Apprentissage de classes déséquilibrées

HAX907X - Apprentissage statistique

SAWADOGO Kader
GERMAIN Marine
LABOURAIL Célia
MARIAC Damien

Université Montpellier
Département de Mathématique

19 octobre 2025

Sommaire

Problème du déséquilibre et motivation du projet

Un jeu de données très déséquilibré

- CréditCard : **284 807 transactions**, dont seulement **492 fraudes** (**0,17 %**).
- $\Rightarrow$ Les modèles ont tendance à ignorer la classe minoritaire.

Illustration du problème

Modèle	**Recall**	**Precision**	**F1-score**
Régression Logistique	0.59	0.89	0.70
Random Forest (200 arbres)	0.60	0.95	0.85

Sommaire

Problématique scientifique

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

Sommaire

Random Over-Sampling

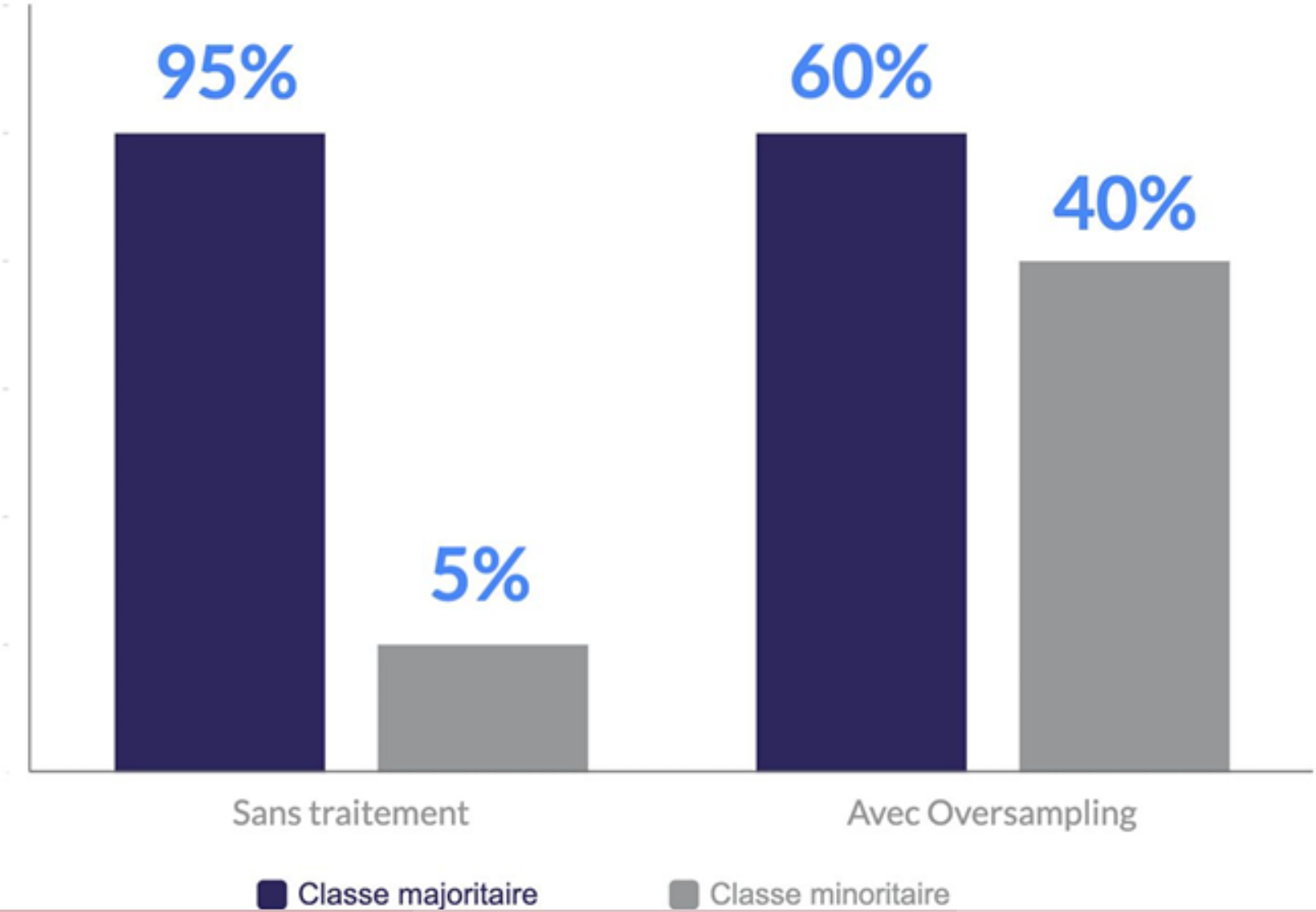


Table – Jeu de données après sur-échantillonnage (ROS)

