Apprentissage de classes déséquilibrées

HAX907X - Apprentissage statistique

SAWADOGO Kader
GERMAIN Marine
LABOURAIL Célia
MARIAC Damien

Université Montpellier
Département de Mathématique

17 octobre 2025

Sommaire

Sommaire

Les classes déséquilibrées

Problème : difficulté à prédire la classe minoritaire

$\Rightarrow$ Le modèle a tendance à ignorer cette classe.

Exemple général

- 99% vs 1%.
- Un modèle naïf prédit la classe majoritaire à une précision de 99 %.
- Mauvais modèle.

Problématique scientifique

Comment atténuer le déséquilibre des classes pour améliorer la performance réelle du modèle ?

Sommaire

Random Over-simpling

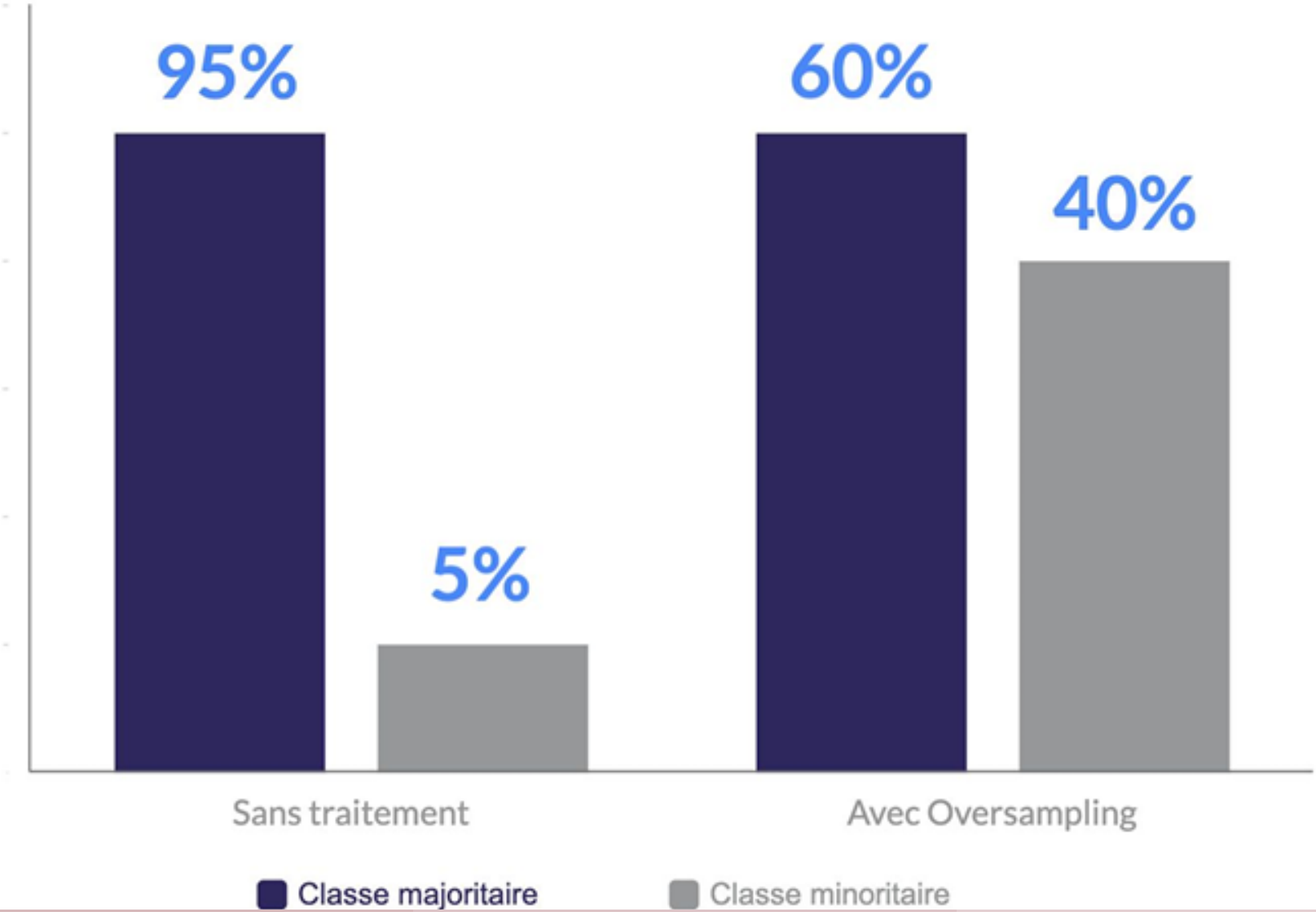


Table – Jeu de données après sur-échantillonnage (ROS)

x	label	source
1	0	original
2	0	original
3	0	original
4	0	original
5	0	original
6	0	original
7	0	original
8	1	original
9	1	original
10	1	original
8	1	duplicated
9	1	duplicated
10	1	duplicated
8	1	duplicated

