Outline

- **Imbalanced data in machine learning and its outcome**
- **SMOTE (Synthetic Minority Oversampling Technique) method**
- **How to install a package in Weka**

https://www.saedsayad.com/zeror.htm

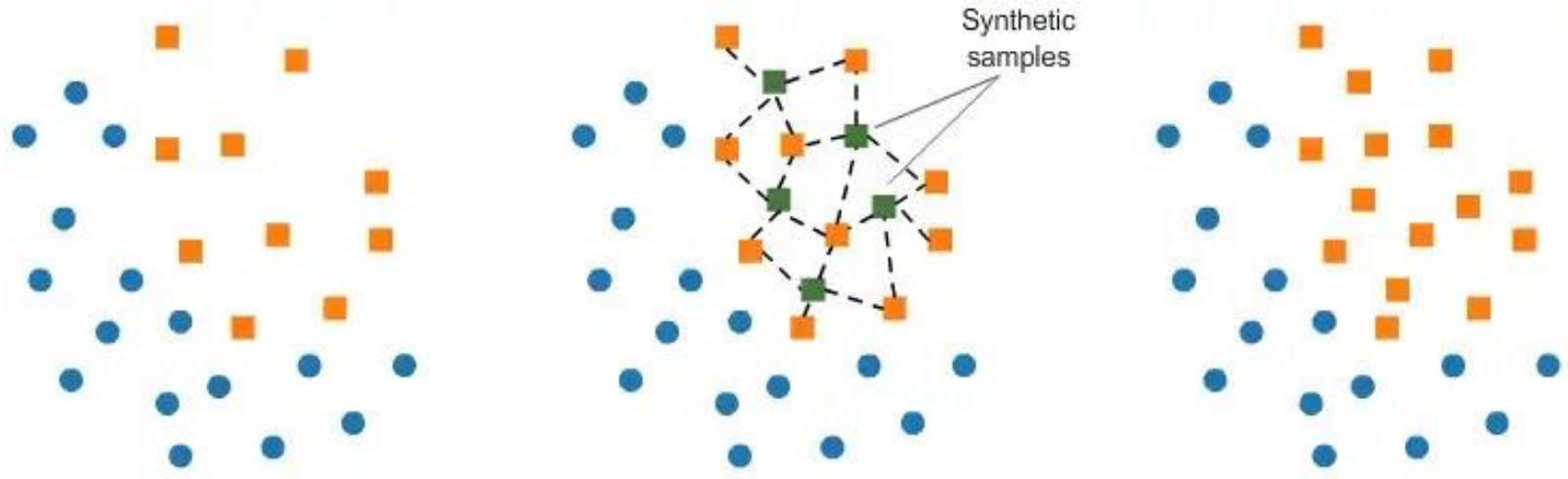
Imbalanced data

Imbalanced data is data in which observed frequencies are very different across the different possible values of a categorical variable. Basically, there are many observations of some type and very few of another type.

Using imbalanced data, we may have a model that appears very accurate for predicting the category that is over-presented but is useless for predicting the category that is under-presented.

https://towardsdatascience.com/smote-fdce2f605729

SMOTE
(Synthetic Minority Oversampling Technique)



SMOTE finds out ‘k’ nearest neighbors of a data point in the minority class. After the nearest data points have been identified, SMOTE then creates some synthetic data points on the lines joining the primary point and the neighbors so that these data points share the similar features/characteristics of the other minority data points.