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	dupliqué
9	1	dupliqué
10	1	dupliqué
8	1	dupliqué

Random Under-Sampling

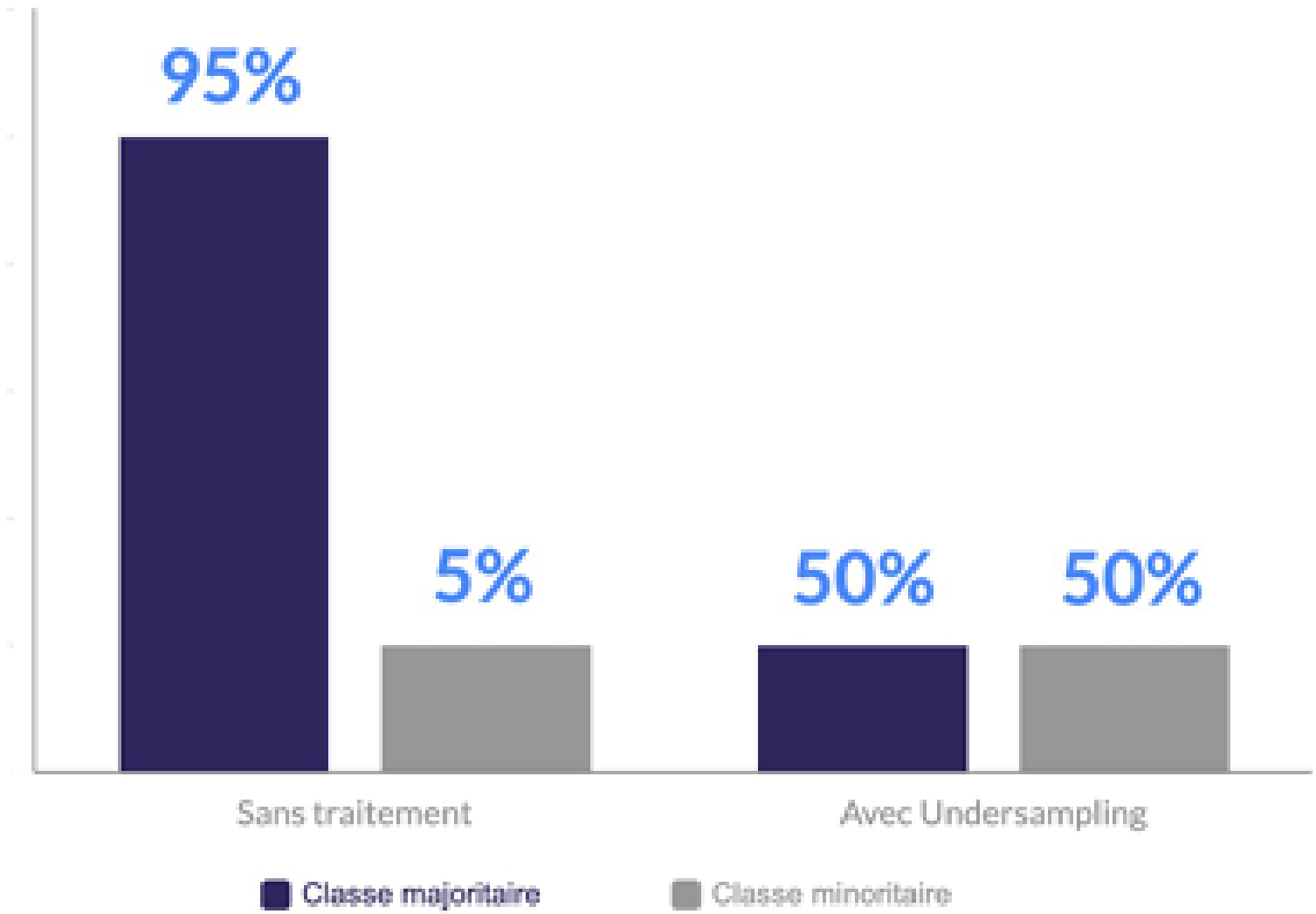


Table – Jeu de données après sous-échantillonnage (RUS)

x	label	source
1	0	supprimé
2	0	supprimé
3	0	supprimé
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

