CHAPITRE 1

Machine Learning

Objectifs : Ce chapitre est une introduction des éléments de contexte du Machine Learning en parcourant son histoire, de ses prémices jusqu'aux évolutions actuelles. Il dresse un panorama des méthodes et se termine en montrant comment ces méthodes, dans notre pratique quotidienne de la recherche, ont permis de réaliser des avancées significatives.

Outils utilisés : ☐ C ☐ Python ☐ Java ☐ Librairies tierces

1.1 Un peu d'histoire

Les statistiques constituent un corpus très ancien. En effet, dès que la civilisation a été suffisamment développée pour que le prélèvement des impôts soit mis en place, il est devenu nécessaire pour les institutions de connaître précisément la richesse de leur communauté. Les moyens de mesurer la richesse disponible, et donc le montant de l'impôt, sont donnés par les statistiques. En Mésopotamie, les archéologues ont mis au jour des quantités de tablettes cunéiformes relatives à des informations qui peuvent être qualifiées de statistiques puisqu'il s'agissait de recueillir des données à propos de la situation des états ou de la société. Plus tard, selon Tacite, l'empereur Auguste aurait demandé un bilan des richesses de l'empire romain (soldats, navires, ressources privées et publiques). Mais des traces beaucoup plus anciennes sont présentes en Égypte, en Mésopotamie et en Chine.

Le mathématicien arabe Al-Kindi (801-873) détaille l'utilisation des statistiques et l'analyse de la fréquence des caractères pour déchiffrer des messages chiffrés en substituant les caractères d'un alphabet par d'autres (généralement du même alphabet).

Les statistiques s'appuient sur les débuts des probabilités développées par Fermat et Pascal, mais il faut attendre Adolphe Quételet au début du XIX^{ème} siècle pour que les statistiques deviennent une science à part entière et indépendante des probabilités.

C'est Pierre-Simon de Laplace qui fait entrer l'analyse dans la théorie des probabilités vers 1812, en donnant une première version d'un théorème devenu le théorème central limite. Celui-ci démontre que, sous de vastes hypothèses, les principales propriétés d'une collection assez grande de données se comportent selon une loi de probabilité d'un type unique : la loi de Laplace-Gauss. La théorie moderne des probabilités est, au contraire,

fort récente puisqu'elle ne prend réellement son essor qu'avec Émile Borel (1871-1956) mais surtout Andreï Kolmogorov (1903-1987).

Les statistiques sont restées longtemps "descriptives", c'est-à-dire se bornant à décrire une situation par quelques caractéristiques (moyenne, médiane, écart-type…) ou à la représenter par un (ou des) graphique(s). L'analyse de données (analyse en composantes principales, analyse hiérarchique…) est la forme moderne la plus aboutie des statistiques descriptives. Cette dernière partie des statistiques est fortement liée à la géométrie puisqu'elle permet d'identifier les caractéristiques principales des données statistiques ou de les regrouper par affinité. Elle s'appuie donc fortement sur le calcul matriciel.

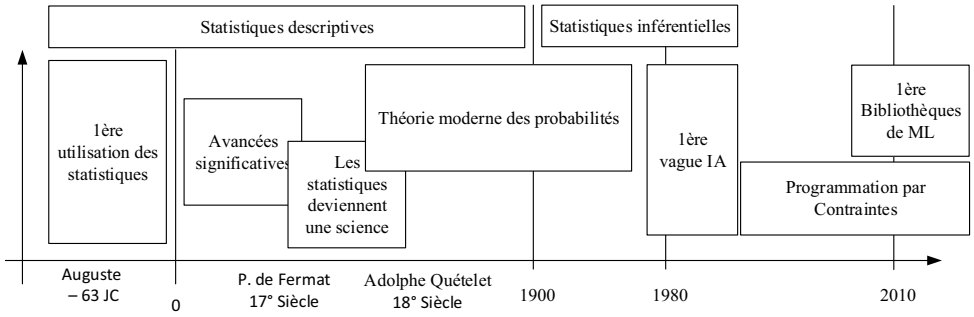

Figure 1-1. Chronologie des événements autour du Machine Learning (ML)

Puis, essentiellement depuis les travaux de William Gosset pour la brasserie Guinness (publiés en 1908 sous le pseudonyme de Student pour préserver le secret industriel !), la statistique est devenue inférentielle (Figure 1-1) : les données disponibles sont considérées comme extraites d'une population et il s'agit, à partir des observations faites sur ces données, d'en déduire des informations sur la population tout entière.

