IRM | institut du risk management

Mémoire présenté devant l'Institut du Risk Management
pour la validation du cursus à la Formation d'Actuaire
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le

Par : Nicolas MAISNIER et François LHENRI

Titre : Cadre éthique de l'utilisation des techniques de data science en actuariat

Analyse et mise en pratique d'une utilisation combinée du machine learning et de l'open data

Confidentialité : ☑ NON ☐ OUI (Durée : ☐ 1an ☐ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

Membres présents du jury de l'Institut du Risk Management :

Secrétariat :

Bibliothèque :

Entreprise :

Nom : CNP MALAKOFF HUMANIS

Signature et Cachet :

CNP ASSURANCES
Société Anonyme
Entreprise régie par le Code des Assurances
4 Place Raoul Dautry - 75716 PARIS Cedex 15

Directeur de mémoire en entreprise :

Nom : Pierre KRAMER Eve RAFFAELLI

Signature :

Invité :

Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels
(après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise

Quentin BOUDOUX
Directeur Technique Groupe

Signature(s) du candidat(s)

Saisissez du texte ici

RESUME

L'essor de la *data science* fait émerger de nouvelles données et techniques de traitement. De par leur métier de gestionnaire du risque, les assureurs et les actuaires peuvent fortement bénéficier de ces innovations notamment pour affiner leurs modèles de prédiction du risque. Néanmoins, les gains attendus ne doivent pas occulter les risques inhérents portés par ces nouvelles technologies, que ceux-ci découlent de sources externes (concurrence des géants du numérique, cyber-attaques), de la donnée (contraintes liées au RGPD), de la conception des algorithmes (effet boîte noire, biais induits) ou encore de leurs utilisations (individualisation du risque conduisant à une démutualisation). Tous ces risques – et tout particulièrement le dernier – doivent être pris en compte et nécessitent la mise en place d'un cadre éthique de l'utilisation de l'intelligence artificielle et du *big data* en assurance.

Afin de vérifier si les technologies de *data science* peuvent être en pratique utilisées de façon éthique, sans que cela n'affecte les bénéfices attendus, un cas d'usage a été étudié : apparier les données de l'*open data* de santé avec celles de l'assureur en conformité avec les principes réglementaires et éthiques, afin de prédire le risque de décès et ainsi d'améliorer le provisionnement en assurance prévoyance.

Pour ce cas d'usage, la base Open DAMIR a été choisie car elle propose des données anonymisées de remboursements de l'assurance maladie obligatoire. Cependant des retraitements conséquents ont été nécessaires du fait de l'existence de nombreuses valeurs manquantes. Un modèle d'imputation a donc été élaboré en testant différents algorithmes de façon à identifier celui qui permettrait de prédire au mieux les valeurs probables de ces données manquantes. En l'occurrence, l'algorithme des forêts aléatoires a permis, au sein du jeu de test, de retrouver plus de 90% des valeurs manquantes.

Suite à ces retraitements, les données issues de l'*open data* de santé ont pu être appariées avec celles de l'assureur. Pour cela, il a fallu notamment sélectionner les données pertinentes côté assureur, anonymiser ces données de façon à éviter tout risque de réidentification et assurer la correspondance entre la codification des deux bases. Par ailleurs, le modèle ayant pour cible la prédiction d'événements rares, il a été nécessaire de retravailler les données en effectuant un rééchantillonnage en amont de l'application de l'algorithme prédictif. Là encore, différents algorithmes ont été testés et la méthode des forêts aléatoires s'est de nouveau révélée être la plus adaptée pour prédire les décès.

Au final, la comparaison entre la prédiction et les sinistres réels démontre que l'utilisation des données d'*open data* couplée aux techniques de *machine learning* permet d'aboutir à un provisionnement bien meilleur que celui obtenu par les tables réglementaires, et ce, même en respectant un cadre éthique et réglementaire contraignant.

Mots clefs : *data science, big data, open data, intelligence artificielle, machine learning, RGPD, provisionnement risque décès, prédiction décès, imputation de valeurs manquantes, rééchantillonnage, k Nearest Neighbors, forêts aléatoires, réseau de neurones, régression logistique, machines à vecteurs de supports, analyse discriminante, boosting.*

ABSTRACT

The rise of data science brings out new data and processing techniques. Because of their risk management profession, insurers and actuaries can greatly benefit from these innovations, specifically to refine their risk prediction models. Nevertheless, the expected gains must not obscure the inherent risks induced by these new technologies, whether such risks arise from external sources (competition from Tech Giants, cyber-attacks) or from the data (constraints linked to the GDPR) or from the algorithms design (black box effect, induced bias) or from their use (individualization of risk leading to demutualization). All of these risks – and most importantly the latter – must be addressed and require the implementation of an ethical framework for the use of artificial intelligence and big data in the insurance sector.

In order to check whether data science technologies can actually be used in an ethical manner, without this affecting the expected benefits, a use case has been studied: matching data from open health data sources with insurer data, in compliance with the ethical and regulatory principles, in order to predict the risk of death.

For the purposes of this use case, the Open DAMIR database has been chosen, since it provides anonymized data relating to compulsory health insurance reimbursements. However, this database has required substantial reprocessing, due to many missing values. An imputation model has thus been developed, by testing various algorithms, in order to determine the one that would allow to best predict likely values for these missing data. In this case, the Random Forest algorithm has made it possible to find, within the test dataset, more than 90% of the missing values.

Further to such reprocessing, it was possible to match the data from open health data with the insurer's data. For this, it has been necessary to select relevant insurer's data, to anonymize such data so as to avoid any reidentification risk and to ensure correspondence between the codification of the two databases. Furthermore, since the model aims at predicting rare events, it has been necessary to rework the dataset, by resampling data prior to applying the predictive algorithm. Here again, various algorithms were tested in order to identify the most efficient one. The Random Forest method proved again to be the most relevant for predicting deaths.

At the end of this work, the comparison between the prediction of the model and the claims actually recorded shows that the use of open data along with machine learning techniques, makes it possible to obtain a much better provisioning than the one obtained by using regulatory mortality tables, even when maintaining strict compliance with a heavy ethical and regulatory framework.

Keywords: *data science, big data, open data, artificial intelligence, machine learning, GDPR, provisioning for death risk, death prediction, imputing missing values, resampling, k Nearest Neighbors, Random Forest, neural networks, logistic regression, support-vector machines, discriminant analysis, boosting.*

SOMMAIRE

REMERCIEMENTS

Nous souhaitons tout d'abord adresser nos sincères remerciements à nos responsables de mémoires, Mme. Eve RAFFAELLI et M. Pierre KRÄMER pour leur aide et leur soutien dans la rédaction de ce mémoire.

Par ailleurs, nous tenons à remercier tout particulièrement M. Arnold MEKONTSO dont les travaux autour de la base Open DAMIR nous ont beaucoup inspirés ainsi que Me. Lorraine MAISNIER-BOCHÉ, avocat spécialiste des questions de protection des données à caractère personnel, qui nous a fait bénéficier de son expertise juridique. Nous remercions également M. Ievgen SAVIN pour son aide dans la réalisation de notre code en Python.

Un grand merci enfin à M. Grégoire CARO, Dr. Rémi DESBUQUOIS, Dr. Elie de PANAFIEU et M. Dorian POTTIER pour leur relecture assidue et leurs suggestions pertinentes.

Une pensée toute particulière pour nos fils Arthur et Alexis, tous deux nés au moment de débuter ce mémoire et qui ont fait tout leur possible pour nous aider à ne pas trop rester concentrés sur nos travaux.

INTRODUCTION GENERALE

L'article 1er de la Loi Informatique et Liberté de 1978 indique : « *L'informatique doit être au service de chaque citoyen. […] Elle ne doit porter atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques.* ». Il existe ainsi une ambivalence de l'intelligence artificielle : comme toute science, elle peut être utilisée à bon ou à mauvais escient. Aussi, la question cruciale n'est pas de savoir si elle est intrinsèquement bonne ou mauvaise mais de déterminer comment utiliser un tel outil avec conscience. C'est ce qui justifie par exemple l'établissement d'une éthique de l'utilisation de l'intelligence artificielle.

Comme beaucoup de secteurs, l'assurance est bien évidemment impactée par les innovations apportées par la *data science*, et ce d'autant plus qu'il s'agit d'une activité au sein de laquelle le traitement des données, la connaissance des informations et la prédiction de métriques prospectives sont déjà au cœur de ses métiers. Il s'agit donc potentiellement d'une formidable opportunité. Dans le même temps, l'assurance présente une responsabilité sociale forte, tout particulièrement en France : elle gère l'épargne d'une grande part de la population, elle est garante de l'accès aux soins de chacun, elle sécurise une grande partie des déplacements et elle est indispensable pour accéder à la propriété. Il est par conséquent indispensable que le marché de l'assurance s'interroge sur l'usage qu'il peut faire de ces nouveaux outils afin de s'assurer qu'il ne crée pas de préjudice à tout ou partie de la population des assurés.

En particulier, l'arrivée des technologies connectées (« *Internet of things* ») et l'accès à d'immenses bases de données nouvelles permises par l'*open data*, accompagnés des algorithmes de *machine learning* ont potentiellement la faculté d'évaluer le risque à la maille de l'individu. Cela représente un pouvoir de prédiction considérable mais cela pourrait remettre en cause également le principe fondateur de l'assurance : la mutualisation. Une réflexion sectorielle autour d'une utilisation éthique de l'intelligence artificielle est donc nécessaire afin que les assureurs ne fassent pas disparaître le fondement même de leur existence.

L'objectif de ce mémoire est ainsi de répondre à la problématique suivante : est-il possible de définir un cadre éthique de l'utilisation des techniques de *data science* en actuariat qui ne bride pas excessivement les gains apportés par ces innovations ? Pour cela, d'une part une analyse des opportunités et des risques portés par les technologies du *big data* et de l'intelligence artificielle pour les assureurs a été conduite afin de dessiner les contours d'une utilisation éthique de ces innovations (cf. partie I). D'autre part un cas d'usage concret est proposé en utilisant des techniques de *machine learning* et une base d'*open data* de santé recoupée avec les données d'un assureur pour prédire le risque de décès sur un portefeuille d'assurés (cf. partie II).

Ce mémoire se situe par conséquent à la croisée des chemins entre les trois formations proposées par l'IRM : l'actuariat (provisionnement du risque décès), la data science (intelligence artificielle et big data) et management des risques en entreprise (classification des risques et mesure pour les limiter).

PARTIE I :

INTERETS ET RISQUES DE L'INTELLIGENCE ARTIFICIELLE ET DU BIG DATA SUR LE MARCHE DE L'ASSURANCE

INTRODUCTION

La *data science* (ou « science des données ») est une matière scientifique relativement récente correspondant à l'utilisation de toutes les disciplines statistiques, mathématiques et informatiques dans l'objectif d'extraire de la connaissance à partir d'un ensemble de données. Les nouvelles techniques regroupées sous le terme de *data science* concernent à la fois les algorithmes d'intelligence artificielle servant à traiter la donnée et la récupération massive de données permises par le *big data*, ainsi que, par extension, l'ensemble des utilisations de ces innovations (cf. § I.1).

Si aujourd'hui les termes de *data science*, d'intelligence artificielle et de *big data* sont particulièrement à la mode, c'est parce que les innovations qu'ils proposaient depuis plusieurs années deviennent accessible grâce à la puissance de calcul des ordinateurs actuels. De plus, les produits et services qui peuvent en découler sont particulièrement prometteurs, notamment pour le secteur de l'assurance qui présente un besoin fondamental de traitement de l'information (cf. § I.2). Néanmoins, l'attrait de ces nouvelles technologies ne doit pas faire oublier le risque inhérent qu'elles comportent (cf. § I.3).

Dès lors, au-delà de l'intérêt évident qu'elles sont susceptibles d'apporter, il semble primordial de définir un cadre éthique d'utilisation des innovations portées par la *data science* afin de ne pas involontairement enclencher une machine infernale qui finirait par détruire les principes qui fondent le métier d'assureur. Dans ce contexte, le terme « éthique » englobe tant de respecter la réglementation que de prévenir des comportements certes licites mais pouvant mettre à mal la relation avec les assurés ou les valeurs portées par l'assurance. Pour illustrer comment pourrait se définir un tel cadre, un cas d'usage d'une utilisation raisonnée de l'intelligence artificielle et du *big data* (au travers de l'*open data* de santé) sera présenté (cf. § I.4). Ce cas consiste à utiliser des algorithmes de *machine learning* sur des données de l'assureur enrichies par des informations d'*open data* de santé afin de prédire les décès dans le cadre du provisionnement en prévoyance.

I.1. EMERGENCE DE LA DATA SCIENCE

I.1.1. Intelligence artificielle

L'intelligence artificielle (ou « IA ») correspond à un « *programme informatique visant à effectuer, au moins aussi bien que des humains, des tâches nécessitant un certain niveau d'intelligence* » [VILLANI, C. (2018)]. Si l'IA est un sujet très à la mode aujourd'hui, il ne s'agit néanmoins pas d'un concept nouveau : dès 1950, Alan Turing pose la question de la faculté d'une machine à penser dans son célèbre test [TURING, A. M. (1950)].

Plus précisément, l'IA consiste à réaliser des « *programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisantes par des êtres humains car elles demandent des processus mentaux de haut niveau tels que*

l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique » [MINSKY, M. L. (1967)]. Cette définition permet de mettre en lumière l'importance du concept d'apprentissage par la machine. C'est ce concept qui permet de distinguer une IA « faible » purement algorithmique et dédiée à une tâche précise, d'une IA « forte » qui est auto-apprenante et donc capable d'évoluer selon le contexte.

I.1.2. Machine learning

Au sein des différentes technologies qui peuvent être caractérisées d'IA, les algorithmes de *machine learning* (ou « apprentissage autonome ») prennent depuis les années 1980 une place toute particulière. Il s'agit d'algorithmes capables d'acquérir de manière automatique de nouvelles connaissances. D'ailleurs, beaucoup d'organismes, en particulier l'ACPR[1], restreignent le concept d'IA aux algorithmes de *machine learning* [ACPR (2018)].

Concrètement, le *machine learning* permet aux programmes informatiques d'apprendre à partir des données qui leur sont injectées ou bien à partir de leur propre expérience afin de faire une prédiction ou de prendre une décision en se basant sur cet apprentissage des données plutôt que sur un algorithme explicite. Par essence, ces programmes sont donc évolutifs car sans cesse exposés à de nouvelles informations qui viennent compléter et affiner leur apprentissage.

Les algorithmes de *machine learning* sont généralement divisés en trois catégories [TANWAR, S. (2019)] :

- L'apprentissage supervisé (« *supervised learning* ») concerne une problématique classique en mathématiques statistiques puisqu'il s'agit d'estimer la valeur d'une variable inconnue à partir d'un certain nombre de variables explicatives. Il s'agit de régression lorsque la variable à estimer est une grandeur quantitative et de classification quand il s'agit d'une variable qualitative (ie. définissant une catégorie ou un attribut).

- L'apprentissage non supervisé (« *unsupervised learning* ») consiste à modéliser des liens ou des distributions sous-jacentes existants au sein de la base de données. Par exemple, le « *clustering* » permet de regrouper les données par groupe homogène et l'association consiste à découvrir des règles ou des liens de causalité (dans le but de réduire la dimension des données en termes de degrés de liberté).

- L'apprentissage par renforcement (« *reinforcement learning* ») représente ce qui s'apparente le plus littéralement à de l'intelligence artificielle puisque ce sont des algorithmes qui vont entraîner un programme sans intervention humaine (la plupart du temps via un jeu de récompenses et de punitions).