Random Under-simpling

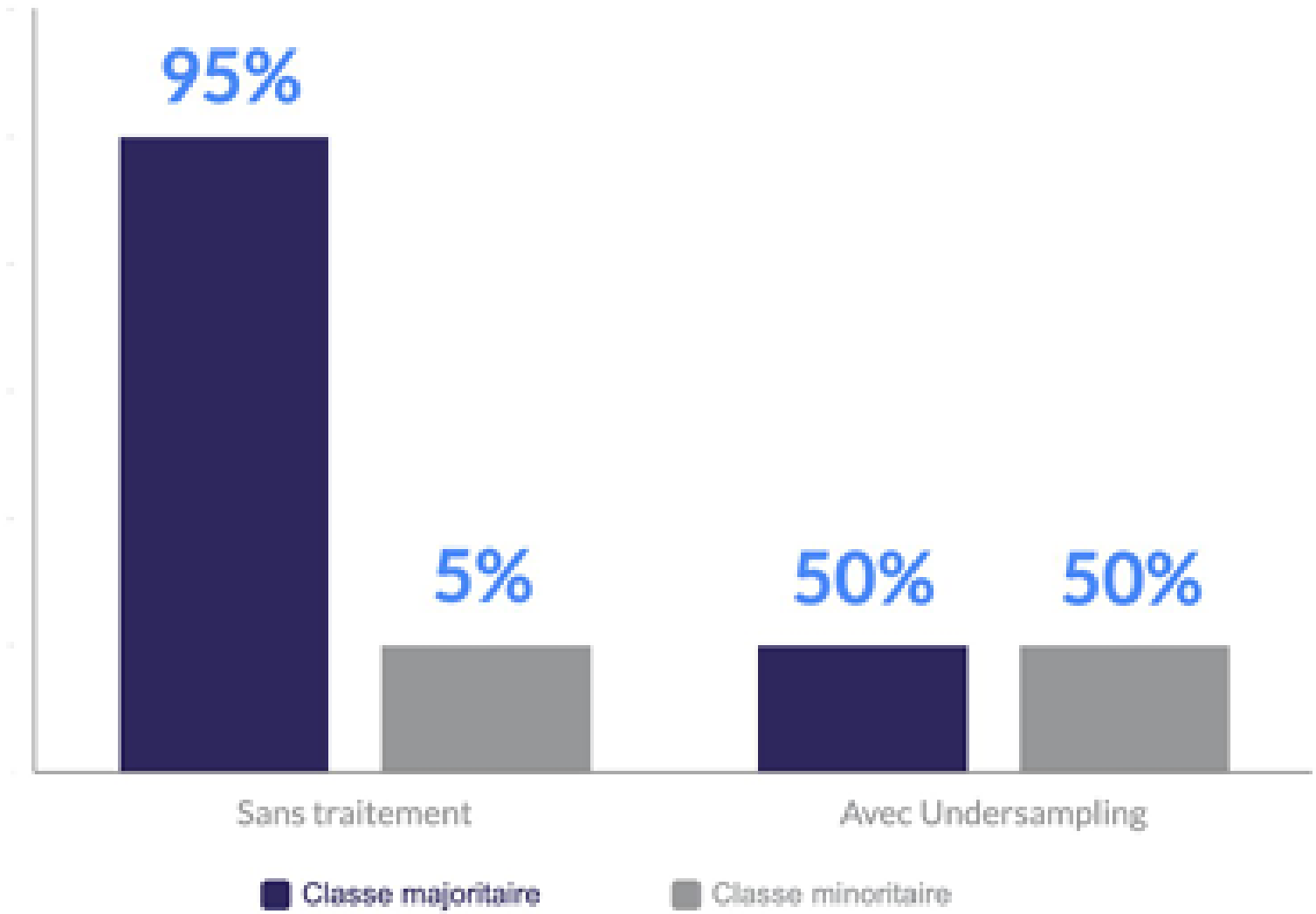


Table – Jeu de données après sous-échantillonnage (RUS)

x	label	source
1	0	supprimé
2	0	suprimé
3	0	suprimé
4	0	original
5	0	original
6	0	original
7	0	original
8	1	original
9	1	original
10	1	original

