# Oversampling methods introduction

1. Oversampling methods

* Process of increasing number of samples from the minority class
* until a desired balancing ratio

1. Random extraction vs sample generation

* Sample extraction: extract samples at random from minority class
* Random
* Sample generation: creates new samples from existing ones
* SMOTE and its variants
* ADASYN

A close-up of a sample

Description automatically generated

# Random oversampling

1. Overview

* Extracts obs at random from the minority class until a certain balancing ratio is reached
* Naïve technique

1. Duplication

* Extracts observations at random from the minority class, with replacement
* Duplicates samples from the minority class
* Increase likelihood of overfitting

1. Implementation

A black text on a white background

Description automatically generated

1. Multiclass

* 1 minority class: sampling\_strategy = ‘auto’
* Multiple minority classes: sampling\_strategy = ‘not majority’ or dict

Random oversampling with smoothing

1. Overview

* Extracts obs at random from the minority class and compounds its value with some noise
* The noise is informed by the class distribution
* Creates new examples -> avoids data duplication
* We can choose how disperse we want the new samples -> shrinkage factor (arbitrary)

1. Procedure

* Take the minority samples and determine their distribution -> Std of each variable
* Extract a value at random from N ~(0, 1)
* Extract at random 1 observation
* Add the noise (with the shrinking factor) = shrinking factor x std x random value

1. Implementation

A screenshot of a computer code

Description automatically generated

A graph of a number of dots

Description automatically generated with medium confidence