[1] Autorité de Contrôle Prudentiel et de Résolution – régulateur chargé de la surveillance des assurances en France

I.1.3. Deep learning

A partir des années 2010, les technologies de *machine learning* ont évoluées vers les technologies de *deep learning* (ou « apprentissage profond ») capables de traiter des données plus complexes et moins structurées avec un haut degré d'abstraction et de restituer leurs résultats sous forme d'images, de sons ou de textes.

L'algorithme de *deep learning* est un sous-type de *machine learning* qui utilise des modèles de type réseau de neurones artificiels : les tâches sont organisées en couches successives dont les sorties des premières viennent alimenter les entrées des suivantes. Ce schéma est nommé ainsi car il est censé reproduire le fonctionnement des neurones au sein d'un cerveau humain.

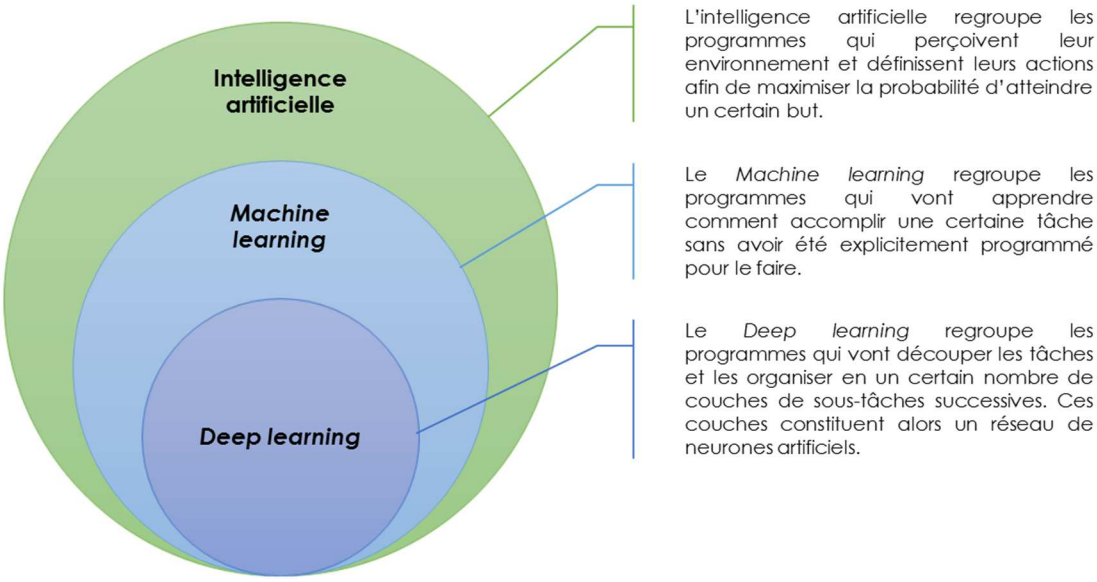


Figure 1 – Imbrication entre IA, machine learning et deep learning [TANWAR, S. (2019)

Les réseaux de neurones ne sont pas des algorithmes particulièrement récents puisque les premiers travaux sur le sujet ont émergé dès 1943 [MC CULLOCH, W. et PITTS, W. (1943)]. Cependant, avant la dernière décennie, ils étaient difficilement utilisables en pratique à cause du manque de puissance de calculs des ordinateurs. L'avènement du *big data* (cf. § I.1.4.) a permis de mettre sur le devant de la scène les technologies de *deep learning* qui se sont rapidement imposées comme les types d'IA les plus populaires et les plus marquantes[2].

[2] Elles sont notamment à la base de la voiture autonome, de l'assistant virtuel, de la colorisation d'image en noir et blanc, de l'IA Alpha Go, de la reconnaissance d'objet ou de visage en temps réel, etc.

I.1.4. Big data

Le terme « *big data* » (ou « mégadonnées ») est défini dans la terminologie officielle comme des « *données structurées ou non dont le très grand volume requiert des outils d'analyse adaptés* » [JORF (2014)]. En pratique, les concepts recouverts par ce terme ne concernent pas uniquement les données mais recouvrent également les techniques spécifiques de traitement de ces données, les technologies qui ont permis ces traitements et par extension les services rendus par l'exploitation de ces données. Il convient donc d'être précis en parlant de *big data* afin de bien spécifier s'il s'agit des données, des processus de traitement ou encore de l'utilisation de ces données.

Ainsi, la différence fondamentale entre un traitement de données classique et l'innovation apportée par le *big data* se situe dans cette notion de « très grand volume ». Dès lors, il convient de définir cet ordre de grandeur. En 2001, l'analyste Doug Laney du cabinet Gardner définissait la désormais célèbre règle des 3V [LANEY, D. (2001)] permettant de dessiner les enjeux de cette nouvelle matière qui allait prendre le nom de *big data* :

- Volume : Le volume mondial des données numériques augmente de façon exponentielle, ce qui a été théorisé par Moore[3], Kryder[4] ou encore Nielsen[5]. Fin 2018, ce volume était estimé à 33 zettaoctets[6] et devrait continuer de se multiplier par 3,5 chaque année [TASSET, M. (2019)].

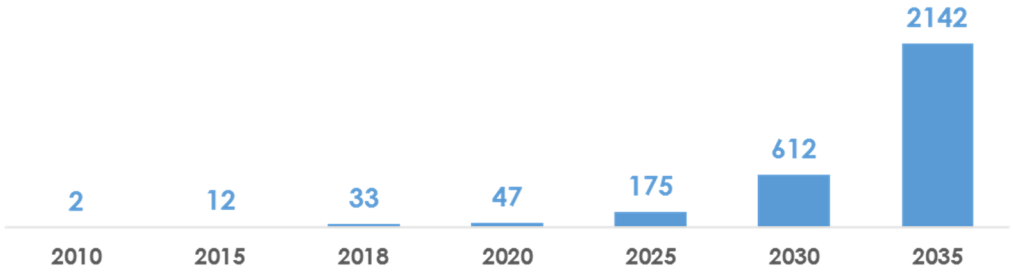


Figure 2 – Prévision du volume de données créées dans le monde depuis 2010 (en zettaoctets)

- Variété : Ne sont plus traitées seulement des bases de données structurées mais également des textes, des images, des sons ou des métadonnées, récupérés auprès d'une multitude de sources que les individus en soient conscients ou non. Cette variété est notamment rendue possible par l'émergence des objets connectés (aussi appelés *Internet of things*) qui ont envahi le quotidien, de la montre à la fourchette, en passant par la voiture ou l'assistant personnel.

[3] Gordon E. Moore a énoncé en 1975 sa loi prédisant que le nombre de transistors intégrés dans un microprocesseur doublerait tous les deux ans.
[4] De manière analogue à la loi de Moore, Mark Kryder a prédit en 2005 que le coût de l'espace de stockage de l'information numérique diminue de moitié tous les deux ans.
[5] De façon similaire, la loi de Jakob Nielsen, énoncée dès 1998, prédisait que le débit que le débit de connexion à un réseau augmenterait de 50% chaque année.
[6] Un zettaoctet équivaut à un milliard de téraoctets. Un téraoctet équivaut à 2^{40} octets soit plus de 10^{12} octets. L'octet correspond à l'unité numérique fondamentale, nécessaire pour coder un caractère.

- Vélocité : L'accumulation, le traitement et l'analyse des données s'effectuent de plus en plus en temps réel. Cette dimension intègre aussi la rapidité d'évolution des techniques et des innovations qui bousculent sans cesse cette nouvelle matière.

Depuis, le nombre de V permettant de caractériser le *big data* s'accroit avec le lyrisme des responsables marketing (Véracité, Valeur, Volatilité, Validité…). Néanmoins, ce qu'il manque encore à la règle des 3V, 4V ou 5V pour véritablement définir le *big data* réside dans sa finalité, à savoir utiliser cette masse de données afin d'améliorer de façon inédite la vision sur une situation passée, présente ou future afin d'aider à la prise de décision.

Plus concrètement, le concept de *big data* englobe généralement les technologies de production des données (tout particulièrement avec l'émergence des objets connectés), le stockage et les capacités de calcul permise par le « *cloud computing* »[7] et les nouvelles techniques d'intelligence artificielle qui viennent se nourrir de cette masse de données et de cette capacité de calcul.

I.1.5. Open data : une nouvelle source de données

I.1.5.1. Qu'est-ce que l'open data ?

Les *open data* (ou « données ouvertes ») correspondent à des données auxquelles l'accès, l'exploitation et la réutilisation sont publics et libres de droit. Selon la définition donnée par l'Open Knowledge Foundation[8] en 2005, l'*open data* doit remplir trois critères [okfn.org] :

- **Disponibilité et accès :** Les données doivent être pleinement accessibles, moyennant un coût de reproduction raisonnable. De préférence, elles se téléchargent sur Internet. Le format doit être confortable et modifiable.

- **Réutilisation et redistribution :** Les données doivent être fournies sous des conditions permettant la réutilisation et la redistribution, incluant le mélange avec d'autres ensembles de données.

- **Participation universelle :** Tout le monde doit être en mesure d'utiliser, de réutiliser et de redistribuer les données. Il ne doit y avoir aucune discrimination concernant les fins d'utilisation, ou contre des personnes ou des groupes. Par exemple, des restrictions non commerciales qui empêchent l'utilisation commerciale, ou les restrictions d'usage à certains secteurs, ne sont pas compatibles avec l'*open data*.

[7] Le *cloud computing* (ou « informatique en nuage ») représente une structure informatique dans laquelle les services ne sont pas stockés sur un serveur local mais sont décentralisés sur un ou plusieurs serveurs distants via Internet. Un service peut correspondre à l'utilisation d'un logiciel (SaaS pour *Software as a Service*), d'un environnement d'exécution de calculs informatique (PaaS pour *Platform as a Service*) ou d'un serveur de stockage (IaaS pour *Infrastructure as a Service*). Il s'agit d'un changement fondamental dans la philosophie d'Internet car l'information n'est plus décentralisée au niveau de l'ordinateur de chaque utilisateur mais elle est au contraire centralisée au sein de quelques grands *data centers*.
[8] L'Open Knowledge Fondation est une association à but non lucratif ayant pour objectif de promouvoir le « savoir libre ».

Néanmoins, cette démarche d'ouvrir l'accès aux jeux de données impose de les retravailler afin qu'ils ne contiennent aucune donnée à caractère personnel. C'est une évidence pour les données directement identifiantes mais la possibilité de croisement oblige également à éviter toute identification indirecte (cf. § I.3.2.). C'est pourquoi les données ouvertes des *open data* qui concernent des individus sont souvent fortement agrégées de manière à éviter au maximum le risque de réidentification.

I.1.5.2. Open data en France

Par nature, le principe d'*open data* concerne principalement le secteur public : l'Etat et les collectivités sont engagés dans une démarche de partage des données publiques s'appuyant notamment sur le droit d'accès aux documents administratifs [Code des relations entre le public et l'administration] qui considère que les données produites ou détenues par les administrations, dans le cadre de leurs missions de service public, doivent être mises à disposition (à l'exclusion des informations personnelles et de celles couvertes par les secrets légaux comme la sécurité nationale).

La France est classée au quatrième rang mondial de l'ouverture des données publiques [okfn.org], derrière Taiwan, le Royaume-Uni et l'Australie. Cette mise à disposition est actuellement dirigée, sous l'autorité du Premier Ministre, par la mission Etalab, à l'origine de la refonte du portail *data.gouv.fr*. Cette mission a pour but de mettre à disposition des données publiques, conformément au principe général de réutilisation libre, facile et gratuite [CIRCULAIRE (2011)]. De nombreuses données publiques sont déjà disponibles, notamment des jeux de données de l'INSEE ou celui du SNIIRAM[9]. Elles sont complétées par les informations publiées par les collectivités territoriales, notamment via l'association Opendata France.

Par ailleurs, certaines entreprises privées doivent publier des données dans le cadre de la réglementation spécifique de leur secteur. C'est le cas dans le domaine de l'énergie [COMMISSION DE REGULATION DE L'ENERGIE (2017)] ou celui des transports [transport.data.gouv.fr]. Plus spécifiquement, le CCSF[10] proposait que les assureurs-vie viennent enrichir les données du GIP Union Retraite avec les données sur leurs produits d'épargne retraite supplémentaire afin que le rassemblement de ces informations puisse permettre de gérer la problématique des contrats en déshérence [CCSF (2020)]. Enfin, dans le secteur associatif, certains acteurs ont pour vocation de rendre disponibles à tous des informations et des jeux de données : *OpenStreetMap* pour les données géographiques, *Open Food Facts* sur les produits alimentaires ou encore *nosdeputes.fr* sur le suivi de l'activité des députés.

[9] Système National Inter Régimes d'Assurance Maladie (cf. § II.1.1.1.)
[10] Comité Consultatif du Secteur Financier, organe consultatif créé par l'article 22 de la loi n° 2003-706 du 1er août 2003 pour analyser les problématiques entre les acteurs et leurs clients

I.1.5.3. Open data de santé

I.1.5.3.1. Health data hub

Dans le domaine de la santé, la démarche d'*open data* se cristallise sous la forme d'un groupement et d'une plateforme des données de santé tous deux nommés *Health data hub* [Code de la santé publique]. Celui-ci prévoit deux formats aux finalités et aux conditions d'accès bien distinctes.

D'une part, le SNDS[11] qui contient des bases de données potentiellement identifiantes et dont l'utilisation est particulièrement encadrée [LOI (2016)] : autorisation préalable de la CNIL après avis du CEREES[12] et de l'INDS[13].

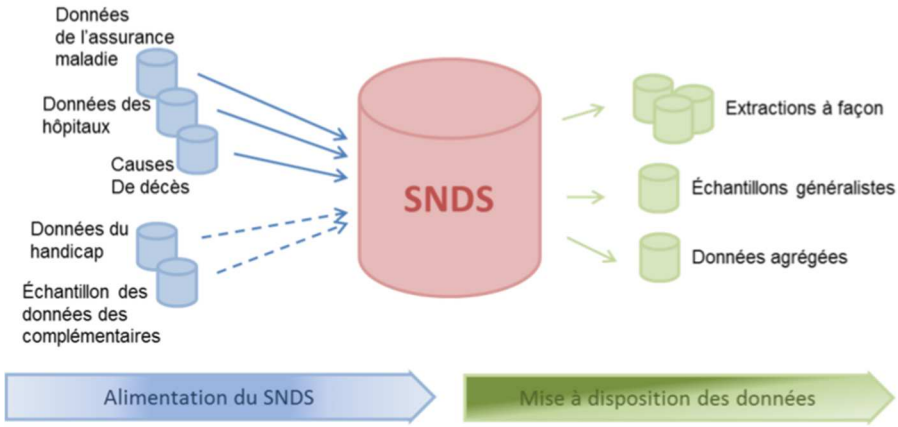


Figure 3 – Construction de la base principale du SNDS [indsante.fr]

D'autres part, les bases qui proposent des jeux de données anonymisées sous forme de statistiques agrégées accessibles à tous (aucune condition d'accès), parmi lesquelles :

- Open DAMIR[14] (dépenses d'assurance maladie) ;
- Open Medic (prescriptions de médicaments) ;
- Open Statines (prescriptions de statines[15]) ;
- Open PHMEV (prescriptions hospitalières) ;
- Open Bio (biologies médicales) ;
- Open LPP[16] (dispositifs médicaux).