SMOTE : Synthetic Minority Over-sampling Technique

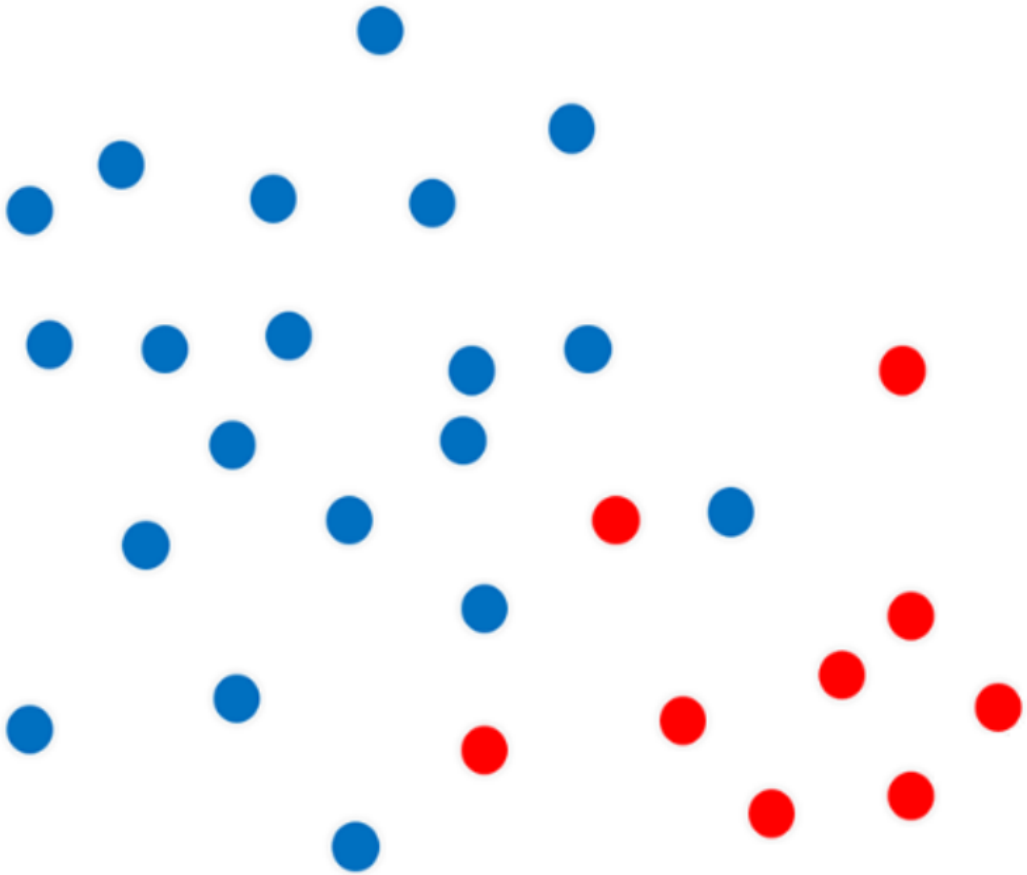


Figure – Schéma de SMOTE

SMOTE : Synthetic Minority Over-sampling Technique

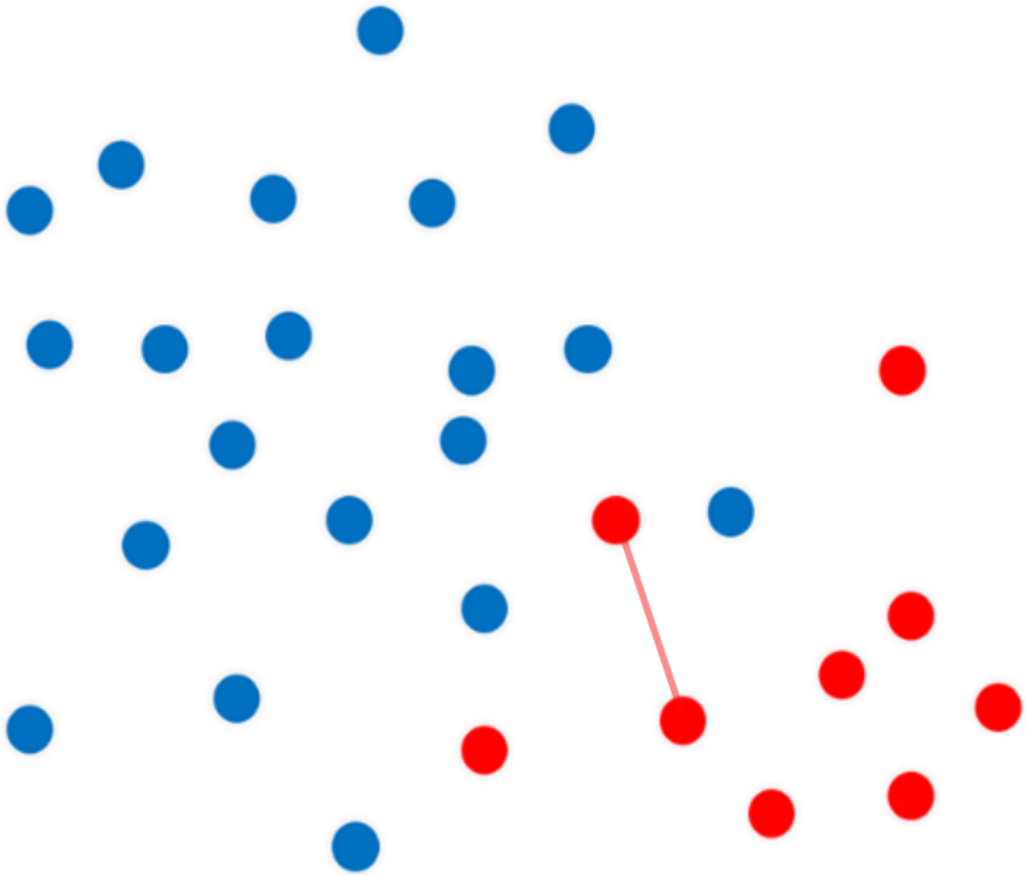


Figure – Schéma de SMOTE

SMOTE : Synthetic Minority Over-sampling Technique

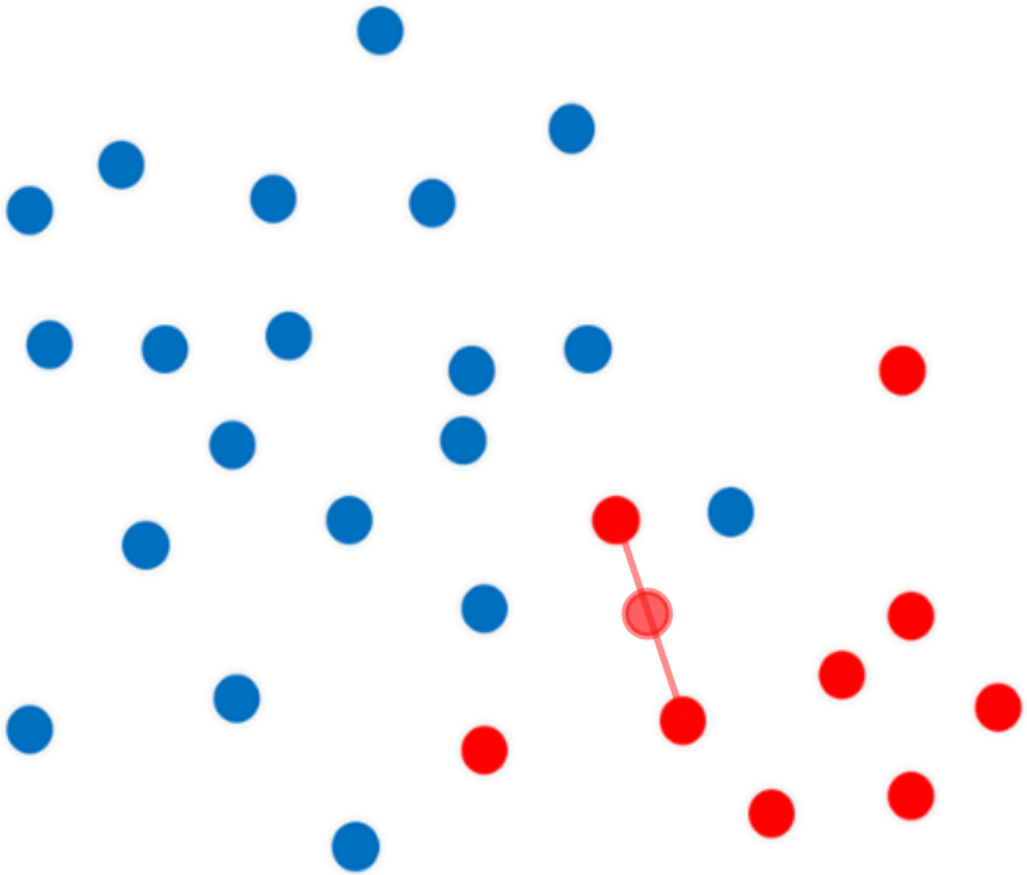


Figure – Schéma de SMOTE

SMOTE : Synthetic Minority Over-sampling Technique