Cette partie des statistiques est intimement liée à la théorie des probabilités et repose principalement sur le théorème de la limite centrale (ou théorème central limit) qui établit la convergence en loi de la somme d'une suite de variables aléatoires (indépendantes, identiquement distribuées) vers la loi normale (ou loi de Laplace-Gauss).

Ainsi, il est possible de connaître (approximativement) la loi d'une somme d'au moins 30 observations quantitatives indépendantes, sans connaître la loi de celles-ci. À titre d'exemple, à partir d'une série statistique de n individus (avec $n \geq 30$) ayant un caractère numérique de moyenne m et d'écart-type σ, la **somme des valeurs** du caractère peut être considérée comme la réalisation d'une variable aléatoire de moyenne $n.m$, d'écart-type $\sigma\sqrt{n}$ et de densité :

$$\frac{1}{\sigma\sqrt{n}.\sqrt{2\pi}} . e^{-\frac{1}{2}\times\left(\frac{x-n.m}{\sigma\sqrt{n}}\right)^2}$$

Par exemple**, la moyenne** des n valeurs de ce caractère est de loi, sur $\mathbb{R}$, de densité :

$$\frac{\sqrt{n}}{\sigma.\sqrt{2\pi}}.e^{-\frac{1}{2}\times\left(\frac{x-m}{\frac{\sigma}{\sqrt{n}}}\right)^2}$$

Ce qui signifie que, si $\bar{X}$ est la moyenne mesurée sur les données, il est possible d'affirmer avec une probabilité de 95% que la moyenne m correspondante (et inconnue) dans la population entière, appartient à l'intervalle : $\left[\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}}; \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right]$.

Avec l'apparition des ordinateurs individuels, les études statistiques se sont rapidement développées. L'agriculture, très en avance sur les statistiques a, très tôt, développé des bibliothèques de programmes statistiques. À la fin du siècle dernier, divers logiciels intégrés de statistiques étaient disponibles, et ont abouti à la création de logiciels libres. Désormais, des bibliothèques faciles d'accès sont disponibles et peuvent être utilisées à partir de langages très répandus tels que Python, Java ou C.

1.2 Les années 90

L'article de Wikipedia est assez explicite sur le sujet et il fixe le début de ce qu'il est possible d'appeler Intelligence Artificielle à l'été 1956 après une conférence tenue sur le campus de Dartmouth :

https://fr.wikipedia.org/wiki/ Histoire_de_l'intelligence_artificielle

Le nombre de scientifiques intéressés par le sujet reste modeste et la croissance des recherches est freinée par les machines de l'époque qui sont à la fois encombrantes et peu puissantes. Quelques avancées significatives ont été obtenues comme, par exemple, la conception d'un jeu d'échecs proposé par Christopher Strachey ou par Dietrich Prinz.

En 1950, Alan Turing imagine son célèbre test (de Turing) : "si une machine peut mener une conversation qu'on ne puisse différencier d'une conversation avec un être humain, alors elle peut être qualifiée d'intelligente". C'est en 1957 qu'est introduite par Frank Rosenblatt, la notion de perceptron (un algorithme d'apprentissage supervisé de classifieurs binaires).

Les subventions ont été stoppées dans les années 1970 et la recherche ne reprit réellement qu'au début des années 1980, essentiellement grâce à l'arrivée de machines de bureau puissantes. Deux grandes tendances se détachent alors :

- L'apparition rapide des systèmes experts qui connaissent un grand succès en particulier pour diagnostiquer des maladies infectieuses du sang.
- Ces années marquent la renaissance du connexionnisme à travers les travaux de David Rumelhart et de John Hopfield. Ce dernier définit et étudie des réseaux neuronaux (désormais appelés réseaux de Hopfield) pour apprendre et traiter de l'information différemment.

Toutefois, les systèmes experts restent difficiles à mettre au point : ils peuvent faire des erreurs importantes quand les paramètres sortent des valeurs habituelles, de sorte que ces systèmes ne sont réellement utiles que dans des contextes très spécifiques.

Quel que soit le type d'algorithme, les machines n'exécutent que des algorithmes et, de ce point de vue l'affirmation de Luc Julia (un des fondateurs de Siri), prend tout son sens : "l'intelligence artificielle (IA) n'existe pas". En résumé, l'IA est l'ensemble des outils permettant à une machine de réaliser des tâches complexes sans avoir été explicitement programmée pour cela.