[11] Système National des Données de Santé
[12] Comité d'Expertise pour les Recherches, les Études et les Évaluations dans le domaine de la Santé
[13] Institut National des Données de Santé, devenu le groupement *Health Data Hub* en 2019
[14] Dépenses d'Assurance Maladie Inter Régimes
[15] Médicament servant à lutter contre l'hypercholestérolémie
[16] Liste des Produits et Prestations

I.1.5.3.2. Conditions d'accès au SNDS[17]

Les bases du SNDS étant principalement composées de données de santé et médicales potentiellement identifiantes, leur accès est particulièrement encadré [drees.solidarites-sante.gouv.fr]. En premier lieu, il convient de prouver que la finalité du traitement des données présente un intérêt public. Une demande d'accès peut présenter un intérêt privé ou commercial dès lors qu'elle justifie également d'un intérêt public et que celui-ci est prépondérant par rapport aux autres. D'après le rapport « *Expertise juridique sur l'intérêt public dans le contexte des données de santé* » [SIMMONS&SIMMONS (2017)] commandé par l'INDS, il n'y a pas de critères bien définis à remplir mais plutôt un faisceau d'indices permettant au *Health data hub* et à la CNIL de déterminer s'il y a ou non un intérêt public selon trois catégories :

- **Le statut du demandeur :** Un organisme public ou de recherche indépendant dispose d'une présomption d'intérêt public par rapport à un organisme privé.

- **La finalité de la demande :** Les finalités de recherche ou de santé publique ainsi que les demandes de données agrégées seront plus facilement acceptables que les finalités commerciales. Il convient surtout d'éviter que la demande ne soit disproportionnée par rapport au besoin.

- **Le financement :** Une demande financée par des acteurs publics pourra potentiellement être mieux accueillie qu'une demande financée par le secteur privé.

Enfin, ce rapport se conclut par des exemples de finalités selon leur droit d'accès :

- **« *Les usages formellement interdits par la loi* » :** les promotions commerciales de produits de santé, l'exclusion de garantie d'assurance (sélection médicale), la modification de cotisations ou de primes d'assurance (tarification).

- **« *Les usages pour lesquels un intérêt public est a priori exclu* » :** le développement commercial, le ciblage de patient, la détection d'usagers ou d'assurés pour une visite médicale.

- **« *Les usages pouvant justifier le cas échéant un intérêt public par application du faisceau d'indices* » :** les traitements effectués pour répondre à une demande de la réglementation, le meilleur suivi médical des patients par les professionnels de santé, la meilleure prise en charge par la Sécurité Sociale, la prévention, la recherche médicale, l'information du grand public.

- **« *Les usages pour lesquels un intérêt public est présumé* » :** les traitements financés par des fonds publics et à l'initiative d'une personne publique, les traitements imposés par l'Etat, par la loi ou par la réglementation, la recherche.

[17] La procédure d'accès aux données du SNDS est reprise en annexe 1.

I.2. INTERETS DE CES INNOVATIONS

I.2.1. Principaux objectifs pour les assureurs

Une étude du secteur [LA FABRIQUE D'ASSURANCE (2019)] a mis en lumière les principales attentes des assureurs vis-à-vis des innovations apportées par le *big data* et l'IA :

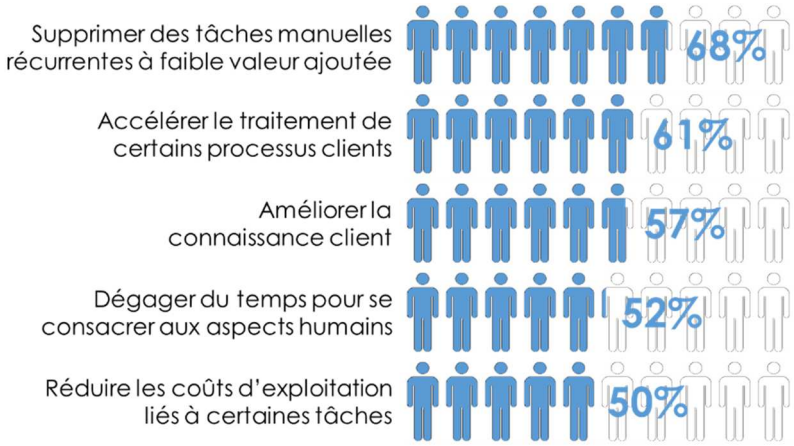


Figure 4 – Principales attentes des assureurs vis-à-vis du big data

Il est ainsi possible de répartir les objectifs des assureurs en trois catégories.

- **Améliorer la relation client** : l'idée principale consiste à améliorer la segmentation des clients afin de leur proposer des offres plus pertinentes et des réponses mieux adaptées selon leur profil et leur situation tout en leur offrant un service plus fluide.

- **Automatiser, optimiser ou supprimer certaines activités** de back office (comme la gestion de projet) voire front office (comme la gestion basique des sinistres) afin de recentrer l'activité humaine sur des actions à forte valeur ajoutée et faire diminuer les coûts. Cela passe également par la fiabilisation des processus, notamment vis-à-vis des dispositifs de lutte contre la fraude.

- **Construire une nouvelle offre de services centrés sur la prévention des risques :** l'assureur sort de son « *rôle d'indemnisation au profit d'une logique d'accompagnement* » [GRÄFE, S. (2019)]. Grâce à cette nouvelle masse de données, l'assureur est potentiellement en mesure de prédire le risque et donc de fournir une offre de service assurantielle personnalisée fondée sur la prévention. Cela peut s'exprimer par la valorisation de certains comportements (grâce à des réductions de tarif) ou bien par le déclenchement d'actions en amont de la survenance du risque (prise de rendez-vous avec un médecin, changement de pièces d'un véhicule, etc.).

I.2.2. Utilisations possibles de l'open data

Concernant spécifiquement les données d'*open data* de santé, celles-ci peuvent servir dans de nombreux cas d'usage pour les assureurs. En particulier :

- **L'évolution de la consommation en santé :** Les données disponibles sont à un niveau assez fin et disposent de suffisamment d'antériorité pour qu'il soit possible de suivre l'historique de l'évolution de la consommation de soins (que ce soit en termes de fréquences ou de montants unitaires) par acte, âge, sexe, région...

- **Les études d'impacts :** Il est possible d'observer des évolutions dans la consommation de santé, suite à l'évolution de la réglementation, par exemple. De même, des projections peuvent être réalisées plus finement en utilisant de données publiques, sans pour autant occulter les spécificités du portefeuille d'un assureur donné.

- **La détection de la fraude :** En recoupant les données publiques et les données propres à l'assureur via une IA, il est possible de détecter des cas de suspicion de fraude qui n'auraient pas été repérés autrement.

- **La tarification :** Les données peuvent être utilisées pour améliorer la finesse de segmentation tarifaire. Ce point demeure néanmoins délicat puisque l'utilisation de l'IA comporte de nombreux risques (cf. § I.3) tandis que la tarification constitue un point crucial du contrat d'assurance. Il reste en revanche envisageable d'utiliser ces techniques afin d'optimiser des barèmes, à critères de tarification fixés par exemple.

- **Le provisionnement :** Les données peuvent permettre à l'assureur de mieux évaluer son risque. Comme la tarification, il faut veiller à conserver le sens de ce qui est produit, mais il est possible d'employer des algorithmes de type *machine learning* pour calculer des provisions plus justes, dans la limite de la réglementation et d'essayer d'anticiper des effets de dérive de sinistralité.

- **La prévention :** Si ces innovations permettent de mieux prévoir le risque, alors l'assureur peut s'en servir pour amplifier et généraliser une transformation qui est déjà observée sur le marché de l'assurance santé : le passage du payeur aveugle au garant d'un plan de prévention optimal pour ses assurés.

I.2.3. Impacts pour les actuaires

De par leur métier de gestionnaire du risque, les assureurs sont habitués à utiliser des techniques statistiques sur de grands volumes de données. Pour autant, est-il correct de dire que les actuaires faisaient du *big data* avant le *big data* ? Ce n'est pas tout à fait exact car les données comme les outils utilisés répondent à une logique fondamentalement différente.

Traditionnellement, pour leurs travaux de tarification ou de provisionnement, les actuaires travaillent sur leurs propres bases de données d'assurés (éventuellement complétées par l'utilisation de tables réglementaires) via des outils statistiques ou stochastiques mais toujours bien encadrés par la théorie mathématique traditionnelle. Pour être utilisées, les données doivent être qualifiées, c'est-à-dire qu'il faut s'assurer de leur qualité et de leur forte densité en information utile.

L'innovation apportée par le *big data* et l'IA concerne à la fois les données collectées et les techniques de traitement de ces données. En effet, les données sont désormais récupérées en masse en dehors des canaux de récupération classiques (nouvelles sources de données extérieures ou nouveaux capteurs). En outre, les techniques utilisées à présent consistent à faire analyser la masse des données (« *data crunching* ») par un programme de *machine learning* qui va apprendre empiriquement la structure d'interdépendance des données, et ce sans aucun biais cognitif sur leur signification. Le programme ainsi entraîné pourra alors effectuer des extrapolations sur cette base de données afin par exemple de trouver des valeurs manquantes ou bien de faire des prédictions. Ainsi, contrairement aux méthodes actuarielles classiques, c'est bien la quantité des données et non leur qualité qui va représenter le carburant des algorithmes de *big data*.

Il s'agit donc d'un passage d'un univers où l'actuaire utilise les mathématiques pour modéliser (de manière plus ou moins complexe) la réalité à un univers où le *data scientist* utilise des programmes autonomes pour trouver un modèle dans les données. Par ailleurs, le fonctionnement des algorithmes de *machine learning* s'affranchit généralement des notions de moyenne et d'espérance[18] sur lesquelles se fondent les approches statistiques classiques. En conséquence, les modèles ne sont plus limités par la loi des grands nombres et peuvent effectuer des analyses à des mailles beaucoup plus fines.

Concrètement, les modèles de l'actuaire, qui sont construits de manière *top-down* (typiquement, le risque de décès est d'abord évalué à partir d'une loi de mortalité par sexe et tranche d'âge ; puis il est affiné via des coefficients correctifs comme le fait de fumer, l'indice de masse corporelle ou l'affection longue durée), se verraient dans un premier temps renforcés par l'IA qui viendrait mettre en lumière des structures beaucoup plus complexes au sein des données (en particulier les jeux de corrélations). Le vrai caractère disruptif de la *data science* apparaîtrait dans un second temps lorsqu'elle permettra l'émergence de nouveaux modèles *bottom-up* qui se baseraient directement sur chaque individu afin de simuler son comportement, son état de santé et *in fine* son décès ou sa survie[19].

L'intérêt pour l'actuaire apparait alors très clairement : non seulement ces nouvelles techniques peuvent permettre d'affiner la prédiction du risque couvert par un contrat d'assurance, mais surtout elles peuvent potentiellement le faire à l'échelle de l'individu (contrairement aux méthodes actuarielles classiques). Ainsi, la connaissance par

[18] Toutefois, les fonctions d'évaluation de l'erreur de prédiction conservent des formes relativement classiques (par exemple, l'erreur moyenne quadratique pour les algorithmes de régression).
[19] Il est possible de comparer ce changement de paradigme au passage d'une évaluation de risque de catastrophes naturelles basé uniquement sur une loi d'expérience à un modèle basé sur la modélisation stochastique de phénomènes physiques du risque étudié.

l'assureur de l'assuré, de ses comportements et de leurs conséquences devient plus importante, rectifiant l'asymétrie d'information et permettant une meilleure segmentation de son portefeuille de risque. Par exemple, l'assureur chinois Ping An utilise l'IA pour déterminer à partir d'une photographie l'indice de masse corporelle et ainsi mesurer une partie des risques non dévoilés par ses assurés avec une efficacité de 90% [PING AN INSURANCE (2019)].

Cependant, afin d'aboutir à l'amélioration de l'évaluation des risques promises, il est indispensable de développer la collaboration entre l'actuaire et le *data scientist*. En effet, comme expliqué ci-dessus, l'actuaire a besoin des nouveaux outils offerts par l'IA et le *big data* tandis que le *data scientist* ne dispose pas de la compétence statistique[20] permettant d'analyser la distribution de lois. Or cette analyse est fondamentale pour l'évaluation du risque car elle permet l'étude des événements rares et le calcul des ratios de solvabilité.

I.3. PRINCIPAUX RISQUES LIES A CES INNOVATIONS

Malgré les intérêts indéniables potentiellement proposés par la *data science*, il est primordial de bien identifier les risques inhérents portés par ces nouvelles techniques[21]. Dans le cadre de ce mémoire, nous proposons une classification de ces risques selon quatre catégories : les risques extérieurs (cf. § I.3.1.), les risques liés aux données (cf. § I.3.2.), les risques liés aux algorithmes (cf. § I.3.3.) et les risques liés aux utilisations (cf. § I.3.4.).

I.3.1. Risques extérieurs

I.3.1.1. Risque concurrentiel

Le secteur de l'assurance et en particulier le marché européen ont accumulé un certain retard dans l'utilisation des technologies de *big data* et d'IA qui ont rendu ses acteurs dépendants des technologies proposées notamment par les GAFAM[22] américains et les BATX[23] chinois. Or ces entreprises risquent de ne pas adopter les mêmes stratégies économiques ou éthiques et pourraient à tout instant passer du statut de partenaire à celui de concurrent. De même, les assureurs traditionnels ne sont pas à l'abri de voir émerger une concurrence profondément disruptive s'appuyant pleinement sur ces nouvelles technologies, comme ce fut le cas dans d'autres secteurs avec les NATU[24].

[20] En particulier, la plupart des algorithmes de *machine learning* s'intéressent uniquement la valeur la plus probable et ne donne pas d'information sur la distribution des valeurs possibles.
[21] Ne sont évoqués ici que les risques liés au monde de l'entreprise. En toute rigueur, il faudrait également citer la problématique écologique. En effet, les algorithmes d'IA sont très consommateurs en énergie machine (un cerveau humain fonctionne avec une puissance de 20 watts alors qu'un programme comme Alpha Go de Google réclame cinq fois plus de puissance). De même, l'avènement du *cloud computing* a multiplié les flux de données (notamment avec les vidéos en streaming via des plateformes comme YouTube ou Netflix) qui sont très couteux en énergie. Nous aurions pu également évoquer les nouveaux risques liés aux services rendus possibles par l'IA (comme ceux liés à la voiture autonome).
[22] Google, Apple, Facebook, Amazon, Microsoft
[23] Baidu, Alibaba, Tencent, Xiaomi
[24] Netflix, Airbnb, Tesla, Uber

A titre d'exemple, Google a récemment investi dans la société Ethos qui utilise la *data science* pour prédire et traiter instantanément les demandes d'assurance vie. Baidu a quant à lui sa propre filiale Bai An Insurance pour couvrir les risques liés aux transactions sur Internet. Enfin, Alibaba a fondé fin 2013 une *joint-venture* avec d'autres assureurs chinois pour créer le premier assureur entièrement en ligne en Chine.

Par ailleurs, dans l'hypothèse où ces innovations ne seraient détenues que par quelques grands acteurs, cela pourrait entraîner une rupture dans l'équilibre de la concurrence sur le marché mondial. Une telle situation aboutirait à une augmentation des prix, une faible maîtrise des données ou de leurs traitements et des possibilités d'audit et de contrôles limitées, sans parler des problématiques de souveraineté vis-à-vis des services proposés par des entreprises issues de pays dont la législation ne présente pas les mêmes garanties que celles prévues par les normes européennes [ACPR (2018)].

I.3.1.2. Risque de cyberattaque

Le terme de cyberattaque regroupe un certain nombre de pratiques (*ransomware*, *phishing*, intrusion, etc.) qui ont en commun de s'attaquer aux dispositifs informatiques des organisations dans un objectif malveillant (espionnage industriel, déstabilisation, vol d'argent, etc.). En 2018, 80% des entreprises françaises ont subi au moins une cyberattaque [CESIN (2019)] ; à tel point que ce risque est désormais considéré, avec l'interruption d'activité comme le plus important [LE GOFF, E. (2018)] (devant les risques incendie, réglementaire, catastrophes naturelles et évolutions de marché).