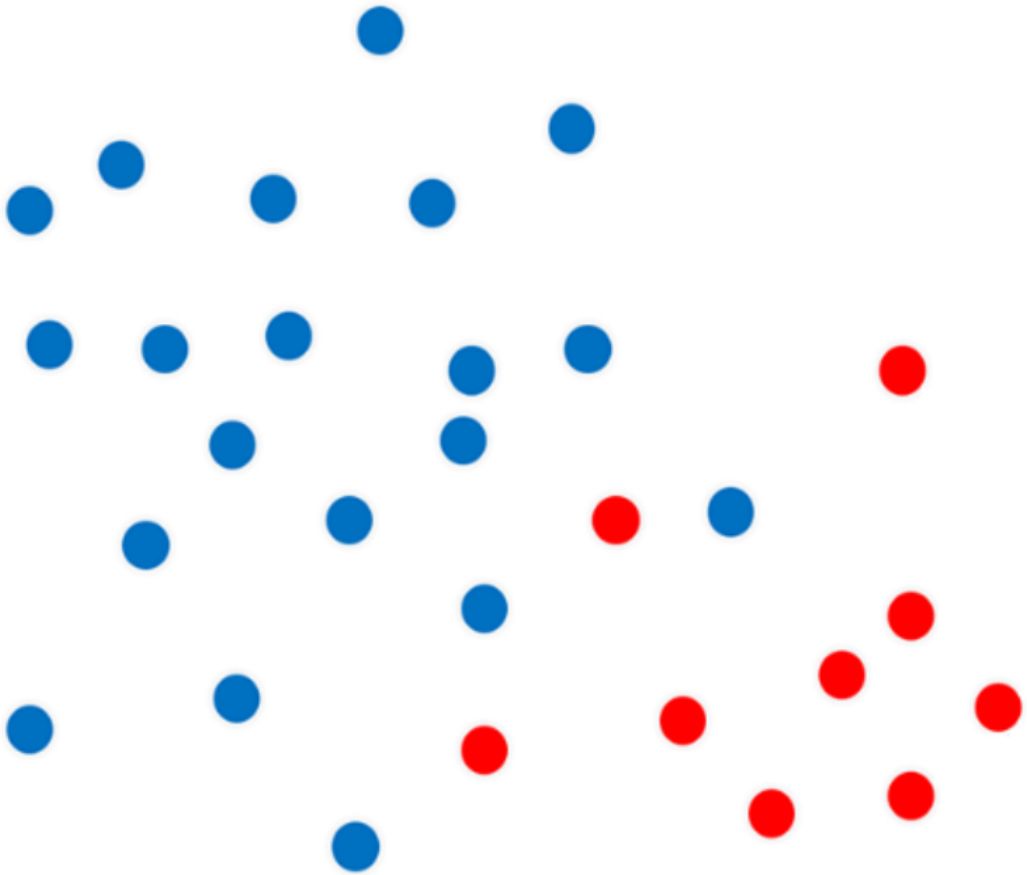


Figure – Schéma de SMOTE

SMOTE : Synthetic Minority Over-sampling Technique

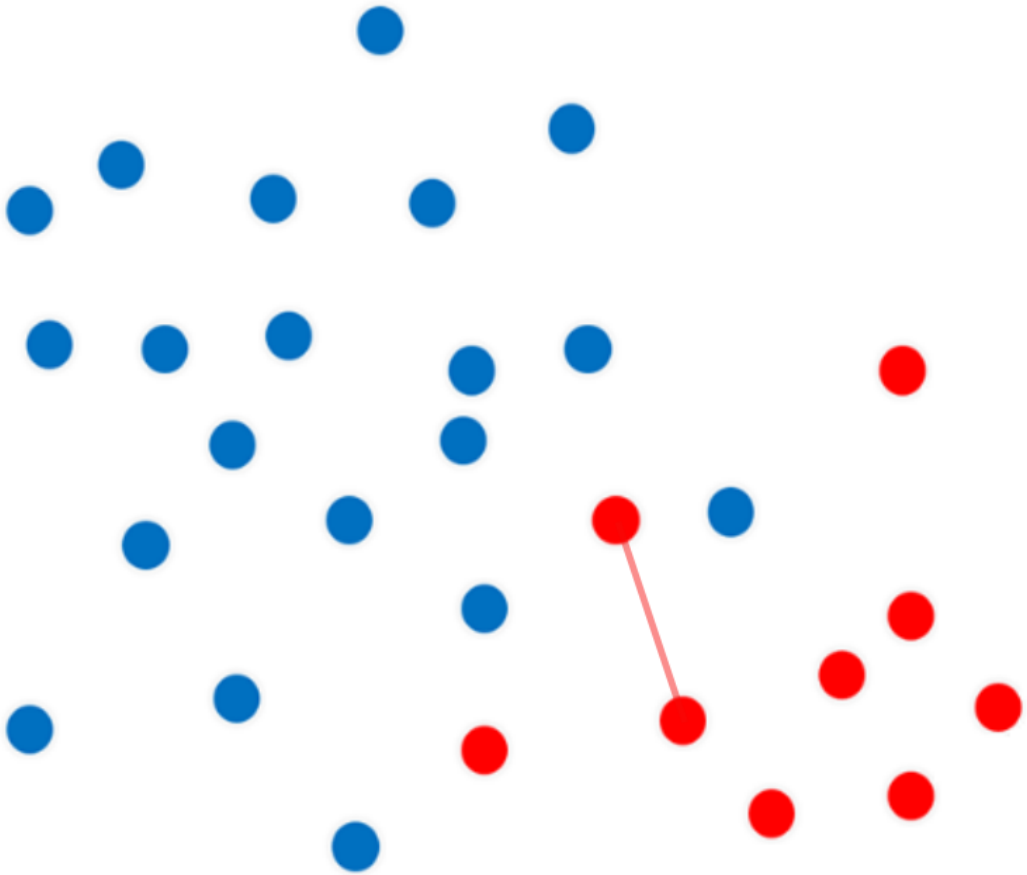


Figure – Schéma de SMOTE