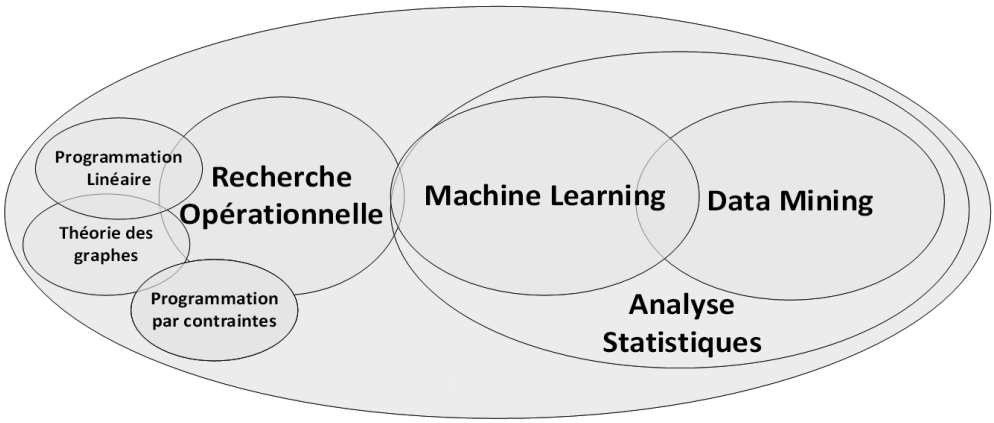

Figure 1-2. L'Intelligence Artificielle

Par abus de langage, le Machine Learning est assimilé à l'Intelligence Artificielle (IA) alors que le Machine Learning fait partie de l'IA au même titre que la Recherche Opérationnelle (Figure 1-2).

1.3 Le Machine Learning pour tous

Le Machine Learning est un sous-ensemble de l'IA qui concerne la création d'algorithmes qui permettent d'apprendre à partir de données précédemment collectées. Le terme Machine Learning a été introduit en 1959 par Arthur Samuel qui en donne la définition suivante "Machine Learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed" ce qui peut se traduire par "Le Machine Learning permet à une machine d'apprendre automatiquement, à partir de données, d'améliorer ses résultats par un processus d'apprentissage et de fournir ensuite des résultats qui n'ont pas été explicitement programmés".

Le Machine Learning regroupe plusieurs thèmes (tout classement est nécessairement discutable) et il est proposé de distinguer ici (Figure 1-3) :

- L'**apprentissage supervisé** qui regroupe sous ce terme toutes les méthodes capables de prendre en compte une base d'apprentissage pour en "retirer" les informations utiles pour, par la suite, manipuler et classer de nouvelles données.
- Les **problèmes de classification**, ce qui englobe une vaste famille de méthodes dont les méthodes de centres mobiles (Kmean), les méthodes SVM (Support Vector Machine), les arbres de décisions et les Réseaux Bayésiens.
- Les **réseaux de neurones**, qui par l'attention qui leur a été portée constituent un domaine à part entière. Ces méthodes peuvent être utilisées, entre autres, pour faire de la classification mais elles possèdent un champ d'application beaucoup plus large.

- Le domaine de la **fouille de données** et tous les éléments qui tournent autour du Pattern Mining au sens large font aussi partie du Machine Learning mais toutefois, pas au même titre que les réseaux de neurones, dans le sens, où les résultats de ces méthodes nécessitent (encore) une part d'interprétation manuelle encore important.