Or, la transformation numérique et la mise en place d'une technologie centrée sur le traitement industrialisé de données massives (en particulier le *big data* et l'*Internet of things*) augmente nécessairement le besoin de cybersécurité dès lors que la masse de données augmente, que les traitements générant des effets financiers importants sont automatisés et que l'entreprise doit faire appel au *cloud computing* afin de disposer de la capacité de stockage et de calcul nécessaire à l'exercice de son activité. La généralisation de ces innovations augmente donc le risque de cyberattaques entrainant potentiellement fuites, corruption ou blocage des données.

C'est ainsi que l'assureur canadien Desjardins a été victime d'une faille de sécurité entraînant la fuite des données personnelles de trois millions d'assurés [PERRIN, G. (2019)]. De même, l'assureur américain Chubb, pourtant spécialiste du risque cyber, a également été victime d'un *ransomware* lié à un vol de données [WHITTAKER, Z. (2020)].

I.3.2. Risques liés aux données

La principale problématique liée aux données utilisées en *big data* et en IA concerne la notion de donnée à caractère personnel (cf. § I.3.2.1.) et plus particulièrement les données de santé (cf. § I.3.2.2.). En effet, cette notion entraine une contrainte réglementaire forte puisque tout traitement de donnée à caractère personnel est encadré par le RGPD [REGLEMENT UE (2016)] (cf. § I.3.2.3.). Dans une telle situation, l'assureur est donc soumis à un risque de ne pas être conforme à la réglementation – avec le risque de sanction induit – mais également d'un risque d'image dès lors que la protection de la vie privée des individus est potentiellement mise en cause. Ces obligations vont générer des contraintes fortes sur l'utilisation des techniques de *data science* s'appuyant sur des données à caractère personnel (cf. § I.3.2.4.). Le plus simple et le plus prudent consiste donc à sortir du périmètre d'application du RGPD en cherchant à anonymiser les données (cf. § I.3.2.5.).

I.3.2.1. Donnée à caractère personnel

Une donnée à caractère personnel correspond à « *toute information se rapportant à […] une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale* » [REGLEMENT UE (2016)].

Sont ainsi considérées comme des données à caractère personnel les données directement identifiantes (comme nom, prénom et adresse par exemple) mais également les données indirectement identifiantes, c'est-à-dire l'ensemble des données qui, recoupées entre elles, permettent d'identifier un individu (telles que numéro de téléphone, adresse IP, numéro de contrat, salaire, etc.). Par exemple, lorsque la population étudiée est de petite taille, le croisement du département d'habitation avec la tranche d'âge et avec le code acte de la Sécurité Sociale pourrait théoriquement permettre de réidentifier un individu : ces données doivent donc être considérées comme des données à caractère personnel.

I.3.2.2. Donnée de santé et donnée médicale

Parmi les données à caractère personnel, il existe des données identifiées comme particulièrement sensibles et dont les obligations de protection sont renforcées. C'est le cas des données de santé y compris des données médicales [REGLEMENT UE (2016)]. Ces données présentent une sensibilité accrue justifiant davantage de vigilance et de sécurisation lors de leur traitement et de leur utilisation. S'entend comme donnée à caractère personnel concernant la santé toute donnée « *liée à la santé physique ou mentale d'une personne, y compris la fourniture de services de soins de santé, qui révèle des informations sur son état de santé* », qu'il soit passé, présent ou futur [REGLEMENT UE (2016)].

Les données de santé peuvent provenir de plusieurs sources (de l'assuré, d'un médecin, d'un professionnel de santé, d'un hôpital, d'un dispositif médical, d'un processus de souscription d'assurance, etc.). Par exemple, pourraient être considérées comme des données de santé :

- le décompte de remboursement santé sauf si les codes actes permettent de deviner une activité produite à l'occasion d'un acte de soin (dans ce cas, le décompte de remboursement devient une donnée médicale) ;

- le devis hospitalisation (chambre particulière) sans indication du service hospitalier concerné ;

- les codes actes détaillés et codes de regroupement utilisés par la CNAM[25] pour autant qu'ils ne permettent pas d'identifier la pathologie (sinon cela devient une donnée médicale) ;

- les bons de livraison (en optique par exemple).

En fonction de sa finalité et de sa provenance [cnil.fr], la donnée de santé peut être qualifiée de donnée médicale. Son utilisation sera alors encore plus encadrée par la réglementation. Selon certaines doctrines [MALAKOFF HUMANIS (2018)], cela est par exemple le cas lorsque la donnée est « *recueillie ou produite à l'occasion des activités de prévention, de diagnostic de soins ou de suivi social et médico-social* » [Code de la santé publique].

Ainsi, les données médicales, en sus des obligations liées au RGPD (cf. § I.3.2.3.), sont soumises au secret professionnel lié à l'obligation de discrétion et de respect de la personne pour créer et assurer une relation de confiance entre le professionnel de santé et le patient qui se confie à lui. Le secret médical est consacré par le code de la santé publique, le code de déontologie médicale, relayé par le code de bonne conduite annexé à la convention AERAS[26] et sa violation est pénalement sanctionnée. L'accès aux données médicales est strictement limité aux professionnels de santé, au médecin-conseil de l'assureur voire aux personnes habilitées par celui-ci.

Dans le domaine de l'assurance de personnes, le secret médical porte donc sur toutes les informations médicales adressées par les assurés ou futurs assurés : de la collecte jusqu'aux traitements des données relatives à l'état de santé en vue de la souscription ou de l'exécution d'un contrat d'assurance.

De manière générale, toute information comportant une pathologie clairement désignée constitue une donnée médicale. Il ne peut être établi une hiérarchisation de la sensibilité au niveau des données médicales. A titre d'illustration :

[25] Caisse Nationale d'Assurance Maladie

[26] La convention AERAS (s'Assurer et Emprunter avec un Risque Aggravé de Santé) est une convention entre les pouvoirs publics, les associations de malades et les acteurs financiers, entrée en vigueur en 2007, ayant pour objectif de faciliter l'accès au crédit et à l'assurance emprunteur qui lui est liée pour les individus présentant un risque aggravé de santé.

- L'information relative à des maux de tête indiquée par un assuré est une donnée de santé. Dès lors que les maux de tête entrent dans le cadre d'un diagnostic opéré/validé par un médecin, par exemple névralgie, l'information devient une donnée médicale.

- Lorsque la donnée de santé est produite par un professionnel de santé, elle constitue une donnée médicale.

En synthèse, la structure réglementaire entourant les données de santé et médicales est définie sur plusieurs niveaux imbriqués les uns dans les autres comme figuré dans le schéma suivant :

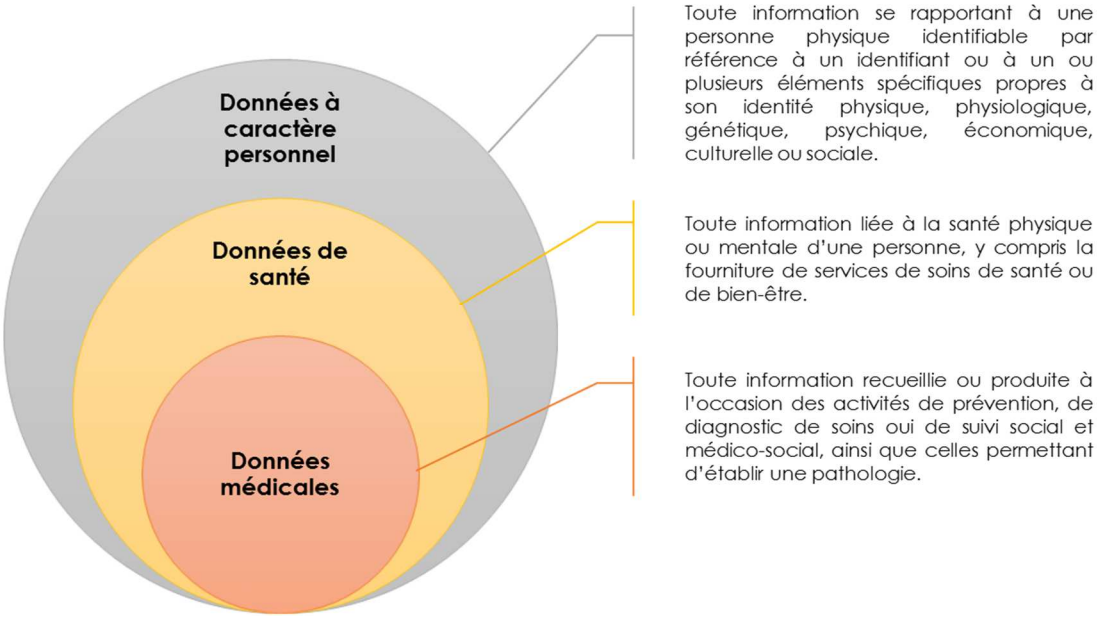


Figure 5 – Périmètres concernés par les différents niveaux de sensibilité des données

I.3.2.3. Obligations induites

Le traitement de données à caractère personnel est défini comme « *toute opération ou tout ensemble d'opérations effectuées ou non à l'aide de procédés automatisés et appliquées à des données ou des ensembles de données à caractère personnel* » [REGLEMENT UE (2016)]. Ainsi, dès que les données utilisées sont des données à caractère personnel, le traitement[27] effectué tombe sous le coup de la LIL [LOI (1978)] et du RGPD[28]. Cette réglementation spécifique vient également traduire des problématiques éthiques lié à la protection de la vie privée des individus. Ces deux questions juridiques et éthiques sont donc par essence indissociables.

[27] Il est important de noter qu'un traitement n'implique pas nécessairement une manipulation de fichier. Par exemple, sont considérés comme des traitements la collecte, l'enregistrement, la conservation, la consultation, la mise à disposition ou encore l'effacement des données.
[28] La LIL (loi française) et le RGPD (règlement européen d'application directe mais laissant des marges de manœuvre au niveau national) constituent tous deux des textes de références pour le droit français en matière de protection des données et s'appliquent de manière cumulative. Par convention, nous parlerons dans la suite de ce mémoire des obligations liées au RGPD, mais ceci comprend les obligations liées à la LIL, au RGPD et aux textes qui en découlent.

De manière synthétique, les obligations liées au RGPD imposent :

- l'identification d'une base légale justifiant le traitement (principe de licéité) ;
- l'identification précise de l'objectif du traitement des données (principe de finalité) ;
- la limitation du traitement des données aux informations strictement nécessaires au regard de la finalité (principes de nécessité et de proportionnalité) ;
- l'information des personnes et, dans certains cas (notamment pour les données de santé), le recueil de leur consentement préalable ;
- le respect des droits des personnes d'accéder, de modifier, de supprimer leurs données et, dans certain cas, celui de s'opposer au traitement ;
- la mise en place d'une durée de conservation des données précise et cohérente vis-à-vis de la nature de l'information et de la finalité du traitement ;
- la documentation des traitements (cartographie, registre et, dans certains cas, analyse d'impact sur la protection des données) ;
- la mise en place de mesures de sécurité et de confidentialité au niveau technique et organisationnel afin de protéger les données.

Il est important de noter que le non-respect de ces contraintes peut être financièrement lourd de conséquences : le règlement européen prévoit en effet que l'auteur d'un manquement encourt une amende pouvant aller jusqu'à 20 millions d'euros ou 4% du chiffre d'affaires annuel mondial [REGLEMENT UE (2016)].

C'est ainsi que la filiale en Bade-Wurtenberg de l'assureur allemand AOK a été récemment condamné par la LfDI[29] du land à une amende de 1,24 M€ pour avoir utilisé des données collectées lors de l'organisation de concours entre 2015 et 2019 à des fins publicitaires quand bien même des mesures (non suffisantes) de recueil de consentement avaient été prises et que l'assureur s'est montré coopératif avec les autorités de protection [VON PRESSESTELLE, G. (2020)].

I.3.2.4. Conséquences pour l'utilisation de la data science

De par l'importance que revêt la donnée dans leur fonctionnement, les technologies du *big data* et de l'IA posent un certain nombre de problématiques juridiques (elles-mêmes étant l'expression de questionnements éthiques) qu'il est important de prendre en compte dès la conception du projet (c'est ce qui s'appelle le *privacy by design*) tant ces aspects sont en interdépendance avec les contraintes techniques. En particulier, il est nécessaire en amont

[29] La *Landesbeauftragter für den Datenschutz* est l'autorité de protection des données à caractère personnel en Allemagne (l'équivalent de la CNIL)

de tout projet de *data science* d'identifier le fondement juridique du traitement envisagé ainsi que les garanties apportées pour répondre aux droits des personnes.

Dans le cadre des assurances de personnes et tout spécialement pour les risques santé et prévoyance, les données traitées sont fréquemment des données sensibles. En particulier, le régulateur peut considérer que les données de remboursement de soins produites par la CNAM et par les assurances sont des données de santé.

Dans le cas d'une base contenant des données personnelles de santé, les traitements de type *machine learning* doivent répondre aux mêmes contraintes [GEFFRAY, E. (2016)] que les autres types de traitements de données (cf. § I.3.2.3.), en particulier l'information des personnes, le droit d'accès aux données et le droit de s'y opposer dans certains cas [LOI (1978)]. En outre, le RGPD impose désormais la conduite d'une étude d'impact de l'effet du traitement sur la protection des données [REGLEMENT UE (2016)] des personnes concernées. Dans le cas des données médicales, l'encadrement du secret médical [Code de la santé publique] est aussi à prendre en compte dans le traitement. A cela s'ajoute enfin les garanties à apporter en termes de sécurité des données.

Néanmoins, le principe des traitements *big data* et IA pose malgré tout une problématique très spécifique : les techniques de *data science* ne sont-elles pas fondamentalement en contradiction avec le principe de minimisation des données, voire de finalité, porté par le RGPD ? En effet, la réglementation impose de ne traiter que les données strictement nécessaires à la finalité du traitement [REGLEMENT UE (2016)], cette finalité devant être identifiée dès le départ. Or, par principe même, comme vu au § I.2.3., les algorithmes utilisés en *machine learning* vont se nourrir des volumes conséquents de données pour éduquer leur intelligence artificielle, sans connaissance *a priori* de l'utilité ou de la finalité de telle ou telle donnée.

I.3.2.5. Anonymisation d'une base de données

En théorie, le fait d'anonymiser une base permet de s'affranchir des contraintes réglementaires liées au traitement des données personnelles, en particulier des données de santé et ainsi d'éviter de mettre en place un régime de consentement préalable ou de devoir limiter dans le temps la conservation des informations obtenues.

Juridiquement, le terme d'anonymisation correspond au cas où il n'est plus possible de rattacher le jeu de données à une personne physique. Ainsi, un simple chiffrement des données identifiantes d'une base ne suffit pas pour considérer que cette base est anonyme. D'après le G29[30], une base de données ne peut être considérée comme anonyme que s'il n'existe aucun risque de réidentification actuel ou futur [GROUPE DE TRAVAIL « ART. 29 » SUR LA PROTECTION DES DONNÉES (2014)]. Cela signifie entre autres

[30] Le groupe de travail « Article 29 » sur la protection des données était un comité indépendant composé principalement d'un regroupement des autorités de contrôles des Etats membres qui donnait son avis à l'Union européenne sur les problématiques de protection des données et de vie privée. Après l'entrée en vigueur du RGPD en mai 2018, il a été remplacé par le Comité européen de la protection des données.

que la clé de transcodage des informations directement identifiantes ne doit pas être conservée et qu'il doit être impossible de croiser les informations (soit directement via les données de la base, soit en la croisant à une autre base de données) pour réidentifier un individu.

Autant dire qu'une anonymisation au sens entendu par les autorités de protection européennes serait quasiment impossible à atteindre. Certaines affaires comme AOL, State of Massachusetts ou Netflix (cf. § annexe 2) ont bien démontré qu'il peut être possible de retrouver des individus au sein de bases pourtant considérées comme anonymisées [OHM, P. (2009)].