SMOTE : Synthetic Minority Over-sampling Technique

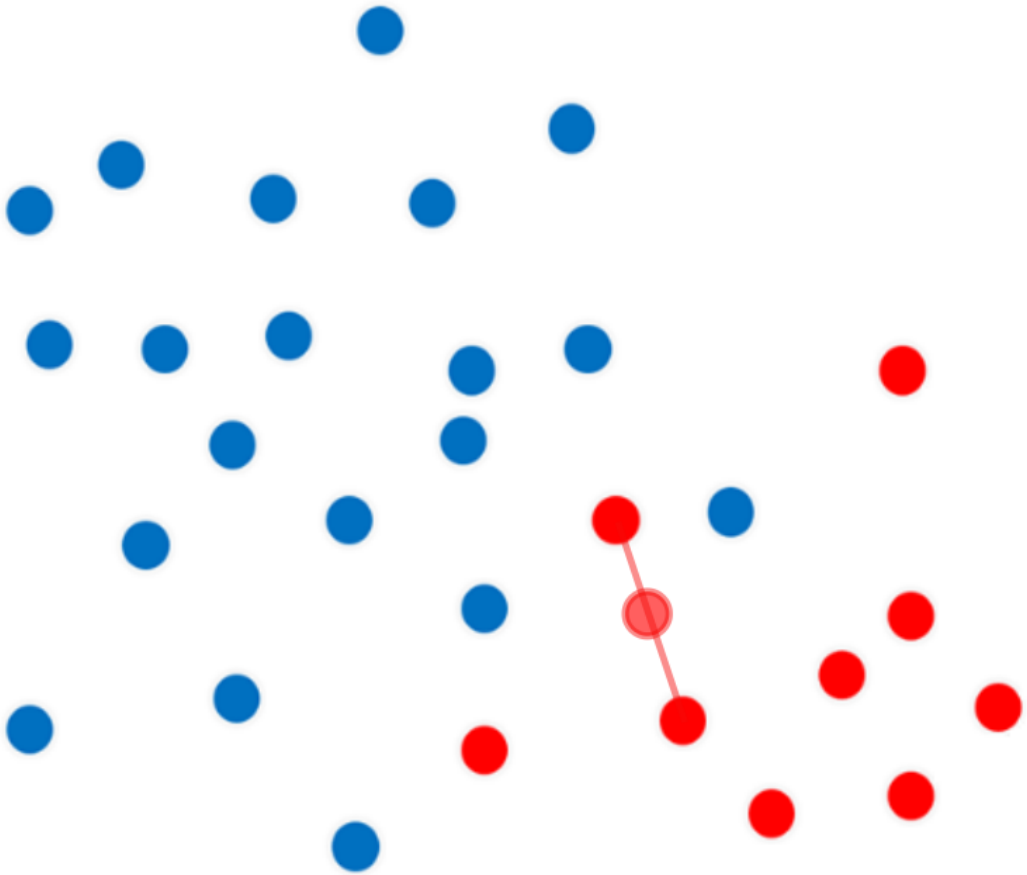


Figure – Schéma de SMOTE

SMOTE : Synthetic Minority Over-sampling Technique

