We can say that the number of positive values and negative values in approximately same.



Balance Dataset

**Imbalanced Dataset: —** If there is the very high different between the positive values and negative values. Then we can say our dataset in Imbalance Dataset.




Imbalance Dataset

## Techniques to Convert Imbalanced Dataset into Balanced Dataset:

Imbalanced data is not always a bad thing, and in real data sets, there is always some degree of imbalance. That said, there should not be any big impact on your model performance if the level of imbalance is relatively low.

Now, let's cover a few techniques to solve the class imbalance problem.

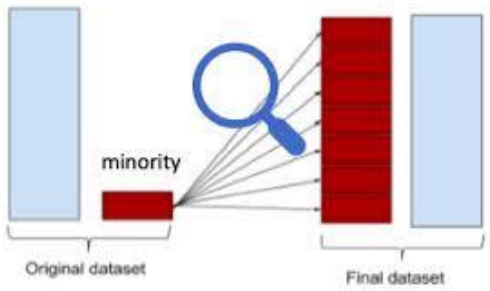
**1 — Use the right evaluation metrics:** Evaluation metrics can be applied such as:

- **Confusion Matrix:** a table showing correct predictions and types of incorrect predictions.
- **Precision:** the number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier's exactness. Low precision indicates a high number of false positives.
- **Recall:** the number of true positives divided by the number of positive values in the test data. Recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier's completeness. Low recall indicates a high number of false negatives.
- **F1-Score:** the weighted average of precision and recall.

**2 — Over-sampling (Up Sampling):** This technique is used to modify the unequal data classes to create balanced datasets. When the quantity of data is insufficient, the oversampling method tries to balance by incrementing the size of rare samples.

(Or)

Over-sampling increases the number of minority class members in the training set. The advantage of over-sampling is that no information from the original training set is lost, as all observations from the minority and majority classes are kept. On the other hand, it is prone to over fitting.




**Advantages:**

- No loss of information
- Mitigate over-fitting caused by oversampling.

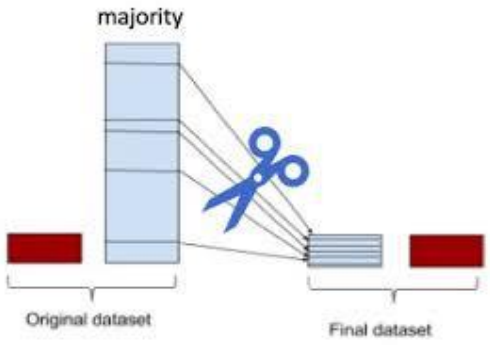
**Disadvantages:**

- Over-fitting.

**3 — Under-sampling (Down Sampling):** Unlike oversampling, this technique balances the imbalance dataset by reducing the size of the class which is in abundance. There are various methods for classification problems such as cluster centroids and Tomek links. The cluster centroid methods replace the cluster of samples by the cluster centroid of a K-means algorithm and the Tomek link method removes unwanted overlap between classes until all minimally distanced nearest neighbors are of the same class.

(Or)

Under-sampling, on contrary to over-sampling, aims to reduce the number of majority samples to balance the class distribution. Since it is removing observations from the original data set, it might discard useful information.




**Advantages**

- Run-time can be improved by decreasing the amount of training dataset.
- Helps in solving the memory problems