L'anonymisation parfaite étant un objectif difficile à atteindre en pratique, afin de pouvoir utiliser une base de données comportant des informations sur les personnes sans être contraint par le champ d'application du RGPD, il est nécessaire de justifier du faible, voire de l'absence de risque de réidentification en réalisant une analyse de ce risque. Celui-ci peut se diviser en trois composantes :

- **L'individualisation :** L'individualisation correspond à la possibilité d'isoler une partie ou la totalité des enregistrements identifiant un individu dans le jeu de données. C'est par exemple le cas d'un fichier d'enquête qui serait uniquement constitué des numéros de téléphone, du sexe et de l'âge des personnes interrogées.

- **La corrélation :** La corrélation consiste dans la capacité de relier entre elles, au moins, deux enregistrements se rapportant à la même personne ou à un groupe de personnes concernées. Si une attaque permet d'établir que deux enregistrements correspondent à un même groupe d'individus, mais ne permet pas d'isoler des individus au sein de ce groupe, la technique résiste alors à « l'individualisation » mais pas à la « corrélation ». C'est par exemple le cas d'une l'étude longitudinale pour laquelle il existe un fichier constitué d'informations associées à un *token* identique sur plusieurs années.

- **L'inférence :** L'inférence représente la possibilité de déduire, avec un degré de probabilité élevé, la valeur d'un attribut à partir des valeurs d'un ensemble d'autres attributs. C'est par exemple le cas d'un tableau de bord de l'absentéisme d'une entreprise pour lequel aucun seuil de restitution ne serait fixé et qui, par le biais du croisement de données, permettrait à l'entreprise de réidentifier ses salariés.

Afin de tendre vers une base de données la plus anonyme possible, il existe plusieurs techniques[31] qui servent à diminuer le risque de réidentification des individus (cf. annexe 2). La plupart de ces techniques vont cependant venir altérer les données et il convient donc de trouver un équilibre entre le risque de réidentification et l'intégrité informationnelle de la base compte-tenu des données et de la finalité de leur traitement. La définition d'anonymisation de la CNIL représentant un objectif difficilement

[31] Dans le cadre de ce mémoire, nous nous sommes surtout concentrés sur les mesures techniques permettant l'anonymisation d'une base de données. Néanmoins, il est important de rappeler que pour atteindre une réelle anonymisation des données, ces mesures techniques doivent impérativement être accompagnées par de mesures organisationnelles (procédures d'accès aux données, mise-à-jour des techniques et des données, etc.) et juridiques (encadrement contractuel des fournisseurs et/ou des clients des données, etc.).

atteignable en pratique, toute solution d'anonymisation passera nécessairement par la combinaison de différentes techniques selon le profil des données, le type de traitement et l'appétence au risque du responsable.

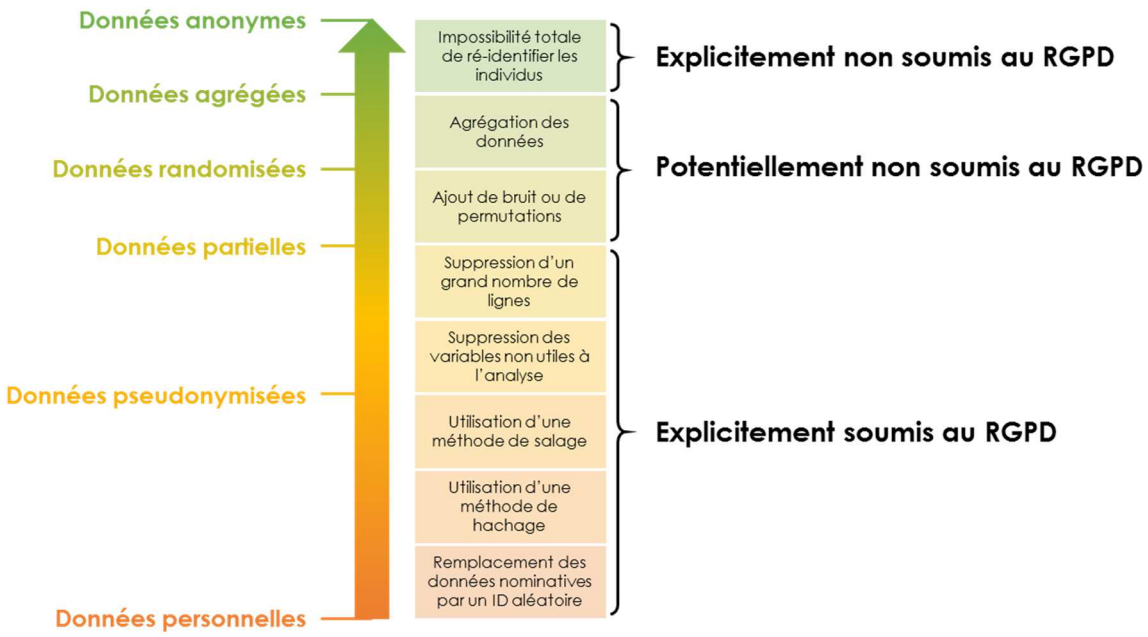


Figure 6 – Faisceau d'indices permettant de tendre vers une anonymisation des données

I.3.3. Risques liés aux algorithmes

Un des grands avantages de l'utilisation de l'IA est l'absence d'intervention humaine qui lui octroie une aura de neutralité et d'universalité. Quoi de plus objectif en effet qu'une suite d'opérations mathématiques ? En réalité, les algorithmes sont tout sauf justes. Tout d'abord parce qu'ils profitent de l'opacité que leur confèrent les mathématiques (cf. § I.3.3.1.). D'autre part parce qu'ils sont programmés par des êtres humains et alimentés par des données qui sont eux, volontairement ou involontairement, perclus de biais (cf. § I.3.3.2.). Ensuite, parce qu'ils tentent modéliser des réalités plus complexes que ce que peut décrire une seule loi mathématique (cf. § I.3.3.3.) ; enfin parce qu'ils cachent toujours une intention qui n'est, elle, jamais neutre (cf. § I.3.3.4.).

I.3.3.1. Mystère mathématique et opacité des algorithmes

La première problématique soulevée par l'utilisation des algorithmes mathématiques en général par le *machine learning* en particulier concerne l'opacité et l'absence d'intervention humaine. Dans son livre [O'NEIL, C. (2016)], Cathy O'Neil[32] explique que

[32] Docteur en mathématiques, data scientist et activiste américaine, Cathy O'Neil est diplômée de l'université d'Harvard. Après avoir effectué des recherches en mathématiques pendant une dizaine d'années, elle quitte le monde universitaire pour rejoindre celui de la finance en tant qu'analyste quantitative au sein d'un fonds d'investissement puis au sein d'une agence d'évaluation des risques avant de travailler comme data scientist dans le domaine de la publicité ciblée. Déçue par l'utilisation des mathématiques, elle devient militante politique et s'engage dans le mouvement Occupy Wallstreet. Son livre

les mathématiques sont utilisées de manière frauduleuse dans les algorithmes d'IA afin de profiter du fait que les gens ont spontanément confiance dans ceux-ci[33] tout en ayant peur des mathématiques. Ainsi, selon elle, les algorithmes mathématiques sont des armes d'intimidation des individus puisque lorsque ceux-ci se plaignent, la réponse est « *vous ne pouvez pas comprendre mais faites-nous confiance : ce sont des mathématiques.* ».

A titre d'illustration, il est intéressant d'étudier l'émergence de la police prédictive aux États-Unis. Ainsi, la police de Chicago a mis en place un algorithme permettant d'indiquer aux policiers dans quels lieux effectuer leur patrouille en priorité [PORTE, D. (2019)]. En réalité, le grand apport de cette IA, n'a pas été le résultat de l'algorithme en lui-même (un policier expert étant capable d'obtenir les mêmes résultats) mais le fait de rendre acceptable des décisions conduisant à un renforcement des patrouilles au sein des « *ghettos* » afro-américains (ce qui aurait été inacceptable si cette conclusion avait été tenue par un humain).

Un autre exemple illustrant la confiance aveugle dans des programmes mathématiques obscur est celui l'algorithme IMPACT[34] aux États-Unis. L'algorithme attribue une note aux enseignants en fonction de la performance de leurs élèves ; les enseignants ayant les notes les plus basses sont renvoyés. Le programme a fonctionné sans être remis en question pendant des années parce que les individus pensaient qu'ils n'avaient pas le droit de remettre en cause les résultats obtenus ni même demander des preuves que les notes étaient attribuées de manière juste et équitable. L'algorithme était en effet tenu secret[35] et personne, y compris le ministère de l'éducation n'avait accès au système. Par ailleurs, les notes étaient attribuées des années après la fin des cours, ce qui ne laissait pas aux enseignants la possibilité de s'améliorer. Mais récemment, six enseignants qui avaient été renvoyés en fonction de leur note ont intenté un procès qu'ils ont remporté : personne n'était capable d'expliquer la décision prise par l'algorithme. Le tribunal a par conséquent conclu que leur droit à un procès équitable n'avait pas été respecté [O'NEIL, C. (2020)].

Dans le cas spécifique du *machine learning* et plus spécifiquement du *deep learning*, l'effet boîte noire est encore amplifié par le fait que le modèle n'est pas construit par un être humain mais directement extrait des données. Cette absence de logique humaine va rendre encore plus difficile la compréhension et l'analyse des résultats obtenus. Or, les actuaires rencontrent déjà des difficultés à faire comprendre le fonctionnement de leurs modèles au niveau de la gouvernance de l'entreprise alors même qu'il s'agit d'une exigence réglementaire [REGLEMENT UE (2016)] et déontologique [INSTITUT DES ACTUAIRES (2014)] forte. Les data scientistes ne peuvent que constater les résultats en sortie de leurs algorithmes. Dès lors, comment justifier ces résultats auprès des commissaires aux comptes et des régulateurs ? A ce propos, l'ACPR recommande de privilégier « *des*

Weapons of Math Destruction a remporté en 2019 le prix Euler qui récompense le meilleur ouvrage de vulgarisation mathématique.

[33] Malgré l'augmentation substantielle du nombre de scandales autour de l'IA et du *big data*, notamment avec l'affaire Facebook-Cambridge Analytica, les opinions et les pratiques numériques ne semblent pas évoluer outre mesure.

[34] Algorithme mis en place en 2009 dans le district de Washington avec pour objectif d'optimiser les compétences de l'équipe des enseignants.

[35] Les seules explications fournies étaient que les notes étaient principalement construites selon les résultats des élèves aux examens par comparaison à ce qui était attendu *a priori*.

algorithmes simples et robustes afin que ces derniers puissent être compris » [ACPR (2018)] par le top management en charge de les valider.

Le fait de rendre les algorithmes de *machine learning* davantage transparents est un véritable enjeu et représente par conséquent un vaste champ de recherche pour la *data science*. Des travaux sont par exemple menés [ANTONIO, K. et al. (2020)] pour créer des modèles interprétables de substitution aux modèles boîtes noires (certains sont déjà implémentés dans Python [shap.readthedocs.io]).

I.3.3.2. Biais induits

Si l'absence d'intervention humaine dans les algorithmes de type *machine learning* permet de faire abstraction de toute idée préconçue, rien ne permet d'assurer *a priori* que les résultats obtenus soient véritablement justes ni même acceptables. En effet, des erreurs ou des biais peuvent apparaître aux différentes étapes : lors de la collecte des données, dans la construction du code informatique ou dans le déroulement de l'algorithme. Les technologies de *big data* et d'IA se nourrissant de masses de données, il n'est que naturel que les résultats issus de ces algorithmes reproduisent et même exacerbent les biais et les erreurs portés en sous-jacent par ces données, alors qu'un être humain a la capacité à corriger ces erreurs et à prendre en compte ces biais.

La CNIL est très au fait de ce risque puisqu'elle écrit dans son rapport : « *La propension des algorithmes et de l'intelligence artificielle à générer des biais pouvant conduire à leur tour à créer ou à renforcer des discriminations, s'est imposée comme un sujet d'inquiétude et de questionnement. Le constat mérite d'autant plus d'être souligné que ces systèmes techniques peuvent également parfois nourrir une croyance en leur objectivité. Une objectivité d'autant plus précieuse qu'elle ferait souvent défaut aux humains.* » [CNIL (2017)].

A titre d'illustration, en 2016, Microsoft a déployé sur les réseaux sociaux et les applications de messagerie Tay [REESE, H. (2016)], une IA auto-apprenante capable de discuter avec les utilisateurs humains tout en enrichissant ses capacités de paroles. Cependant, au bout de quelques temps, alimenté par les propos tenus par les internautes, le *chatbot* serait devenu vulgaire et raciste si bien que Microsoft a dû mettre fin au projet.

De manière plus inquiétante, cela prouve que les algorithmes pourraient dans certaines circonstances venir creuser les inégalités présentes dans la société [O'NEIL, C. (2020)]. En effet, beaucoup d'algorithmes fonctionnent sur le principe de classification des individus en fonction de leur classe sociale et plus particulièrement en fonction de leur pouvoir d'achat. Est-ce qu'ils sont blancs ? Est-ce qu'ils sont bien éduqués ? Est-ce qu'ils vivent au bon endroit ? En d'autres termes, l'IA est souvent utilisée pour placer les individus sur une échelle de la chance et de la réussite puis va sélectionner ceux dépassant un certain seuil pour leur proposer de nouvelles opportunités (qui va avoir un bon taux de crédit, qui va avoir une bonne assurance, etc.). Inversement ceux qui bénéficient le moins d'opportunités ont tendance à être maintenus dans cette condition. Cet effet potentiel est

d'autant plus insidieux que chaque entreprise qui utilise ce type d'IA imagine n'utiliser qu'un simple algorithme prédictif pour maximiser son profit sans nécessairement penser à mal. Cependant, comme le *big data* se généralise à très grande échelle et à tous les secteurs, l'effet potentiellement négatif obtenu au niveau de la société est massif.

Un autre élément qui prouve que l'IA présente un risque d'augmentation des inégalités est le fait que les algorithmes sont souvent nourris par des données historiques : ils analysent le passé afin de prédire ce qui peut se passer dans le futur. C'est en soit un fondement qui semble raisonnable et qui est même à la base de l'éducation humaine. Cependant, le passé des hommes fourmille d'exemples de faits qu'il n'est pas souhaitable de reproduire car considérés de nos jours comme injustes ou inégalitaires. Or une IA éduquée avec ces données historiques ne pourrait qu'amplifier ces phénomènes.

Par exemple, en 2014, Amazon a réalisé un programme d'IA pour le recrutement de ses ingénieurs. En utilisant leurs propres données (profils des gens ayant postulés et de ceux qui ont obtenus un poste), ils ont essayé de caractériser des critères de carrière réussie au sein de l'entreprise (durée dans le poste, nombre de promotions, etc.) afin d'alimenter leur algorithme. Cela semblait somme toute des données très raisonnables. Or, très vite, il s'est avéré que l'IA privilégiait fortement les candidatures des hommes par rapport à celle des femmes. En réalité, l'IA ne faisait qu'automatiser le processus de recrutement tel qu'il a été conduit par le passé, en reproduisant les inégalités homme/femme. Qui a décroché le poste d'ingénieur au départ ? Qui a obtenu une promotion ? Qui a le meilleur salaire ? Toutes ces données comportaient des biais au départ. Il est donc logique qu'en utilisant une IA qui se nourrisse de ces données, le résultat obtenu soit identiquement biaisé, même si ce n'était clairement pas l'intention d'Amazon au départ.