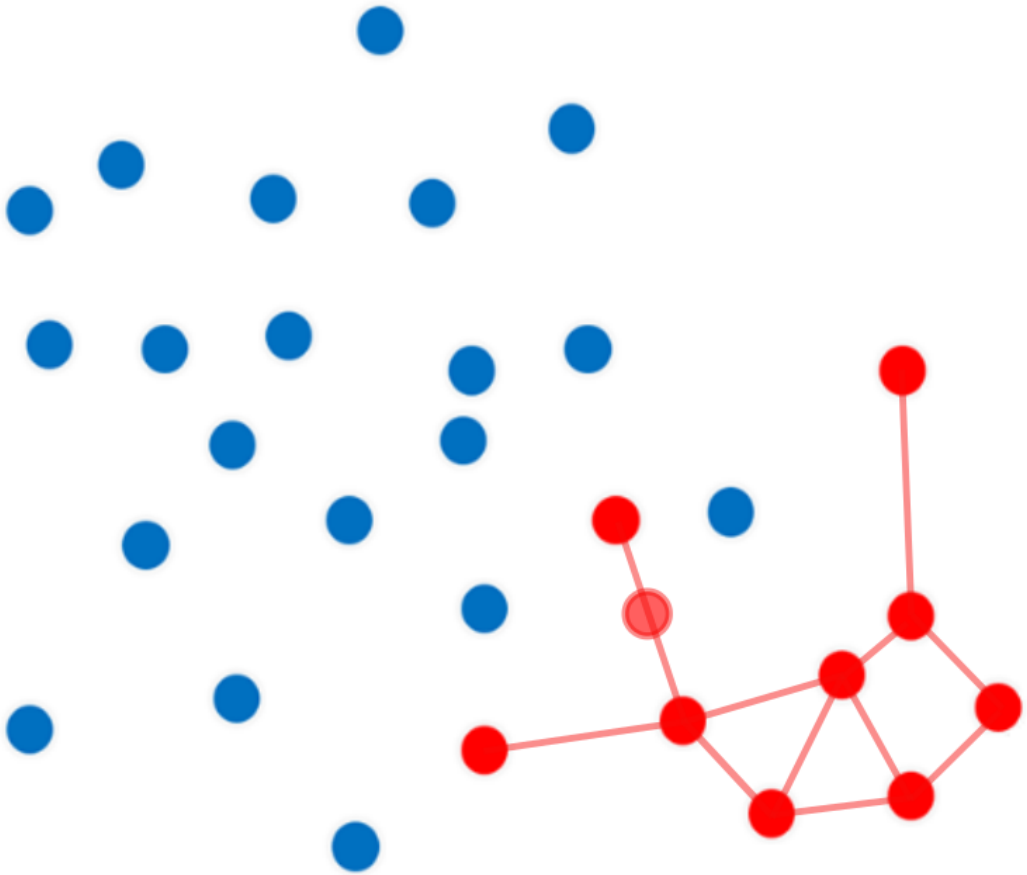


Figure – Schéma de SMOTE

SMOTE : Synthetic Minority Over-sampling Technique

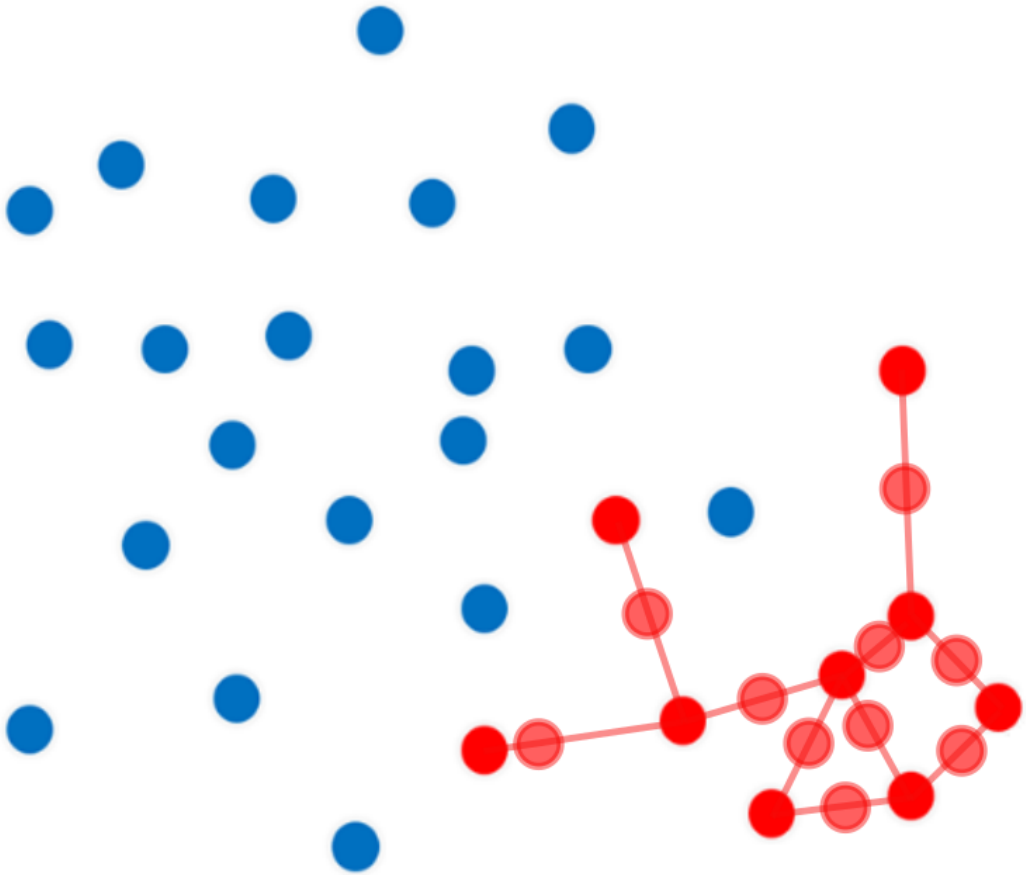


Figure – Schéma de SMOTE