**Disadvantages**

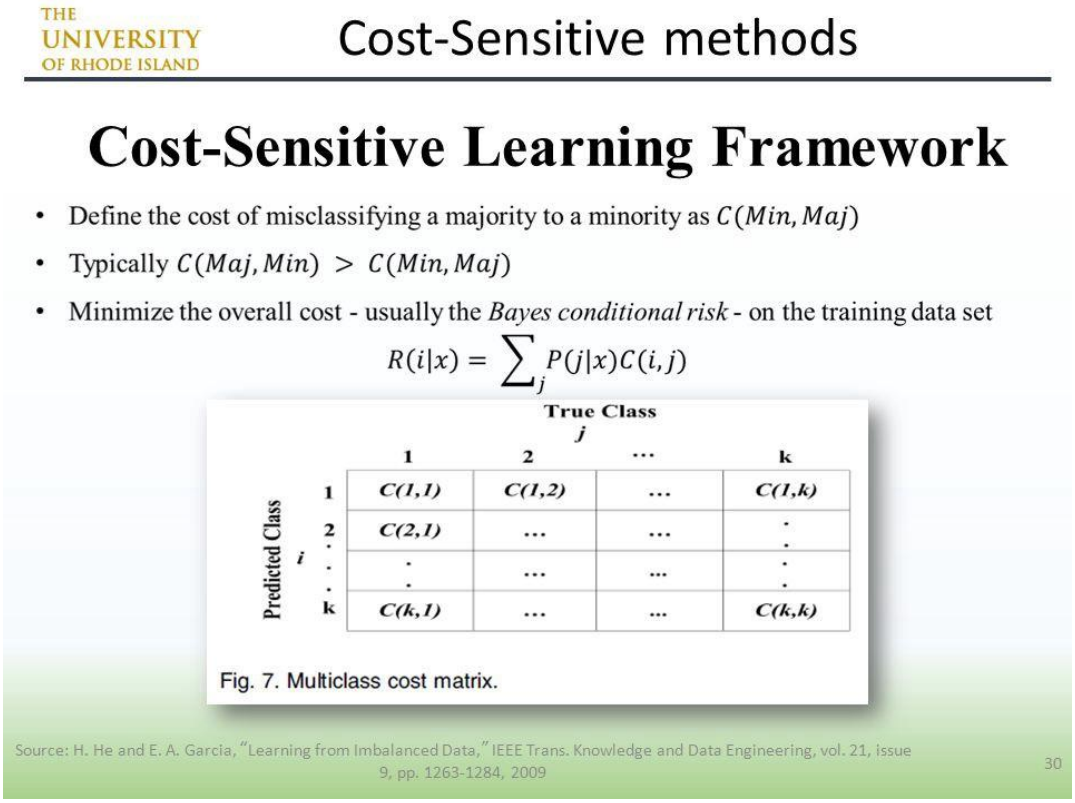
- Losing some critical information

**4 — Feature selection:** In order to tackle the imbalance problem, we calculate the one-sided metric such as correlation coefficient (CC) and odds ratios (OR) or two-sided metric evaluation such as information gain (IG) and chi-square (CHI) on both the positive class and negative class. Based on the scores, we then identify the significant features from each class and take the union of these features to obtain the final set of features. Then, we use this data to classify the problem.

Identifying these features will help us generate a clear decision boundary with respect to each class. This helps the models to classify the data more accurately. This performs the function of intelligent subsampling and potentially helps reduce the imbalance problem.

**5 — Cost-Sensitive Learning Technique:**

The Cost-Sensitive Learning (CSL) takes the misclassification costs into consideration by minimizing the total cost. The goal of this technique is mainly to pursue a high accuracy of classifying examples into a set of known classes. It is playing as one of the important roles in the machine learning algorithms including the real-world data mining applications.

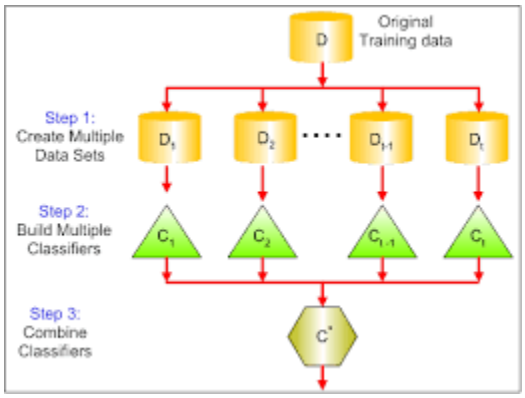



**Advantages:**

- This technique avoids pre-selection of parameters and auto-adjust the decision hyperplane.

## 5 — Ensemble Learning Techniques:

The ensemble-based method is another technique which is used to deal with imbalanced data sets, and the ensemble technique is combined the result or performance of several classifiers to improve the performance of single classifier. This method modifies the generalisation ability of individual classifiers by assembling various classifiers. It mainly combines the outputs of multiple base learners. There are various approaches in ensemble learning such as Bagging, Boosting, etc.




### Advantages

- This is a more stable model.
- The prediction is better.

### Conclusion:

Imbalanced data is one of the potential problems in the field of data mining and machine learning. This problem can be approached by properly analyzing the data. A few approaches that help us in tackling the problem at the data point level are under-sampling, oversampling, and feature selection. Moving forward, there is still a lot of research required in handling the data imbalance problem more efficiently.