Outre les biais contenus dans les données, les biais peuvent aussi provenir dans la conception même de l'algorithme car celui-ci a été pensé par un être humain. A titre d'illustration, nous pouvons citer un autre exemple de police prédictive aux Etats-Unis : il s'agit de l'attribution par une IA à un condamné d'un score de risque de récidive[36]. Au-delà de la question philosophique soulevée par ce système[37], il y a potentiellement un problème d'équité dans la conception même de ce genre algorithme car les critères qui vont servir à déterminer ce score peuvent être orientés selon des fins politiques. Par exemple, en construisant ces critères spécifiquement autour de crimes non violents caractéristiques de gens en situation précaire (consommer de la drogue, resquiller dans le métro, uriner sur la voie publique, présenter des pathologies mentales, etc.), le résultat du programme sera nécessairement biaisé envers les personnes en situation de pauvreté, non pas à cause des données mais parce qu'il a été conçu comme tel [O'NEIL, C. (2016)].

Enfin, comme évoqué au § I.3.2.3., lorsque les bases traitées comportent des données à caractère personnel, le RGPD peut imposer dans certains cas (notamment pour le traitement de données de santé) le recueil préalable du consentement libre et éclairé des personnes vis-à-vis du traitement de leurs données. Or, nous pouvons nous interroger sur les biais potentiellement induits par ce recueil de consentement dans les données traitées.

[36] Risque d'être à nouveau arrêté et condamné dans les deux années suivant sa libération.
[37] Les détenus obtenant un score élevé vont recevoir des peines plus longues ; autrement dit, ils vont être condamnés pour un crime qu'ils n'ont pas encore commis.

En effet, il semble raisonnable de penser que les personnes ayant donné leur consentement n'aient pas le même profil que le reste de la population ; ou bien qu'en accordant son consentement, l'individu décide sciemment de changer son comportement ou de « mentir » sur les données qu'il fournit

Ce biais dans le recueil des données peut également se traduire par des problématiques de données manquantes qui vont venir altérer la vision de la réalité si les informations absentes ne sont pas purement le fruit du hasard (cf. § II.1.2.). Pour illustrer ce point, reprenons une dernière fois l'exemple de la police prédictive aux États-Unis. En partant de l'hypothèse que les hommes réalisant des actions illicites (tel que l'usage de drogue) présentent *a priori* moins de chance de se faire arrêter par la police lorsqu'ils proviennent d'un environnement aisé et non issus d'une minorité ethnique que s'ils sont afro-américains et issus d'un milieu défavorisé. Dès lors, si les algorithmes utilisés pour faire de la police prédictive ne cherchent pas en amont à traiter ses valeurs manquantes d'une manière ou d'une autre, l'IA ne pourra qu'amplifier un système défaillant.

Ainsi, est-il raisonnable d'aliéner l'expertise de l'actuaire au profit d'une boîte noire sous prétexte que cette dernière est auréolée de l'aura magique de l'objectivité mathématique et du marketing à la mode ? Au vu de ce qui précède, il apparaît que les données servant à éduquer les IA doivent être retraitées afin d'éliminer au maximum les biais qu'elles peuvent contenir et que les résultats obtenus par les algorithmes de type *machine learning* doivent être analysés finement avant de pouvoir être utilisés afin d'éviter par exemple que l'assureur ne discrimine sans le vouloir une partie de la population. Dans son rapport [ACPR (2018)], l'ACPR indique clairement sa volonté de contrôler le cadre de validation des résultats obtenus par le bais de l'IA.

I.3.3.3. Modélisation d'une réalité complexe

L'objectif principal d'un modèle est de représenter une partie du monde réel. De la mécanique newtonienne à la mécanique quantique, les mathématiques sont un formidable outil pour modéliser les lois universelles de la nature. La théorie mathématique, la puissance de calcul des ordinateurs et l'émergence du *big data* ont permis de chercher à modéliser des phénomènes de plus en plus complexes, en particulier ceux faisant intervenir plusieurs êtres humains.

Cependant, dès lors que ce sont des comportements humains qui sont modélisés, en particulier dans le domaine de la finance, la plupart des modèles sont faux. Soit parce qu'ils prennent des hypothèses simplificatrices (distribution normale, concurrence pure et parfaite, absence d'opportunité d'arbitrage, etc.) qui sont éloignées de la réalité. Soit parce que le modèle est utilisé en dehors de son domaine de validité. Soit encore parce qu'un modèle plus juste serait incompréhensible pour la gouvernance de l'entreprise ou bien mettrait en lumière un profil de risque nettement plus élevé que ce que l'entreprise est prête à reconnaître.

Ce dernier point serait d'ailleurs l'une des causes de la crise financière des *Subprimes* de 2008 : certains acteurs financiers ont préféré conserver des modèles faux afin d'afficher un profil de risque beaucoup moins élevé et optimiser que ce qu'ils étaient prêts à accepter en réalité dès lors qu'ils se croyaient protégés par le principe du « *too big to fail* ». Cathy O'Neil dit à ce propos : « *Les mathématiques s'étaient associées à la technologie pour décupler le chaos et le malheur conférant une ampleur et une efficacité redoutable à des systèmes que je savais désormais défectueux.* » [O'NEIL, C. (2016)].

Malgré tout et comme vu précédemment, la confiance des gens dans les modèles mathématiques et l'IA demeure inchangé et de nombreux programmes continuent de modéliser de manière erronée des réalités trop complexes[38]. Si nous reprenons l'exemple de l'algorithme IMPACT de notation des enseignants en mettant de côté la problématique liée à l'effet boîte noire, l'analyse du programme basée sur une comparaison entre la réussite des élèves et les résultats attendus compte-tenu de leur niveau à l'entrée peut sembler de prime abord pertinente. Néanmoins, dans la réalité, cela a incité les enseignants à éduquer leurs élèves uniquement dans l'optique d'obtenir les meilleures performances aux examens (ce qui n'est pas la méthode d'éducation la plus épanouissante) et – plus grave – cela les a incités à tricher. De plus, cela générait un effet pervers au sein de l'algorithme : puisque les élèves d'un enseignant ayant cédé aux sirènes de la fraude à l'examen avaient obtenu de très bons résultats, leur attendu pour l'examen de la classe supérieure étaient d'autant plus important. Par voie de conséquence, l'enseignant de l'année suivante se trouvait pénalisé *de facto* dans sa notation par l'algorithme avant même d'avoir commencé à dispenser ses cours. Au final, selon Cathy O'Neil, l'algorithme IMPACT se serait révélé aussi efficace qu'un algorithme attribuant des notes purement aléatoires [O'NEIL, C. (2016)] et n'aurait pas été en mesure de saisir toutes les subtilités qui distinguent un « bon » enseignant d'un « mauvais ».

Face à la nécessité de modéliser un monde complexe, le *big data* a apporté une réponse : l'augmentation massive des données qui vont venir alimenter les algorithmes. En utilisant ainsi toujours davantage de données, y compris celles qui ne présentent pas d'intérêt *a priori*, l'information disponible pour l'algorithme est maximisée et par conséquent le résultat obtenu est nécessairement meilleur. Malheureusement, cette affirmation est fausse [O'NEIL, C. (2020)] : le fait de ne pas sélectionner les données pertinentes va venir diluer l'information importante et favoriser l'émergence de résultats déviants portés par les données. Dans l'exemple de l'algorithme de recrutement développé par Amazon, la prise en compte de toutes les données qui tentaient de capter de l'information sur les tendances expliquant le succès a complètement occulté la prise en compte des compétences et des qualifications des individus.

[38] De plus, ces modèles imparfaits sont parfois utilisés pour dicter le pilotage de l'entreprise, comme cela pourrait potentiellement être le cas en assurance sous la nouvelle norme comptable IFRS 17.

I.3.3.4. Prise de décision automatisée

Dans le monde économique, lorsqu'il y a prise de décision, cela signifie qu'il existe une différence de pouvoir entre deux acteurs : un individu recherche un service auprès d'une entreprise ou d'une institution (un emploi, une place à l'université, un emprunt, une police d'assurance) et sera contraint de répondre à des questions. Les données ainsi collectées peuvent aboutir à une décision qui devient de plus en plus fréquemment automatisée.

La prise de décision automatisée par une IA soulève une problématique fondamentale : celui du critère de décision. Si un algorithme peut être considéré comme neutre, il cache en effet toujours derrière une intention subjective. Ce critère de réussite sera bien évidemment à la main de celui qui détient le pouvoir de décision (l'entreprise ou l'institution acceptant ou non de vendre un service) et l'autre partie prenante (le plus souvent l'individu) ne peut qu'espérer que ce critère sera compatible avec son propre besoin. Or, selon le critère de décision (qui reflète l'intention de l'auteur et qui est par conséquent nécessaire subjectif), un même algorithme pourra donner des résultats diamétralement opposés. Dès lors, se pose la question de la responsabilité de ceux qui choisissent ce critère de réussite : vont-ils utiliser leur pouvoir pour servir le surplus global du système ou bien uniquement leur propre profit aux détriments des autres acteurs ?

Un exemple très parlant est celui de l'algorithme de construction du fil d'actualités sur Facebook. L'objectif de l'utilisateur est évidemment de s'informer de manière simple et efficace sur l'actualité du monde et sur celle de leurs connaissances tandis que l'objectif de Facebook est de maintenir l'utilisateur le plus longtemps possible sur son site. L'algorithme va ainsi privilégier les actualités qui ont le plus de chance de générer des clics et non les actualités les plus pertinentes ou les plus éclectiques. Ainsi, au lieu de répondre au besoin d'information de l'utilisateur, l'algorithme de Facebook va avoir tendance à l'enfermer dans une bulle de désinformation qui ne montre plus l'actualité telle qu'elle est mais telle qu'il aimerait la voir, avec un risque avéré de dérives qui peuvent entraîner dans le pire des cas vers des réseaux conspirationnistes [YATES, J. (2017)] voire terroristes [FRANCOIS, M. (2018)].

Avec des algorithmes opaques, des résultats qui peuvent comporter des biais et des modèles potentiellement faux, dans quelle mesure les assureurs peuvent-ils utiliser les technologies du *big data* et de l'IA lorsque celles-ci vont produire des effets contractuels pour les assurés, en particulier en faisant évoluer leur niveau de primes ? Le cas échéant, est-ce que le critère de décision retenu est-il le reflet de la recherche d'une amélioration du système dans son ensemble ? L'assureur doit ainsi bien vérifier le fonctionnement de son algorithme (en particulier sur les cas minoritaires) et être à tout moment en mesure de justifier et d'expliquer à l'assuré la décision prise. Par ailleurs, ce dernier doit être en mesure de pouvoir facilement la contester [thepublicvoice.org]. A contrario, si l'assureur ayant refusé une assurance à un individu sur la base d'une IA n'est pas capable de démontrer de manière intelligible les raisons de ce refus alors cette décision sera jugée non recevable [GROUPE DE TRAVAIL « ART. 29 » SUR LA PROTECTION DES DONNÉES (2018)].

Dans tous les cas, il semble nécessaire de bien séparer le champ d'intervention de l'IA et celui de l'actuaire : le premier doit être cantonné à un rôle de suggestion et le second doit pouvoir disposer de la compétence, de l'expertise et de la latitude nécessaire afin de confirmer ou d'invalider le résultat produit par l'algorithme. En particulier, il est primordial que la capacité de remise en cause de la machine par l'expert soit réelle et soutenue par la gouvernance de l'entreprise. En outre, s'il est avéré que l'algorithme s'est trompé dans sa prise de décision, il est important de pouvoir disposer des moyens d'analyser les motifs ayant conduit à ce résultat afin d'en tirer les conséquences.

I.3.4. Risques liés aux utilisations

I.3.4.1. Conséquences induites par les décisions de l'assureur

Quelles sont les conséquences que peuvent avoir les décisions de l'assureur suite à une prédiction par les innovations permises par le *big data* et l'IA ? Dès 1956, Philip K. Dick[39] s'interrogeait sur la validité d'une prédiction dès lors que la connaissance de celle-ci entraîne des actions spécifiques : « *Il ne peut y avoir de réelle connaissance du futur. Dès qu'une information précognitive est livrée, elle s'annule d'elle-même. L'affirmation selon laquelle un homme commettra un crime dans l'avenir est un paradoxe. Le simple fait de posséder cette donnée la fausse.* » [DICK, P. K. (1956)]. Cette réflexion, certes un peu romancée, pose la question de l'impact non négligeable de l'observation et de la prédiction sur l'évolution de la réalité modélisée, ce qui est clairement le cas dans le domaine financier (une prédiction de chute des marchés va entraîner un affolement des acteurs et *in fine* devenir auto-réalisatrice).

Imaginons par exemple qu'un programme de type *machine learning* servant à évaluer le risque assuré détecte une catégorie des assurés comme plus risquée que le reste du portefeuille. L'assureur s'empressera d'augmenter les primes de ces individus, les poussant vraisemblablement à la résiliation. Dès lors, il influencera les comportements de ces assurés et les résultats du modèle utilisés ne sont plus fiables puisque l'état du portefeuille assuré a substantiellement changé par rapport à ce qu'il aurait été si la prédiction n'avait pas été énoncée. Certes, la machine apprenante va intégrer ce nouvel état des choses progressivement pour mettre à jour ses prédictions mais cela démontre qu'à un instant t, rien ne prouve que l'algorithme ne soit pas sous-entraîné par rapport aux données et que ses prédictions ne soient pas erronées (un raisonnement similaire peut être réalisé avec des cas de surentraînement). Or comme déjà évoqué, l'effet boîte noire inhérent aux algorithmes de *machine learning* empêche l'actuaire de détecter facilement ce biais.

[39] Philip K. Dick (1928-1982) est un auteur américain de romans, nouvelles et essais de science-fiction. Une grande partie de son œuvre pose la question de la signification de la réalité.

I.3.4.2. Démutualisation du risque assurantiel

I.3.4.2.1. Mutualisation VS segmentation

Il est possible d'imaginer un second risque lié à l'utilisation de la *data science* encore plus dangereux pour l'assureur : la démutualisation du risque. Depuis son origine, l'assurance se fonde sur le principe de mutualisation[40] du risque : le métier de l'assureur consiste généralement à réunir un nombre important de personnes présentant des risques similaires afin de mettre en place un système d'inter-financement permettant de couvrir l'aléa (sauf par exemple dans le cadre de l'activité d'épargne en unités de compte ou pour l'aérospatial).

Cette mutualisation possède à la fois un rôle mathématique (elle permet de se rapprocher de la loi des grands nombres et ainsi rendre le risque modélisable par des outils statistiques), économique (elle permet de rendre assurable des risques qui ne le serait pas autrement) et social (elle crée un mécanisme de solidarité entre les assurés).

En parallèle, afin de mieux connaître les risques qu'il assure, l'assureur est naturellement incité à réaliser une segmentation. Il s'agit donc de récupérer un maximum d'information afin de regrouper ensemble des profils de risques similaires. Cela lui permet ainsi non seulement d'optimiser les montants à provisionner dans ses comptes mais également de proposer des tarifs plus compétitifs ou des produits mieux adaptés à ses clients.

Ces deux mouvements s'opposent : l'un cherche à regrouper les risques, l'autre à les dissocier. Cependant, tant que la maille de segmentation est suffisamment grande et que l'effet de l'aléa reste prépondérant, il reste possible de les réconcilier.

I.3.4.2.2. Vers une individualisation du risque

Avec les techniques de *big data* et d'IA, une tendance à l'individualisation du risque peut être observé puisque ce dernier est calculé à une maille de plus en plus fine avec tout un jeu de nouvelles données disponibles. Ces innovations peuvent donc permettre aux assureurs de mieux connaître leurs clients afin de les segmenter de manière plus fine, rétablissant ainsi l'asymétrie d'information entre l'assureur et l'assuré [SCOR (2018)]. Ces avancées dans les possibilités de segmentation par les assureurs s'illustrent particulièrement bien dans le domaine de l'assurance automobile où nous sommes passés d'une segmentation selon la catégorie de véhicule[41], ensuite l'introduction du

[40] Outre la mutualisation opérée pour chaque sous-ensemble de risque décrite ici, il existe également un effet de mutualisation lié à la diversification des risques couverts par l'assureur. En effet, l'aléa sur le résultat d'un ensemble de contrats est indubitablement plus aisé à supporter lorsqu'il peut être compensé par les aléas sur le résultat des autres types de risques. Ainsi un assureur généraliste dispose généralement d'un levier de pilotage supplémentaire par rapport à un assureur spécialisé.