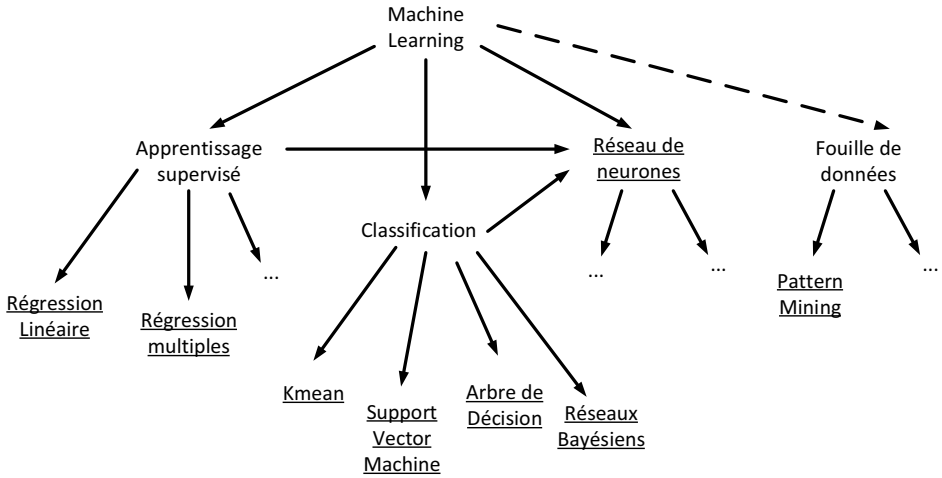

Figure 1-3. Domaines du Machine Learning

L'application de ces méthodes peut se faire dans différents domaines. En particulier, l'étude des **séries chronologiques** fait appel à de nombreuses méthodes qui relèvent en partie du Machine Learning. Les séries chronologiques constituent un domaine à elles seules et c'est la raison qui a conduit à leur consacrer le dernier chapitre de ce livre.

1.4 Les outils présentés dans ce livre

1.4.1 Les réseaux de neurones

Les concepts des réseaux de neurones sont relativement anciens et ils ont été remis sur le devant de la scène avec l'arrivée de nouveaux types de réseaux (qui ont permis de traiter efficacement de nombreux problèmes de reconnaissance d'images) et ensuite avec l'arrivée de la librairie TensorFlow de Google. Les méthodes de classification les plus connues sont des classifieurs linéaires qui découpent linéairement un ensemble de données en définissant un hyperplan (en 2 dimensions il s'agit d'une droite). Les cas où les données ne sont pas linéairement séparables sont très nombreux et dans ce cas, il faut avoir recours à des classifieurs non linéaires. Les réseaux de neurones font partie de la famille des classifieurs non linéaires.

Un réseau de neurones est une version très simplifiée des neurones naturels dans lequel, les entrées sont transformées en une sortie et les interconnexions entre neurones définissent un réseau. Un réseau se caractérise principalement par 3 informations :

- le nombre d'entrées et le nombre de sorties (nombre de neurones sur la couche d'entrée et sur la couche de sortie) ;
- le nombre de neurones par couche et le nombre de couches ;
- les interconnexions entre les neurones.

Visuellement un réseau de neurones se présente donc comme un réseau ordonné par couches, comme le montre la Figure 1-4, qui propose un réseau avec 3 entrées, 3 sorties et une couche cachée, elle-même composée de deux neurones.

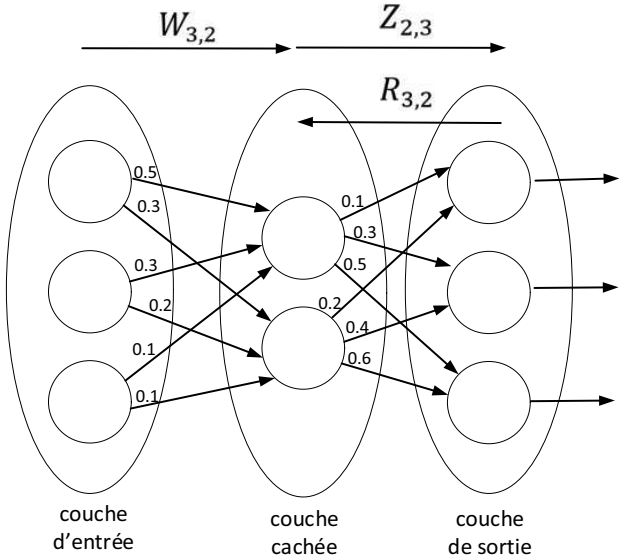

Figure 1-4. Exemple de réseau de neurones

Un réseau de neurones nécessite la manipulation de vecteurs et de matrices pour faire les calculs. Deux types de calculs sont réalisés avec les vecteurs et les matrices : le premier permet de mettre à jour itérativement les matrices, pour rechercher la meilleure matrice possible, c'est-à-dire, celle minimisant l'erreur entre le résultat obtenu et le résultat souhaité ; le deuxième permet de calculer les valeurs de sortie en fonction des valeurs d'entrée (il s'agit d'évaluer les sorties).