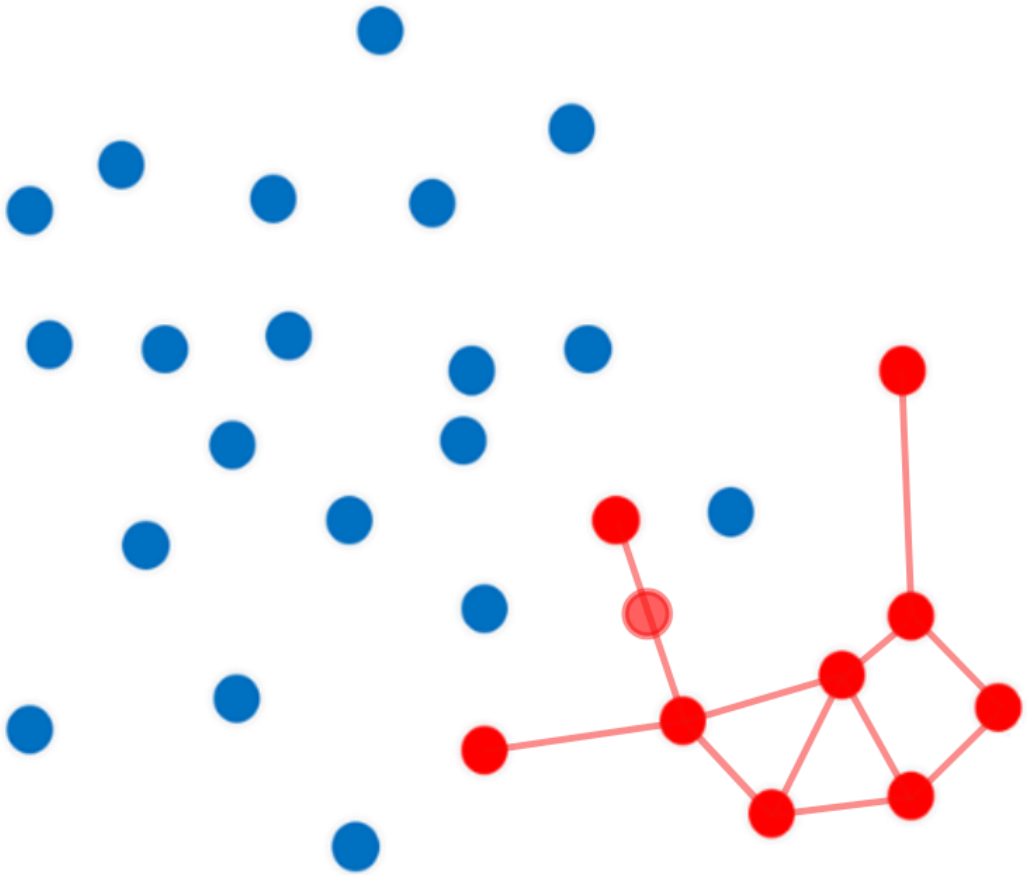


Figure – Schéma de SMOTE

SMOTE : Synthetic Minority Over-sampling Technique

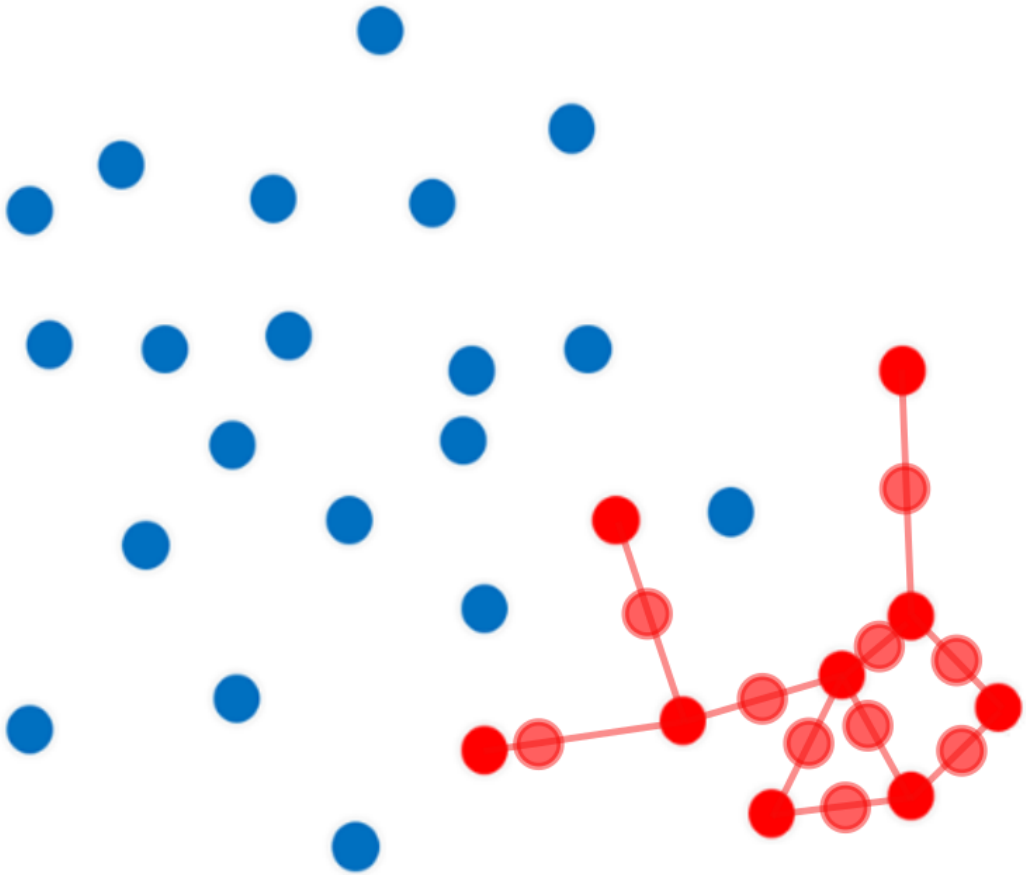


Figure – Schéma de SMOTE

SMOTE : Synthetic Minority Over-sampling Technique

Notations

- n : nb **total** d'observations
- n_{min} : nb d'observations **minoritaires**
- d : dimension (nb de variables)
- k : nb de plus proches voisins
- M : nb de points synthétiques générés

Étapes dominantes & complexité (naïf)

1. **Recherche des k-PPV (vers tous les points)** : coût d'une distance $\mathcal{O}(d)$ $\Rightarrow$ comparaison à n points $\mathcal{O}(n\,d)$ $\Rightarrow$ pour n_{min} points minoritaires $\boxed{\mathcal{O}(n_{\text{min}}\,n\,d)}$.
2. **Génération** : $x_{\text{new}} = x_i + \lambda(x_j - x_i)$, $\lambda \sim \mathcal{U}(0,1)$ $\boxed{\mathcal{O}(M\,d)}$