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

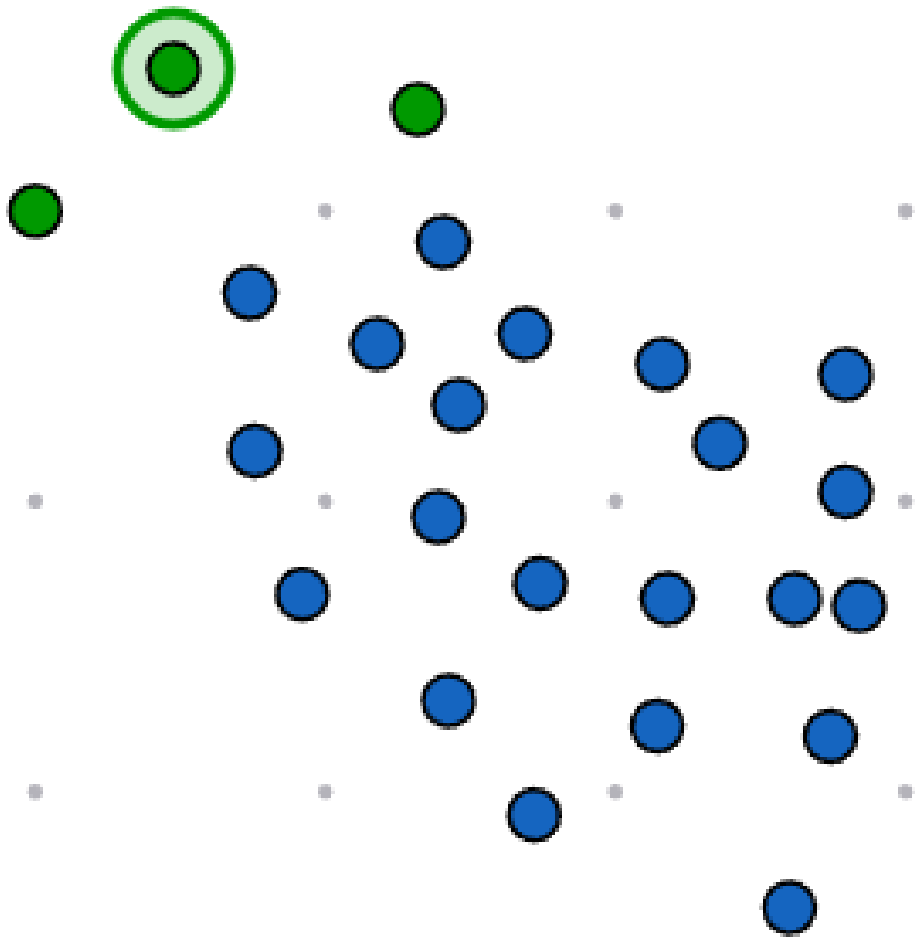
$\boxed{T_{\text{SMOTE}} = \mathcal{O}(n_{\text{min}}\,n\,d) + \mathcal{O}(M\,d)}$ (recherche kNN dominante).

