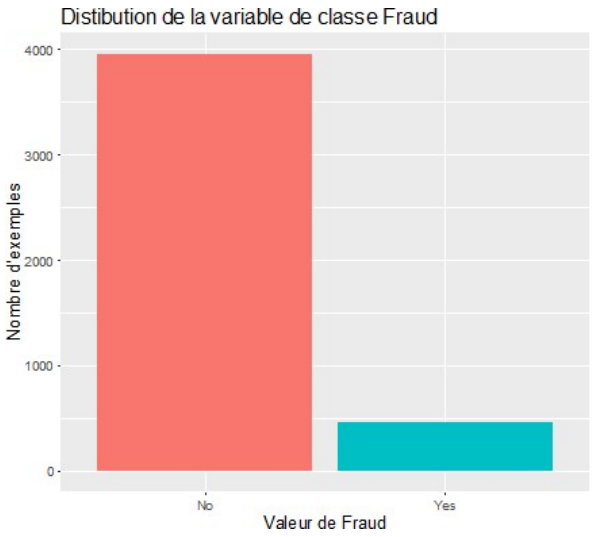
Rééquilibrage des Classes

Cette notice concerne le problème du déséquilibre des classes lors d'une application de classification supervisée. La première section fournit une description de ce problème et la seconde section concerne les solutions possibles par rééquilibrage des classes.

1. Problème du Déséquilibre des Classes

Certaines applications de classification supervisées (e.g. détection de fraudes, prédiction d'incidents, gestion de l'attrition) génèrent par leur nature des données dans lesquelles les classes, c-à-d les valeurs de la variable cible à prédire, sont très déséquilibrées.

Cela signifie que la proportions d'exemples de l'ensemble de données correspondant à chacune des classes sont très différents : l'une des classes est très minoritaire (cas rares) et l'autre classe très majoritaire (cas très fréquents).



L'affichage d'un histogramme des valeurs de la variable de classe permet d'observer simplement ce phénomène. Dans le graphique à droite par exemple, la classe *Fraud=No* (barre rouge) est très majoritaire et la classe *Fraud=Yes* (barre turquoise) est très minoritaire dans l'ensemble de données.

Un effet fréquent de ce déséquilibre des classes est que l'apprentissage d'un arbre de décision ne va générer qu'un nœud racine prédisant systématiquement la classe la plus fréquente (i.e. très majoritaire) dans l'ensemble de données.

2. Solution par Rééquilibrage des Classes

La solution consiste à rééquilibrer les classes de manière que la classe minoritaire soit suffisamment représentée dans l'ensemble d'apprentissage pour que les algorithmes d'apprentissage de classifieurs puissent la reconnaître et la prédire en la distinguant de la classe majoritaire.

Les deux méthodes de rééquilibrage des classes possibles sont :

- Sur-échantillonnage : les exemples de la classe minoritaire dans l'ensemble de données sont dupliqués, éventuellement plusieurs fois en fonction de l'importance du déséquilibre des classes. Cette solution est privilégiée si on dispose de peu d'exemples de la classe minoritaire, i.e. quelques dizaines ou centaines selon la taille totale de l'ensemble de données.
- Sous-échantillonnage : des exemples de la classe majoritaire dans l'ensemble de données sont supprimés, le nombre d'exemples supprimés étant proportionnel à l'importance du déséquilibre des classes. Cette solution est privilégiée si on dispose de très nombreux exemples, i.e. un ensemble de données de plusieurs milliers ou dizaines de milliers d'exemples.

Plusieurs librairies R fournissent différentes fonctions de rééquilibrage des classes :

- Librairie **smotefamily** (https://www.rdocumentation.org/packages/smotefamily) : plusieurs variantes de la méthode SMOTE (Synthetic Minority Oversampling Technique) de sur-échantillonnage pour les données numériques.
- Librairie **DMwR** (https://www.rdocumentation.org/packages/DMwR/) : fonction `SMOTE()` de sur-échantillonnage pour les données numériques et catégorielles.
- Librairie **imbalance** (https://www.rdocumentation.org/packages/imbalance/) : fonctions `racog()` et `rwo()` de sur-échantillonnage et fonction `neater()` de sous-échantillonnage.
- Librairie **UBL** (https://www.rdocumentation.org/packages/UBL/) : fonctions `RandOverClassif()` de sur-échantillonnage aléatoire et `RandUnderClassif()` de sous-échantillonnage aléatoire.
- Librairie **unbalanced** (https://cran.r-project.org/web/packages/unbalanced/) : fonctions `ubOver()` de sur-échantillonnage aléatoire et `ubUnder()` de sous-échantillonnage aléatoire.