[41] Outre la valeur du véhicule, le montant de la prime d'assurance dépend de la catégorie du véhicule, de sa marque et de sa couleur.

bonus/malus[42] puis le « *pay as you drive* »[43] et enfin l'émergence aujourd'hui du « *pay how you drive* »[44].

Cet exemple illustre bien que la disponibilité de données plus variées et plus proches de l'assuré a permis l'émergence d'une assurance comportementale : la tarification ne s'appuie plus seulement sur la nature du risque, sur les caractéristiques de l'assuré ou sur l'historique du contrat, mais intègre désormais un système de récompense des « bons » comportements[45] sous forme de réductions en amont de la réalisation ou non du risque couvert.

C'est ainsi que l'assureur américain John Hancock fait évoluer ses primes d'assurance décès de 15% selon le mode de vie (activités physiques, habitudes alimentaires, temps de sommeil, etc.) des assurés qui sont « *trackés* » par des objets connectés (comme la montre Appel Watch ou le bracelet Fitbit) [BARLYN, S. (2018)].

I.3.4.2.3. Qui entraine la fin de la mutualisation

Cependant, avec cette mesure du risque beaucoup plus fine, les innovations apportées par le *big data* et par l'IA rompent *de facto* le principe de la mutualisation. Ainsi, plutôt que répartir collectivement la charge du risque, grâce à des prévisions rendues possible à la maille individuelle, l'assureur sera incité à segmenter ses tarifs au plus près de son évaluation du risque, à savoir potentiellement au niveau de chaque assuré.

La conséquence (en admettant que les modèles soient justes) est que chaque individu paiera la valeur de son propre risque. Ce faisant, l'assureur proposera une baisse du niveau global de ses primes puisque les bons assurés auront une prime plus faible (ce qui attirera d'autres bons risques) et les mauvais seront résiliés (ou incités à partir vers un autre assureur n'ayant pas encore réalisé sa révolution *big data*). Cependant, une fois que tous les acteurs du marché se seront adaptés (ou auront disparu), il n'y aura théoriquement plus d'utilité à l'assurance : les bons risques auront tout intérêt à s'auto-assurer (pour ne pas avoir à payer les frais de l'assureur) et les mauvais risques ne pourront tout simplement plus être assurés [SCOR (2018)]. Afin d'illustrer ce phénomène liée à la démutualisation du risque, nous avons développé un modèle simplifié en Python. Celui-ci est présenté en annexe 3.

---

[42] Ajustement du montant de la prime d'assurance en fonction du nombre de sinistres impliquant la responsabilité de l'assuré, conduisant à faire payer davantage les « mauvais » conducteurs *a posteriori*.

[43] Ajustement du montant de la prime d'assurance en fonction de l'utilisation réelle du véhicule, la plupart du temps mesuré en temps réel à l'aide d'un logiciel de géolocalisation, conduisant à faire payer davantage les conducteurs fréquents.

[44] Ajustement du montant de la prime d'assurance en fonction de la manière de conduire de l'assuré (façon de freiner ou d'accélérer, anticipation des virages, etc.) grâce à un boîtier électronique installé sur le véhicule, conduisant à faire payer davantage les « mauvais » conducteurs *a priori*.

[45] Il serait intéressant de s'interroger sur qui est en droit de définir la frontière entre un « bon » comportement et un « mauvais » (dès lors que ce comportement n'est pas puni par la loi). Dans le cas du marché de l'assurance, il est probable que ces décisions soient basées soit sur des études scientifiques soit sur des algorithmes de *machine learning*. Or il existe un risque non négligeable que les premières soient fausses (par exemple le couchage des nourrissons sur le ventre dans les années 80) ou partisanes (par exemple, le lait maternel par rapport au lait artificiel) et que les seconds peuvent comporter des biais socialement inacceptables (cf. § I.3.3.).

Ce modèle de segmentation à outrance qui exclut *de facto* la mutualisation du risque porterait ainsi en germe la mort programmée de la majorité du marché de l'assurance dès lors que les uniques raisons qui pousseraient les gens à s'assurer seraient l'obligation légale (pour la responsabilité civile par exemple) et l'aversion au risque. Même en prenant l'hypothèse que cette disparition serait bénéfique au surplus[46] global de la société, ce modèle remet en cause le rôle de la protection sociale tel qu'il existe en France depuis 1945.

Nous voyons donc que par essence, l'assurance ne peut fonctionner que sur une population et non sur des risques individuels. Évidemment, le rôle de l'assureur consiste à identifier, sélectionner et diviser les risques. Afin d'avoir une mesure fine du risque, il va effectivement chercher à procéder à une individualisation du risque. Mais celle-ci repose sur une individualité moyenne permise par le regroupement de risques similaires et qui n'a de sens que relativement à l'ensemble de la population assurée. Le marché de l'assurance doit donc réussir à trouver un équilibre subtil entre le principe original de mutualisation des risques et la segmentation fine des assurés permises par les nouvelles technologies.

Ce point de vue d'ailleurs est partagé par Charpentier et al. dans leur article : « *[La] spirale de la segmentation » aboutit à une situation qui n'est pas claire. Alors que certains évoquent une éventuelle responsabilité des actuaires dans la limitation, pour des raisons sociales, de la segmentation (les « mauvais » risques pouvant être incités à ne plus s'assurer), on notera que, d'un point de vue économique, la conclusion n'est pas aussi évidente. La recherche de niches tarifaires « rentables » peut conduire à la situation où quelques rares sociétés sont (potentiellement) à l'équilibre, avec des parts de marché très faibles et une variabilité du résultat importante, et d'autres perdent de l'argent (en moyenne) avec des parts de marché beaucoup plus importantes. Ce jeu pourrait être dangereux à moyen terme, et beaucoup risquent de découvrir que l'art de la tarification est plus subtil qu'il n'y paraît.* » [CHARPENTIER, A. et al. (2015)].

I.4. PRESENTATION DU CAS D'USAGE MIS EN ŒUVRE

I.4.1. Problématique traitée

Dans le cadre de ce mémoire, nous avons souhaité proposer une utilisation des innovations permises par le *big data* et l'IA qui puisse apporter de la valeur à l'assureur dans son activité d'actuariat et ce de manière éthique. Nous entendons par le terme « éthique » bien évidemment le respect de la réglementation mais également le respect de lois non écrites permettant de garantir la bienveillance et l'intérêt de l'assuré dans le traitement et l'utilisation de ses données. En particulier, nous avons accordé une importance particulière au risque de démutualisation.

[46] Dans la théorie microéconomique, le concept de surplus (ou de bien-être) permet d'évaluer quantitativement l'intérêt d'un modèle du point de vue du producteur ou du consommateur d'un produit ou d'un service. Le surplus global est obtenu en additionnant le surplus de tous les acteurs du marché et permet ainsi de déterminer le modèle théorique optimal pour la société.

L'idée principale du cas proposé consiste à utiliser les techniques apportées par la *data science* sur une base de l'*open data* de santé appariée à la base de données des assurés d'un assureur en santé et en prévoyance dans le but de prédire les décès et ainsi d'améliorer le provisionnement en prévoyance. L'hypothèse sous-jacente est que le profil de consommation de soins d'un assuré apporte de l'information sur son risque de décès. Il s'agit d'une hypothèse instinctive et raisonnable pour ce qui concerne les décès qui ne sont pas dus à une cause accidentelle.

Par ailleurs, le risque décès est un risque rare et lourd, ce qui rend délicat son estimation avec les techniques actuarielles classiques[47], en particulier lorsque la population assurée n'est pas très nombreuse (par exemple à l'échelle d'une entreprise) ou lorsque la série temporelle est courte (par exemple pour un nouveau produit). En conséquence, selon le jeu de l'aléa, les assureurs sont amenés à constater des boni ou des mali importants par rapport aux montants qu'ils ont provisionnés.

I.4.2. Etapes de la solution proposée

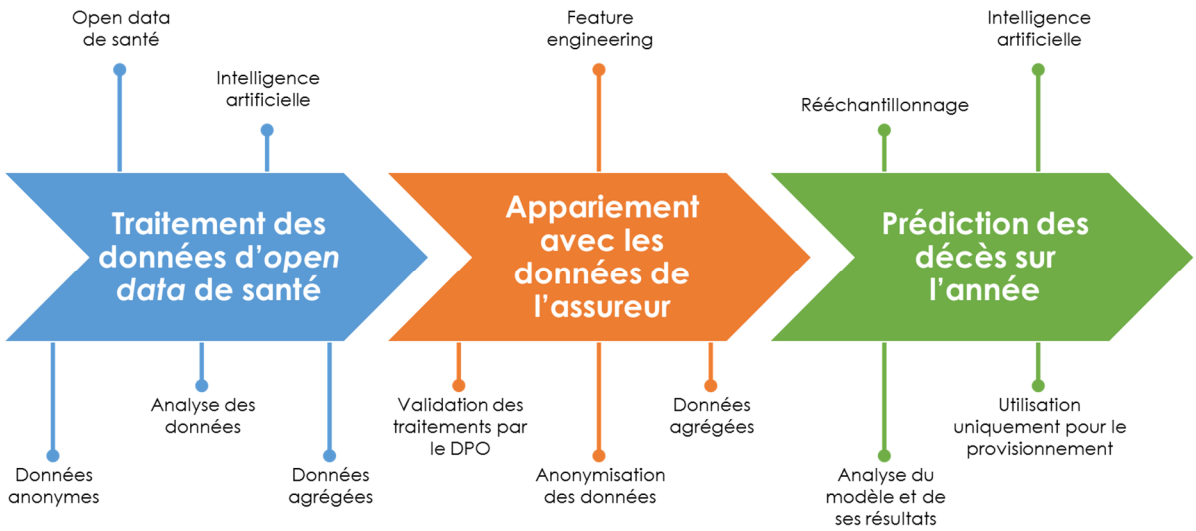


Figure 7 – Etapes du cas d'usage mis-en-œuvre

La première étape de notre solution consiste à récupérer et à traiter les données de l'*open data* de santé (cf. § II.1). Ce traitement aura notamment pour objectif de vérifier l'intégrité des données, en particulier ce qui concerne le traitement des données manquantes et des valeurs aberrantes. Cette étape est indispensable afin de s'assurer que la base de données utilisée est fiable et complète pour la suite des traitements.

[47] Traditionnellement, les actuaires estiment le risque décès soit via l'application d'un triangle de liquidation basé sur les années passées (méthodes de type *chain ladder*), soit via une loi de décès qui peut être réglementaire ou construite à partir des données de l'assureur.

Les données ainsi traitées devront ensuite être rapprochées des données de l'assureur (cf. § II.2.). Pour cela, il faudra trouver une méthode et des critères de rapprochement, et transformer les données de manière à les appliquer. Il s'agit d'un traitement qui va avoir un fort impact sur le résultat du modèle puisqu'il s'agit à la fois de conserver le plus grand volume de données possible et de les mettre sous une forme qui permettra exploiter au mieux l'information disponible lors de l'étape de prédiction.

Enfin, nous utiliserons la totalité des données consolidées pour mettre en place une méthode de provisionnement du risque décès par un algorithme prédictif (cf. § II.3.). Plusieurs algorithmes de *machine learning* seront testés et il faudra pouvoir les comparer entre eux afin de déterminer le plus pertinent. Afin de valider la solution, il faudra également éprouver les résultats de cet algorithme par rapport à une méthode similaire appliquée sans données d'*open data* (uniquement les données de l'assureur), ainsi que par rapport aux méthodes traditionnelles.

I.4.3. Base de données d'open data de santé utilisée

Parmi toutes les bases mises à disposition par le *Health data hub*, il aurait été évidemment particulièrement intéressant d'obtenir un accès afin de travailler sur les bases de données du SNDS. Le cas d'un mémoire d'actuariat n'est pas clairement répertorié dans les exemples d'utilisation qui possèdent ou non un intérêt public mais celui-ci aurait pu potentiellement être démontré. Cependant, après une prise de contact avec l'INDS, il s'avère qu'actuellement, des salariés d'entreprises d'assurance, même dans le cadre d'un mémoire de recherche, ne peuvent fournir des garanties suffisantes pour pouvoir accéder aux données du SNDS, même à un niveau agrégé. Cela prouve que les assureurs sont encore considérés (à tort ou à raison) par les autorités comme de simples acteurs privés avec un fort *a priori* négatif et non comme de véritables partenaires de l'assurance maladie obligatoire dans l'accès aux soins et la prévention des risques. A ce sujet, il est intéressant de noter que les finalités d'exclusion de garanties et de modulation de primes sont explicitement exclues des traitements possibles des données issues du SNDS [Code de la santé publique].

Cela étant dit, les bases du *Health data hub* directement accessibles restent des sources de données très intéressantes. En particulier, la base Open DAMIR correspond aux informations qui semblent répondre à la problématique traitée et contient un volume de données conséquent (cf. § II.1.1.).

I.4.4. Eléments de réponses aux risques identifiés

I.4.4.1. Risques extérieurs

Dans le cadre d'une étude réalisée pour un mémoire d'actuariat, l'analyse d'impact des risques liés à la concurrence ou à la cybersécurité ne nous a pas paru pertinente à ce stade. Cette question devra cependant se reposer en cas de projet en entreprise.

I.4.4.2. Risques liés aux données

De manière assez naturelle, nous nous sommes orientés vers une solution qui n'utilise pas de données à caractère personnel afin de s'absoudre des contraintes réglementaires liées au RGPD. Il est donc indispensable de s'assurer que les données soient bien anonymisées.

En ce qui concerne la partie *open data*, comme évoqué au § I.1.5.3.1., la base Open DAMIR est complètement anonymisée (aucune donnée identifiante et lignes agrégées). Concernant la base de données de l'assureur (Malakoff Humanis[48]), la contrainte de l'anonymisation plus est complexe puisqu'il s'agit de déduire un risque de décès à partir de données suffisamment fines pour pouvoir identifier des signaux faibles. Nous avons donc mis en place plusieurs mécanismes (cf. § II.2.) afin de pseudonymiser les données et réduire au minimum le risque de réidentification. Par ailleurs, nous avons procédé à la validation de l'ensemble de notre traitement par le DPO[49] de l'assureur.

Il est important de noter néanmoins que l'analyse juridique du traitement consistant à rapprocher des données de contrats santé avec des données d'un contrat prévoyance n'a pas été réalisée. En effet, ce traitement s'effectue nécessairement en amont de l'anonymisation et est donc soumis aux obligations du RGPD. Or cette problématique est loin d'être triviale et va dépendre fortement de l'information des personnes présente dans les clauses contractuelles. Il conviendrait en outre de vérifier que le fondement pour effectuer ce traitement est licite. Cela étant, dans le cadre de travaux de recherche, le risque semble très limité. Par ailleurs, il est certain que l'autorisation formelle par le régulateur d'un tel rapprochement entre les données santé et prévoyance ne sera accordé que si son efficacité est prouvée. Or cela n'est pas possible sans la réalisation d'études comme celle que nous présentons dans ce mémoire.

[48] A noter qu'à tout moment, seul l'auteur de ce mémoire qui est salarié de Malakoff Humanis a pu avoir accès aux données de l'assureur.
[49] Le délégué à la protection des données (ou *Data Protection Officer*) est la personne chargée dans une entreprise de traiter toutes les problématiques ayant trait à la protection des données à caractère personnel. Ce poste a été rendu obligatoire par le RGPD.