Synthèse

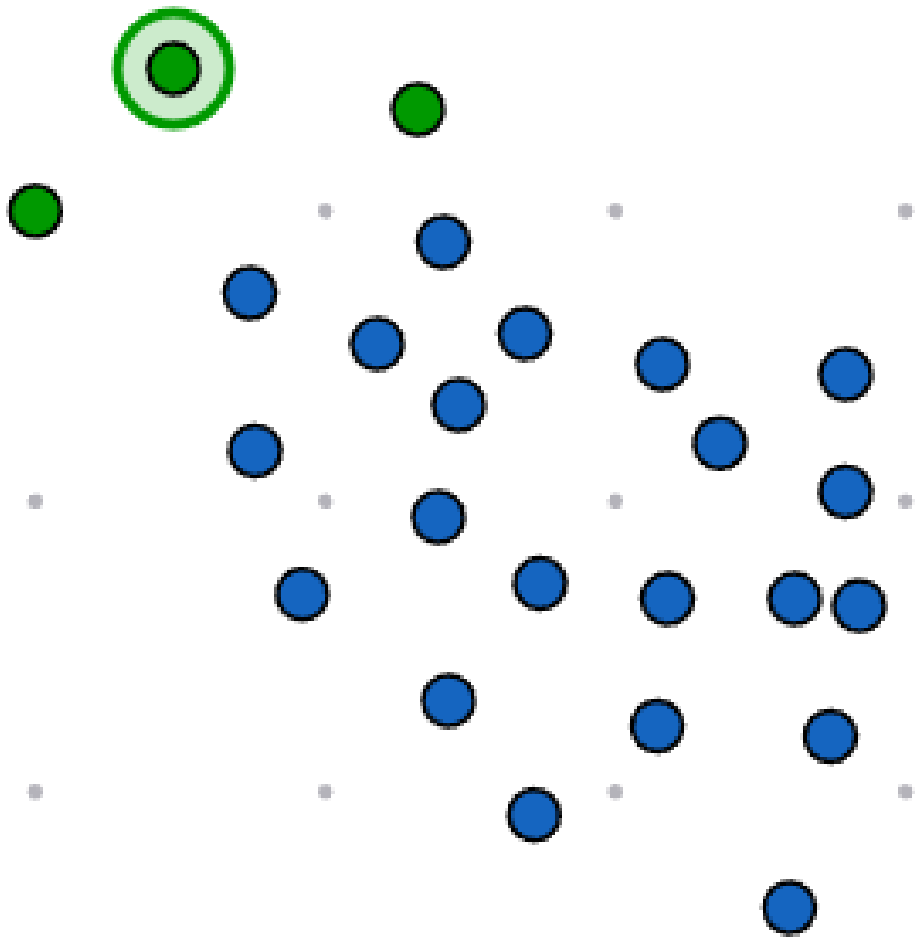
$\boxed{T_{\text{SMOTE}} = \mathcal{O}(n_{\text{min}}\,n\,d) + \mathcal{O}(M\,d)}$ (recherche kNN dominante).