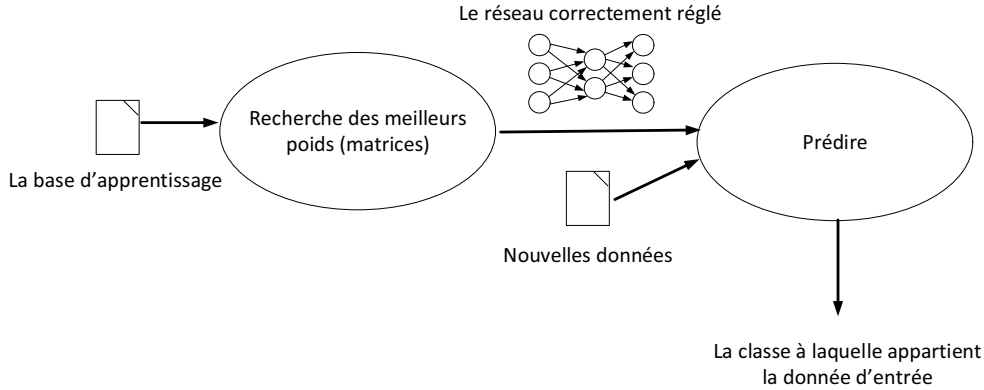

Figure 1-5. Processus d'utilisation d'un réseau de neurones pour un problème de classification

Ainsi pour pouvoir utiliser un réseau de neurones, il faut d'abord, comme le montre la Figure 1-5 rechercher les meilleurs poids possible sur une base d'apprentissage. Ceci

sous-entend qu'il faut disposer de suffisamment de données déjà classées par l'intermédiaire d'un autre processus qui peut être, par exemple un processus manuel : pour chaque entrée, il faut disposer de la classe associée. Le processus d'apprentissage se fait sur ces données d'entrée. Une fois le réseau correctement entraîné à partir des données de la base d'apprentissage, il est alors possible de lui fournir des données nouvelles (elles n'appartiennent pas à la base d'apprentissage) et de lui faire classer ces données : le réseau prédit la classe la plus probable à laquelle la donnée devrait appartenir. Il faut souligner l'importance des données qui définissent la base d'apprentissage du réseau, données qui soulèvent beaucoup de problèmes qui ne sont pas abordés dans ce livre, livre qui se concentre sur "comment" définir la phase d'apprentissage à la fois en C et en Python et qui donne des exemples d'utilisation simples à coder et à appréhender.

1.4.2 Les réseaux Bayésiens

Un réseau Bayésien est un graphe dont chaque sommet porte une variable aléatoire et qui permet d'identifier et de comprendre les dépendances existantes entre plusieurs données d'un problème. Dans la suite du livre, l'exemple proposé est emprunté au milieu médical, et il s'agit d'étudier les liens potentiels entre l'apparition de pathologies, des données comportementales et des données biologiques.

