Comparaison de méthodes de réduction de dimension pour des analyses de données biologiques

Reunan Bellec & Malo Gillard
Enseignant référent : David Rousseau

17 mars 2019

Table des matières

Introduction

Le dérèglement climatique, le développement de nouvelles maladies des plantes, la maîtrise des rendements, amènent les états et l'ensemble des acteurs de l'agriculture en charge de la sélection variétale, à identifier des semences performantes, résistantes aux maladies, à des périodes de sécheresse ou de brusques variations environnementales durant leur développement. Ces travaux peuvent bénéficier d'avancées technologiques récentes en matière de traitements de l'information, applicables sur de larges populations de plantes. Une échelle particulièrement importante est celle de la graine, dont la qualité germinative conditionne la suite du développement de la plante. Dans ce projet annuel, nous nous intéressons à des graines de betterave sucrière pour laquelle la France est l'un des plus gros producteurs au monde.

L'objectif de l'expérience est d'élargir la variabilité génétique de la betterave, dans le but de la rendre plus compétitive, en doublant le rythme de croissance annuelle de son rendement en sucre. Nous allons donc tester différents génotypes de betteraves (200 individus) et étudier leur germination. À partir des semences des populations de betteraves sélectionnées et des résultats de leur germination, nous observons plusieurs différences, que l'on va chercher à caractériser. Parmi les 200 génotypes, les variables à expliquer sont la surface, la longueur, la largeur, l'imbibition, la vitesse de germination, etc.

Notre rôle ici sera de comparer plusieurs méthodes de réduction de dimension de l'espace de toutes ces variables, de manière à classifier les différents génotypes par la suite. Nous allons donc considérer plusieurs méthodes afin de répondre à cette problématique : PCA, t-SNE, Random Forest, LDA, etc.

Dans un premier temps, nous étudierons le principe de chaque méthode, ce qui nous permettra d'aborder les notions d'apprentissages supervisé et non supervisé, mais également de classer toutes ces méthodes selon leur fonctionnement. Dans un second temps, après avoir compris chacun des précédés, nous les utiliserons sur les données issues de l'expérience de germination et nous comparerons les résultats.

Première partie

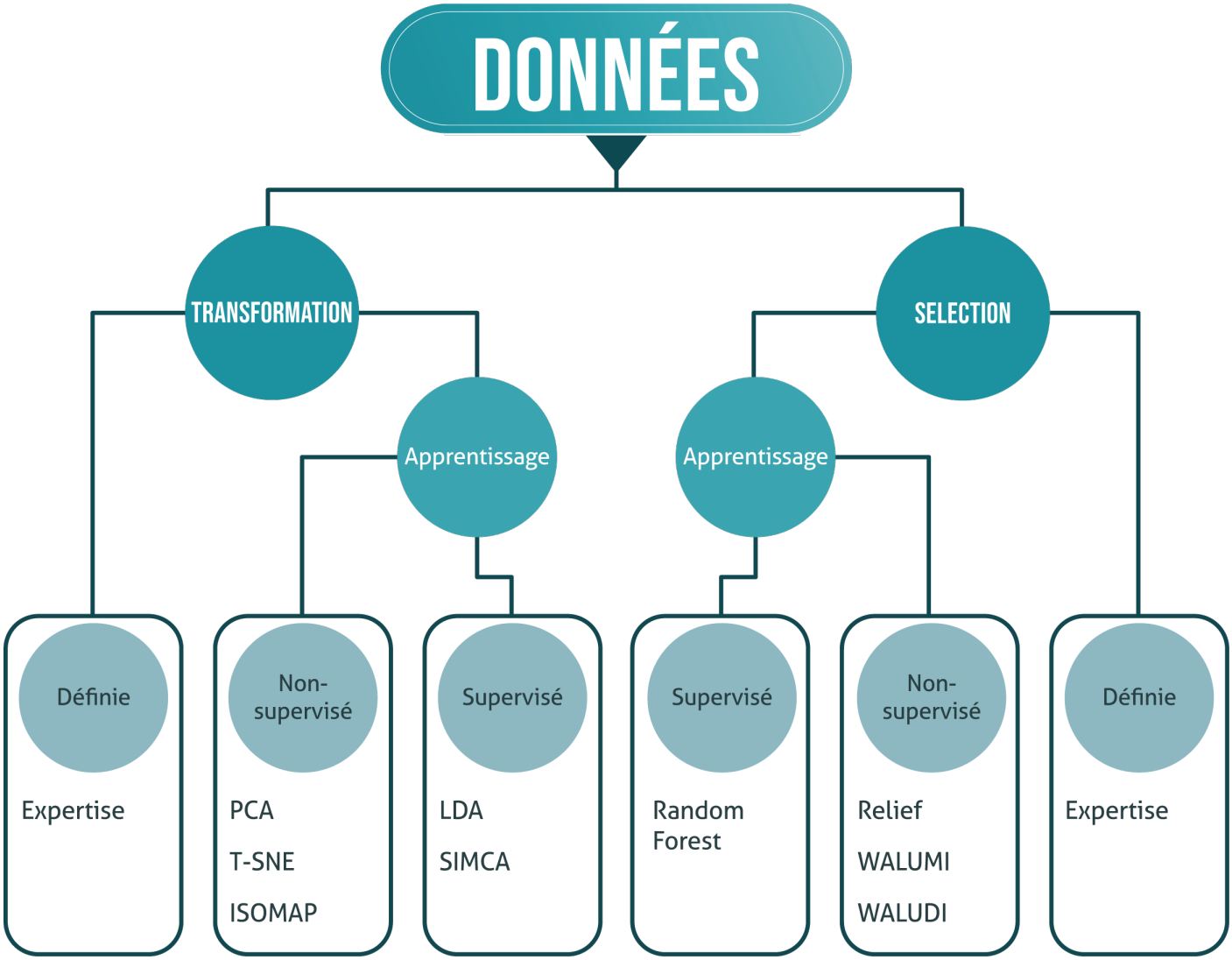
Principes des méthodes

0.1 Groupe 1

0.2 Groupe 2

0.3 Groupe 3

0.4 Groupe 4



Deuxième partie

Réduction de dimension sur des données biologiques

0.5 Expertise

La première tâche à réaliser avant de travailler sur les données au travers des différentes méthodes de réduction de dimension est l'expertise, c'est à dire la préparation, le nettoyage des données. Nous avons réfléchis avec un point de vue autre que celui de data scientist : qu'est-il bon de garder ou d'écarter dans le jeu de données, faut-il rajouter des variables explicatives, et pourquoi.

Comme nous nous intéressons à l'évolution des germinations jour par jour, nous avons écarté les variables qui s'exprimaient en heures, c'est à dire les variables **5°C TMG (h)** et **5°C T50 (h)**, et gardé celles qui s'exprimaient en jours, **5°C TMG (j)** et **5°C T50 (j)**.
Cependant, cette dernière comprenait un nombre important d'entrées "Nan" (plus de la moitié), ce qui signifie qu'il n'a pas été possible de calculer le délai nécessaire pour obtenir 50% de germination pour plus de la moitié des individus. Nous l'avons donc également écartée car elle présentait peu d'intérêt dans notre étude.

Nous avons ensuite rajouté 6 variables **'v15-16j'**, **'v16-17j'**, **'v17-18j'**, **'v18-19j'**, **'v19-20j'**, et **'v20-21j'**, pour exprimer la vitesse de germination de chaque individu jour après jour à partir du jour 15.