Sommaire

Limites des méthodes

ROS

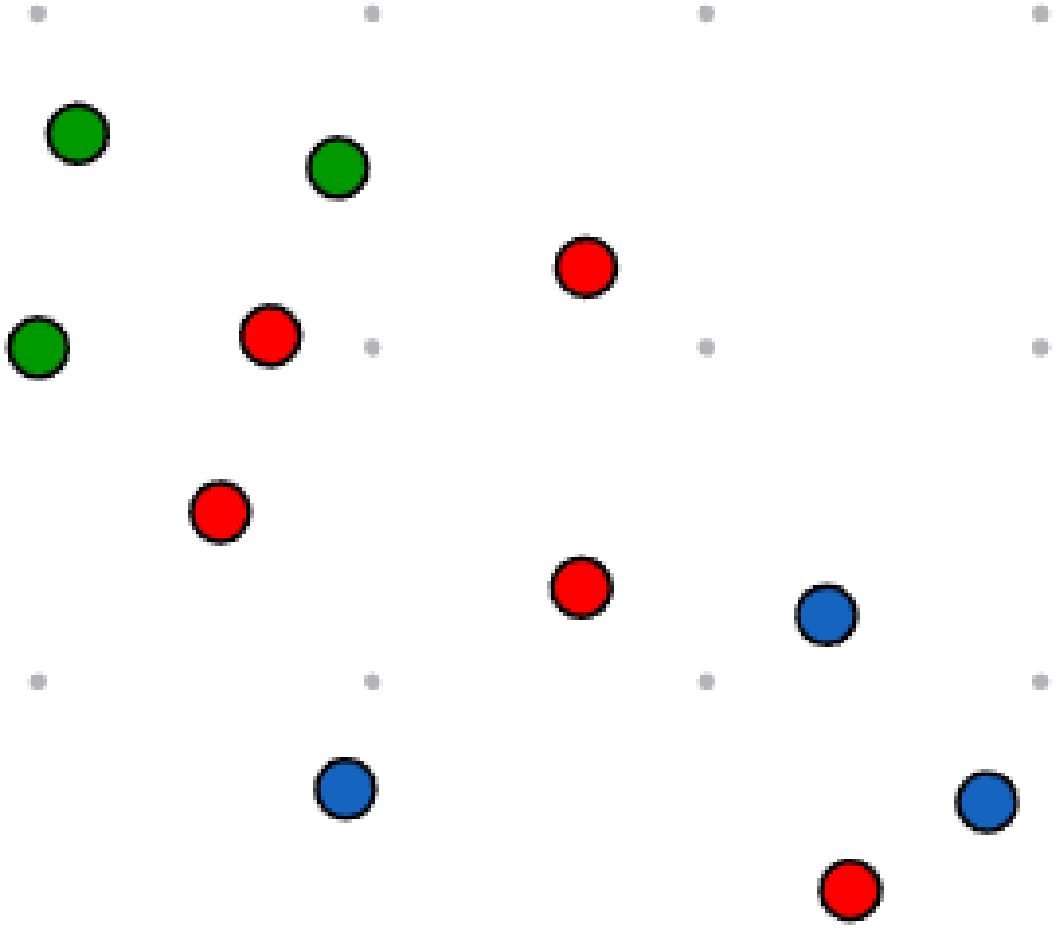
overfiting



Limites des méthodes

RUS

Overfiting



Limites des méthodes

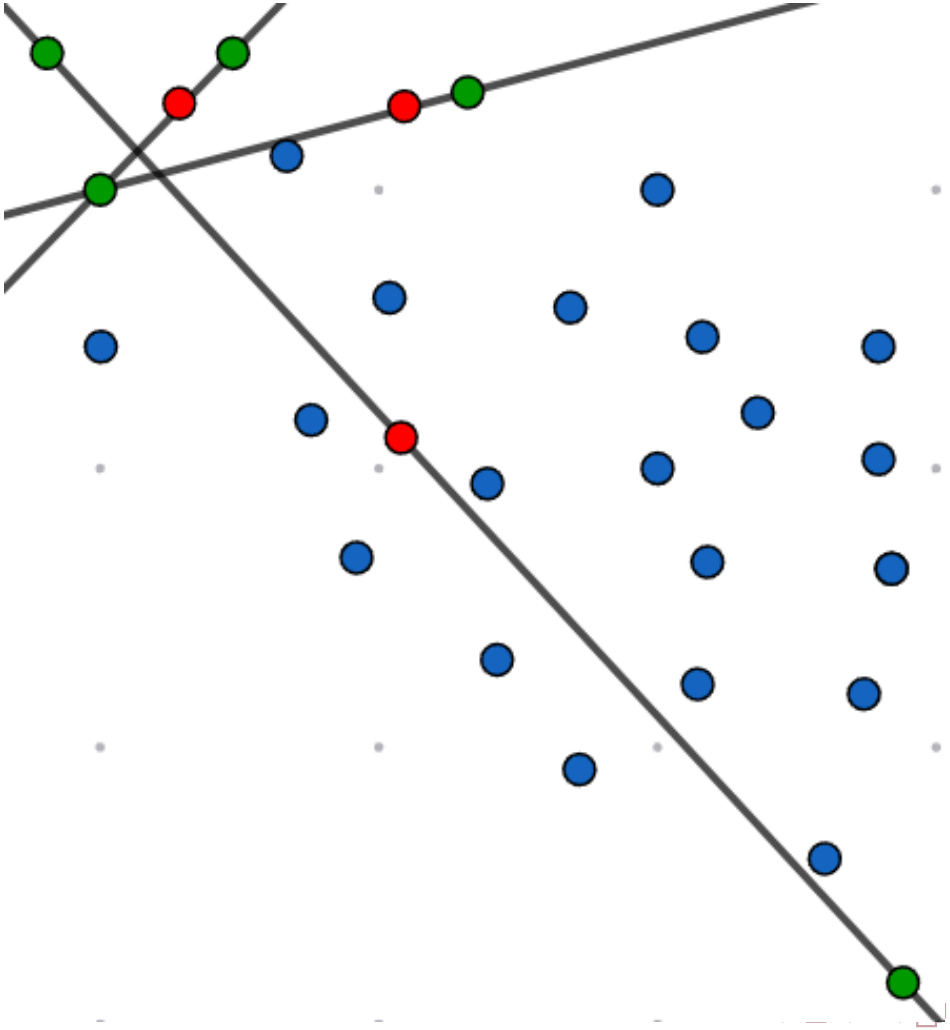
SMOTE

En grande dimension le temps de calculs peut vite augumenter. De plus plus la création de points abérant

Limites des méthodes

SMOTE

Points aberrant



Sommaire

Sommaire

Bilan général des méthodes

Méthode	**Points forts**	**Limites**
ROS	Simplicité, conserve toutes les données	Overfitting, grand volume de données
RUS	Rapide et réduit le biais	perte d'information et représentativité
SMOTE	Données synthétiques variées	Coût élevé, sensible aux outliers

Aucune méthode n'est universelle :
le choix dépend du jeu de données et du modèle.

Conclusion et perspectives

- Pour notre jeu de données, la méthode la plus efficace est SMOTE.
- Pour aller plus loin : il serait pertinent de combiner des méthodes existantes ou de pondérer les modèles.

Merci pour votre attention !