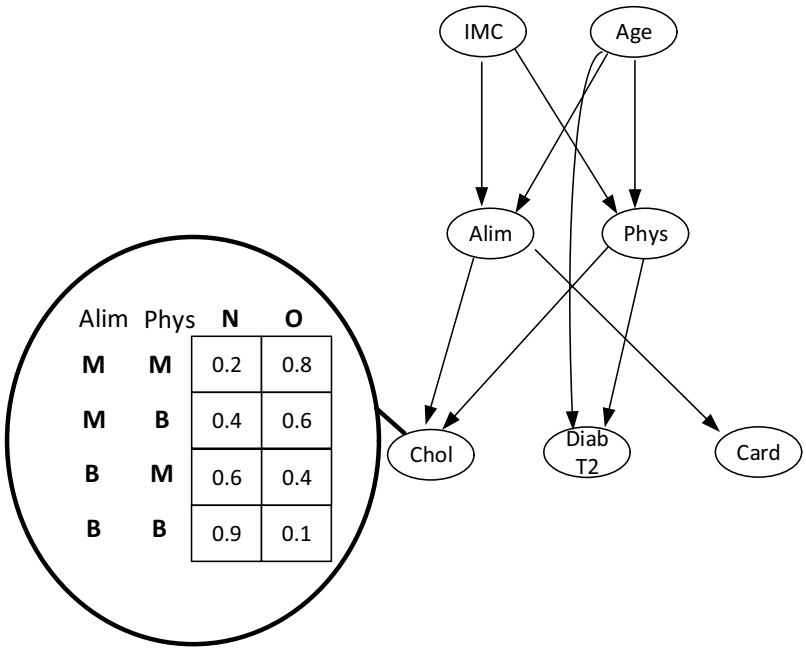

Figure 1-6. Exemple de réseau Bayésien

Un arc définit une relation de dépendance entre la variable aléatoire à l'origine de l'arc et la variable aléatoire à l'extrémité finale de l'arc. À chaque nœud est associée une table de probabilité : la probabilité d'avoir du cholestérol (nœud en bas à gauche du graphe de la Figure 1-6) dépend de deux variables qui sont Alim (pour Alimentation) et de Phys (pour le niveau d'activité Physique). Ainsi la Figure 1-6 définit une probabilité de 80% d'avoir du cholestérol pour un individu ayant une Alimentation mauvaise (Alim = M) et une activité physique mauvaise (Phys = M), et donc seulement de 20% de ne pas avoir de cholestérol.

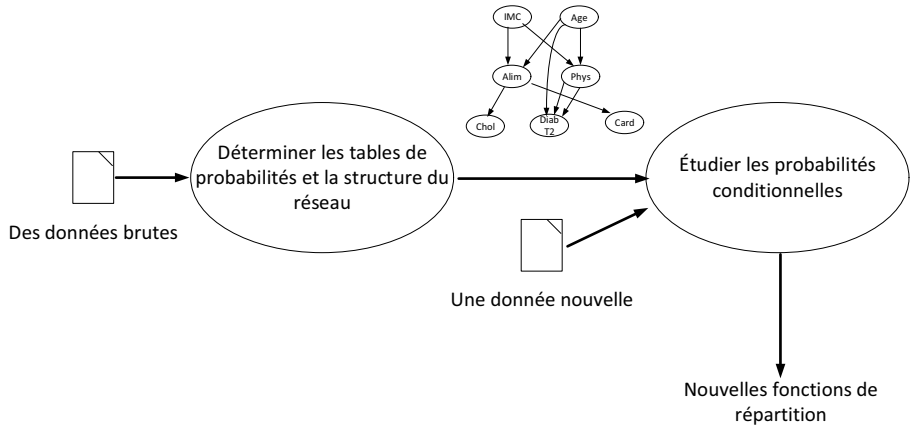

Figure 1-7. Principes de construction et d'utilisation d'un réseau Bayésien

En pratique, les données qui figurent dans ces tableaux sont issues de statistiques réalisées sur des données brutes (Figure 1-7) et doivent être réalisées avec des experts, seuls capables de construire le réseau bayésien adapté du problème. Une fois construit, le réseau Bayésien est utilisé comme un outil de calcul "pseudo-automatique" des distributions de probabilité en présence d'information nouvelle qu'il est possible "d'injecter" dans le réseau. Il est alors possible de l'utiliser pour répondre à des questions du type :

> "Sachant qu'un individu à une mauvaise alimentation et qu'il a des problèmes cardiaques, quelle est la probabilité que cet individu ait en plus du diabète ?".

Ce type d'outil est indispensable pour réaliser des calculs complexes et comprendre les dépendances entre les variables, sans se laisser induire en erreur par "l'intuition" ou une méconnaissance des phénomènes observés. L'exemple pseudo-médical dont il est question dans ce paragraphe sert de fil conducteur au chapitre 4.

1.4.3 Les méthodes de classification

Elles constituent un vaste domaine dans lequel il est possible d'inclure les réseaux de neurones, qui sont traités dans ce livre dans deux chapitres dédiés compte tenu de l'importance qu'ils ont pris dans le Machine Learning. Les méthodes de classification se divisent en deux branches distinctes : les méthodes supervisées pour lesquelles il faut disposer d'une base d'apprentissage (les arbres de décision appartiennent à cette catégorie) et les méthodes non supervisées, où il s'agit de créer des classes avec comme objectif de mieux analyser les données.

Les résultats récents en Machine Learning font apparaître les arbres de décision (et leur extension "random forest") comme un élément incontournable des méthodes de classification. Ceci tient essentiellement au fait que ces méthodes se montrent capables de traiter efficacement des données de grande taille dans des temps de calcul très courts et qu'elles fournissent ensuite un outil graphique de présentation des résultats qui facilite les échanges avec les experts du domaine, comme des médecins ou des biologistes.