Sommaire

Limites des méthodes

ROS

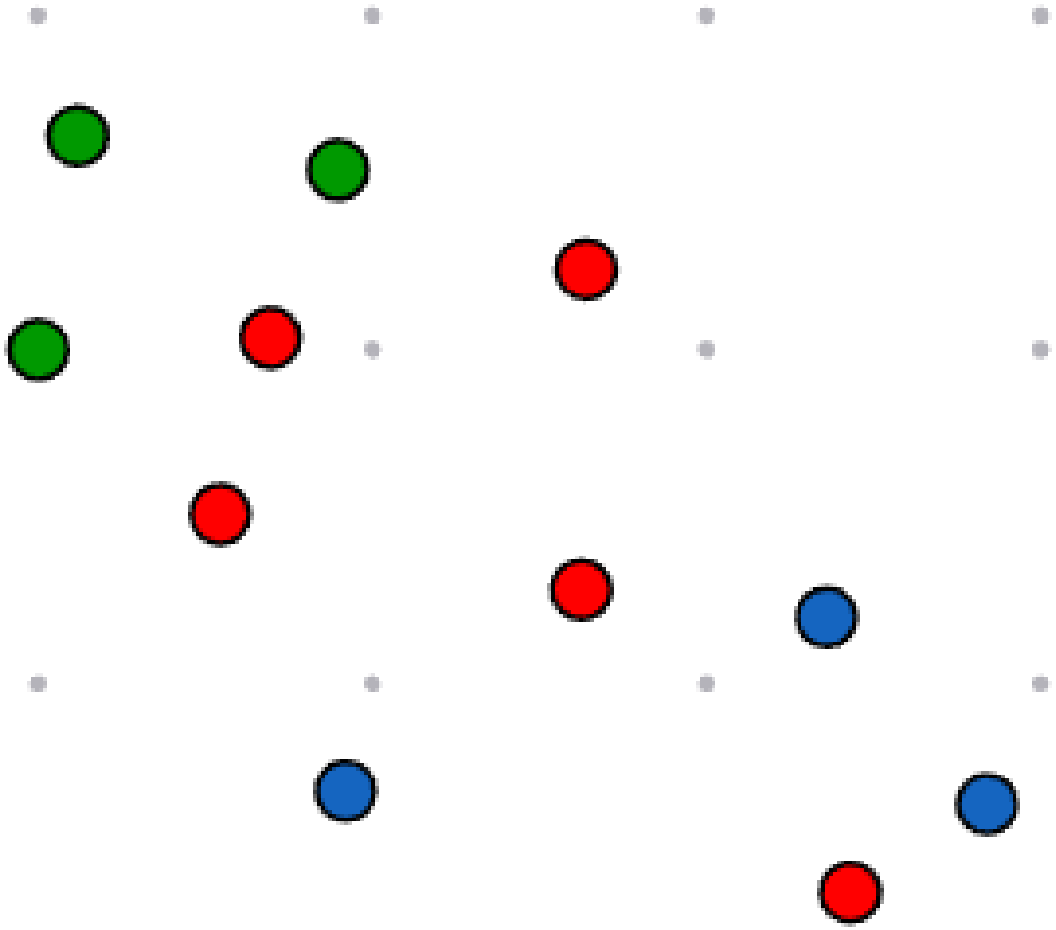
- Overfitting



Limites des méthodes

RUS

- Perte d'information



Limites des méthodes

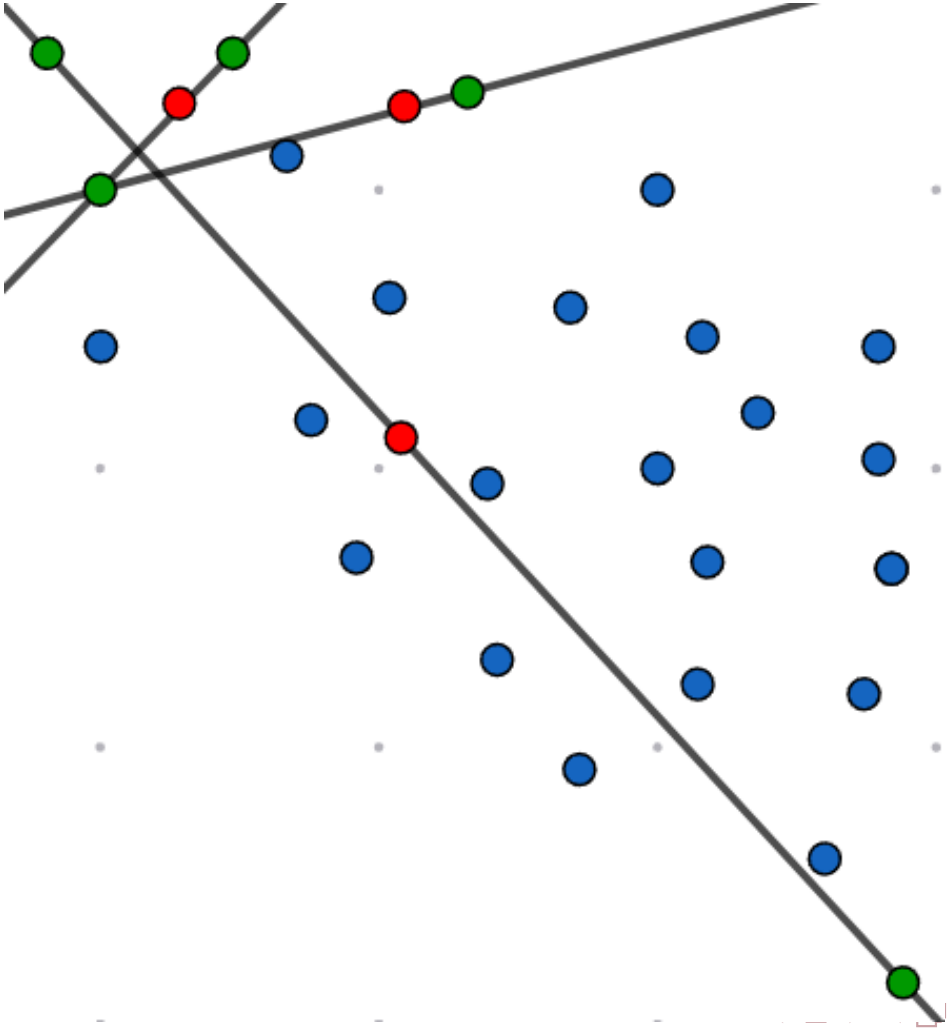
SMOTE

- Temps de calculs
- Création de points aberrants
- Hyperparamètre k
- Variables qualitatives

Limites des méthodes

SMOTE

- Points aberrants



Sommaire

Comparaison des matrices de confusion

	RUS		ROS		SMOTE	
	P-F	F	P-F	F	P-F	F
Pas-Fraude	231	24	233	13	261	12
Fraude	13	232	7	247	0	227

Comparaison des méthodes de rééchantillonnage (Régression Logistique)

Méthode	Rappel (Fraude)	Spécificité	Exactitude équilibrée
RUS	0.906	0.947	0.926
ROS	0.950	0.971	0.960
SMOTE	0.950	1.000	0.975

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 !