I.4.4.3. Risques liés aux algorithmes

Les principaux risques liés aux algorithmes (opacité, biais et décisions automatisées) peuvent être adressés en comparant les résultats obtenus avec ce qui a aurait été estimé en utilisant les tables de mortalité réglementaires. Ce faisant, nous nous assurons que nous n'introduisons pas une déformation non justifiée du provisionnement selon les profils des assurés.

Cela étant dit, le fait que les données utilisées (à la fois celles de la base *open data* de santé et celles de l'assureur) concernent des flux de remboursement de soin réduit substantiellement le risque de biais (puisqu'il ne s'agit pas de données libres ni déclaratives mais administratives). Par ailleurs, comme nous n'utilisons le résultat des algorithmes uniquement à des fins de provisionnement (et non de tarification), il n'y a pas d'effet contractuel automatiquement applicable pour les assurés. Dans l'hypothèse où l'algorithme entraînerait une forte différence dans le niveau de provisionnement, la répercussion sur la prime payée par l'assuré ne pourrait s'effectuer sans une analyse préalable par l'assureur et une négociation commerciale avec son client.

I.4.4.4. Risques liés aux utilisations

Comme évoqué au I.3.4.2., le principal risque identifié est celui de la démutualisation introduite par un calcul qui serait effectué à la maille de l'individu. La technique de l'agrégation des données (déjà utile pour réduire drastiquement le risque de réidentification des données) permet de répondre en partie à cette problématique. Ainsi, en agrégeant les individus présentant des caractéristiques similaires au sein de mêmes lignes (que ce soit au sein de la base en *open data* ou au niveau de la base après appariement avec les données de l'assureur), une partie de la mutualisation est conservée.

Enfin, comme déjà évoqué au § I.4.4.3., le cas présenté a été volontairement restreint à des problématiques de provisionnement sans jamais s'intéresser aux questions de tarification. Le fait de ne pas toucher à la prime payée par les assurés *in fine* permet de limiter les effets des utilisations à un périmètre strictement interne à l'assureur : la connaissance du risque par l'assureur peut évoluer ce qui permettra d'optimiser le provisionnement mais le lien avec le profilage commercial n'est pas réalisé. Cela permet à la fois de limiter drastiquement le risque lié à la démutualisation mais également celui de prendre de mauvaises décisions commerciales sur la base des résultats de l'algorithme.

CONCLUSION

Les nouvelles techniques de la *data science* révolutionnent actuellement la récupération et le traitement de l'information : le *big data* (et en particulier l'*open data*) met à disposition de nouvelles sources de données et les algorithmes d'intelligence artificielle de type *machine learning* permettent d'extraire une information nouvelle de cette masse de données. Face à cette révolution numérique, il n'est pas étonnant que les assureurs, dont le cœur de métier était déjà centré autour de la connaissance et de la prédiction du risque, entrevoient de formidables opportunités, notamment pour leurs travaux d'actuariat.

Il est néanmoins primordial de rester vigilant par rapport à ces innovations qui comportent potentiellement de nombreux risques graves. Ainsi, si l'assureur ne veut pas se retrouver dans une situation délicate vis-à-vis de la loi, du régulateur ou de ses assurés, il ne doit pas céder aux sirènes du marketing et bien analyser les méthodes et les conséquences de ces nouveaux modèles afin de créer un cadre éthique de l'utilisation de l'intelligence artificielle. En particulier, il semble nécessaire que les assureurs prennent une position de place vis-à-vis de l'individualisation de l'évaluation du risque permises par ces nouvelles techniques qui peut mettre à mal le principe de mutualisation et remettre en cause le principe même de l'assurance.

Il nous a semblé important d'insister sur les risques portés par l'IA et le *big data* car ceux-ci sont généralement insuffisamment analysés. Cela ne doit pas occulter le fait indéniable que ces innovations proposent de formidables possibilités qui seront probablement très positives dès lors qu'elles sont correctement encadrées. Par exemple, la mesure à une maille très fine du risque pourrait permettre dans certains cas de proposer des produits spécifiques – certes plus chers – à des populations qui sont jusqu'à présent considérées comme techniquement non assurables (personnes en situation d'obésité ou atteintes de cancers par exemple).

Pour concilier ces opportunités et ces risques, nous présentons dans le cadre de ce mémoire un exemple concret de l'utilisation de l'intelligence artificielle et de l'*open data* de santé afin d'améliorer la prédiction du risque décès en prévoyance tout en respectant les contraintes éthiques et réglementaires identifiées.

PARTIE II :

CAS D'USAGE : UTILISATION DE LA BASE OPEN DAMIR POUR PREDIRE LE RISQUE DE DECES

INTRODUCTION

L'*open data* de santé permet l'accès à de nouvelles sources de données d'une grande richesse. Ainsi, la base Open DAMIR offre une quantité impressionnante d'informations dont la plupart n'était pas disponible pour les assureurs auparavant. Cependant, comme pour toute base, il est nécessaire de procéder à des analyses et des traitements des données qu'elle contient avant de pouvoir correctement l'utiliser. En particulier, le traitement des valeurs manquantes[50] au sein de la base représente une problématique importante et complexe qu'il est nécessaire d'adresser (cf. § II.1.).

Ensuite, la base d'étude est constituée en regroupant toutes les informations disponibles : les données d'*open data* et celles de l'assureur. La difficulté va résider dans le fait que les données issues de différentes sources sont parfois difficilement réconciliables. Il faudra donc trouver une méthode pour réaliser leur appariement. Par ailleurs, la constitution de cette base va dépendre de la cible de prédiction du modèle final qui sera donc définie de manière précise (cf. § II.2.).

Enfin, la dernière étape consiste en l'élaboration du modèle de prédiction du risque de décès qui utilisera différentes techniques de data science. La méthode de construction ainsi que l'analyse des résultats obtenus y seront présentées (cf. § II.3.).

II.1. TRAITEMENTS NECESSAIRES SUR LA BASE OPEN DAMIR

II.1.1. Présentation de la base Open DAMIR

II.1.1.1. Qu'est-ce que la base Open DAMIR ?

La base de données Open DAMIR est une base issue du SNIIRAM qui regroupe l'ensemble des dépenses de santé de la population française prises en charge par l'assurance maladie obligatoire quel que soit le régime. Dans le cadre de l'*open data* de santé, cette base a été mise à disposition en téléchargement direct depuis 2015. Elle couvre aujourd'hui une période temporelle allant du 1er janvier 2009 au 31 décembre 2019 [data.gouv.fr].

Il n'est pas prévu dans le cadre de ce mémoire de rappeler le fonctionnement du système de l'assurance santé en France, en particulier l'articulation entre le régime obligatoire proposé par la Sécurité Sociale et le régime complémentaire porté par les assurances, les mutuelles et les instituts de prévoyance. Le lecteur intéressé pourra se rapporter à l'excellent mémoire réalisé par Arnold Mekontso [MEKONTSO FOTSING, A. C. (2018)] sur le sujet.

[50] Le traitement des valeurs aberrantes est une autre problématique importante à analyser au sein d'une base de données et est évoqué en annexe 4.

La base de données du SNIIRAM a été créée en 1999 [Code de la sécurité sociale] tout d'abord dans un objectif de gestion des remboursements traités par l'assurance maladie obligatoire mais surtout dans un but de connaissance et de recherche afin de développer des politiques de santé publique et des campagnes de prévention.

Les données du SNIIRAM proviennent des fichiers administratifs de la Sécurité Sociale, des professionnels de santé et des établissements de soin qui permettent de récupérer « *des informations sur les bénéficiaires, des répertoires de professionnels qui renseignent sur les prestataires de service médicaux, des feuilles de soins et des remboursements, des bordereaux de facturation des cliniques, des arrêts de travail et des indemnités journalières et des résumés de sortie hospitaliers transmis depuis 2007 par l'Agence Technique d'Information sur l'Hospitalisation (ATIH).* » [MEKONTSO FOTSING, A. C. (2018)].

Il est important de noter que la base du SNIIRAM est unique en Europe de par sa richesse et la profondeur de son historique. Bien évidemment, les données qui y sont traités sont des données à caractère personnel sensibles concernant aussi bien les assurés (nom, prénom, date et lieu de naissance, adresse, NIR[51], soins reçus, etc.) que les professionnels de santé (nom, prénom, adresse, spécialité, numéro FINESS[52] et RPPS[53], tarifs pratiqués, etc.). Ces données sont évidemment pseudonymisées au moment de l'intégration dans le SNIIRAM. Néanmoins, comme vu au § I.3.2.5., le fait de simplement retirer les informations directement identifiantes ne suffit pas à garantir l'anonymisation de la base : les données de la base sont « *bien anonymes prises une par une, en ce sens qu'elles ne comportent pas l'identité des personnes, mais elles ne peuvent pas être en accès libre parce qu'en croisant certaines informations qui y figurent, on peut identifier des personnes connues par ailleurs (des proches, des collègues ou des célébrités)* » [BRAS, P.-L. (2013)].

L'accès à la base du SNIIRAM est par conséquent très restreint et encadré par la CNIL [CNAM (2015)]. Les organismes pouvant accéder à la base complète sont déterminés par arrêté et chaque utilisateur est formellement et nominativement habilité. Chaque requête effectuée est tracée. Ce traçage permet de vérifier que les utilisateurs ne puissent pas croiser des données aboutissant potentiellement à la réidentification d'individus (interdiction de traitements portant sur moins de 10 individus).

[51] Numéro d'Inscription au Registre National d'Identification des Personnes Physique (RNIPP) géré par l'INSEE, utilisé comme numéro identifiant pour la Sécurité Sociale. Il s'agit d'une donnée extrêmement sensible car elle est attribuée de manière unique à chaque individu et elle est chargée en information (sexe, date et lieu de naissance, nationalité)

[52] Fichier National des Établissements Sanitaires et Sociaux qui assure l'immatriculation des établissements sanitaires, sociaux, médico-sociaux, et de formation aux professions de ces secteurs.

[53] Répertoire Partagé des Professionnels de Santé qui assure l'immatriculation des professionnels de santé (médecins, infirmiers, pharmaciens, sages-femmes, chirurgiens-dentistes, kinésithérapeutes et pédicures-podologues)

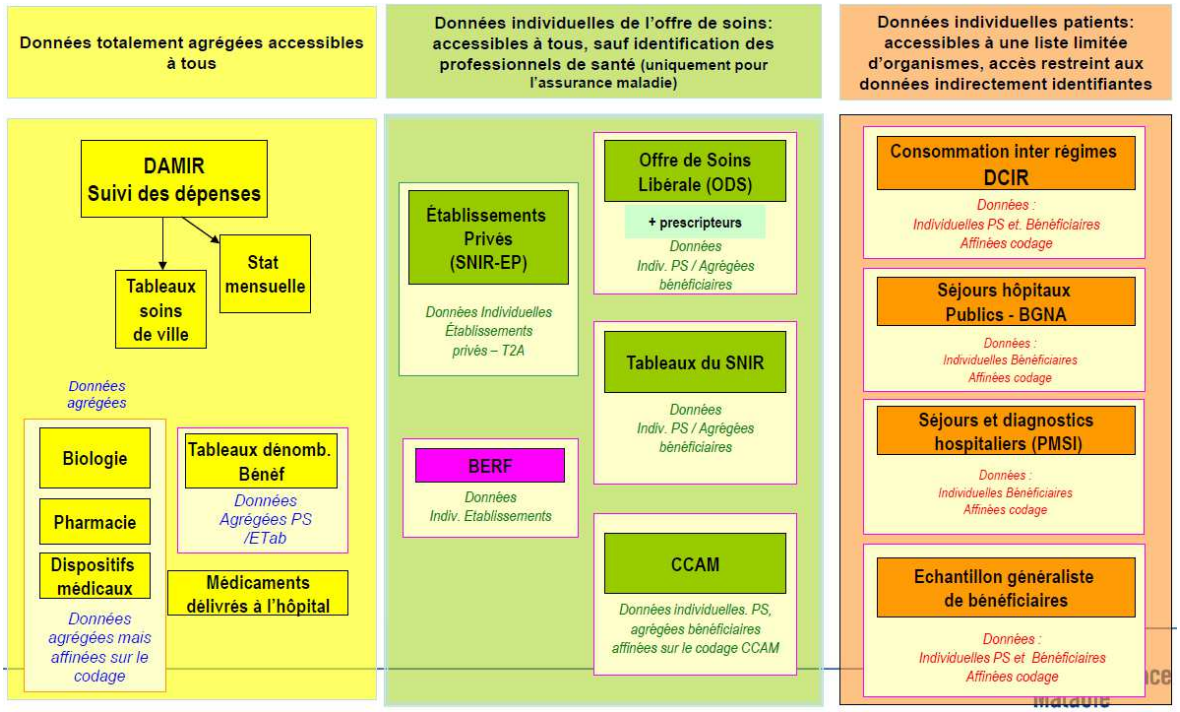


Figure 8 – Les trois niveaux d'accès aux données du SNIIRAM [CNAM (2015)]

Afin d'aboutir à la base Open DAMIR, il a fallu véritablement anonymiser une partie de la base du SNIIRAM (celle qui concerne spécifiquement les dépenses de santé). Comme vu au § I.3.2.5., cette anonymisation passe nécessairement par une agrégation des données afin d'éviter tout risque de réidentification.

II.1.1.2. Présentation du contenu de la base

La base Open DAMIR comporte 55 variables (le détail est disponible sur data.gouv.fr [data.gouv.fr]) :

- 13 données quantitatives représentant des montants de dépense de santé ou des nombres d'actes ;
- 42 données qualitatives (codées sur des entiers) représentant des catégories de classification selon 4 axes : l'acte de soin, le bénéficiaire, le professionnel de santé et l'organisme d'assurance maladie.

La base Open DAMIR regroupe 10 années de remboursement de frais de santé. Chaque mois de données pèse en moyenne 5 Go en format CSV et compte entre 15 et 35 millions de lignes[54]. Au total, la base compte quasiment 530 To de données et comporte 3,34 milliards de lignes.

[54] Il est intéressant de noter que le nombre de lignes a sensiblement augmenté (+36% en 2015 par rapport à 2014) à partir de 2015. Cela est dû au changement des 6 variables liées à la répartition régionale qui sont passées de 9 à 13 modalités.

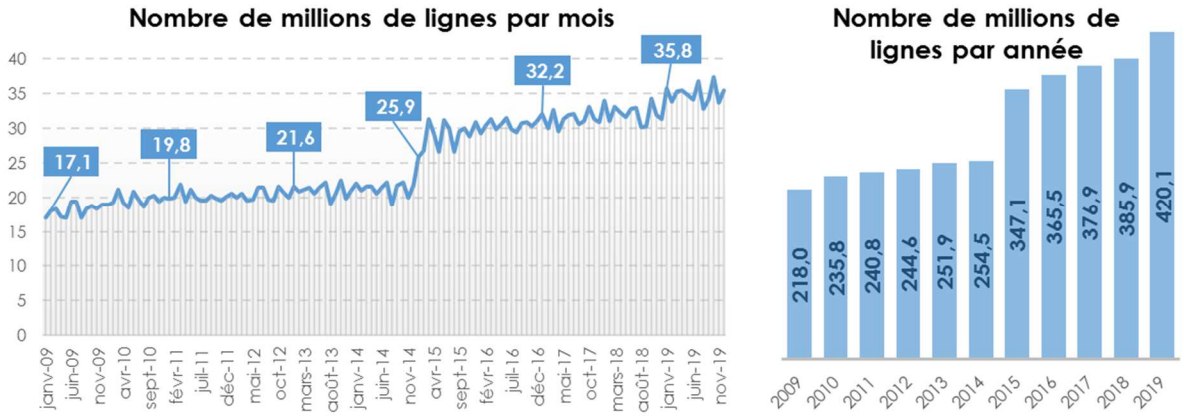


Figure 9 – Evolution temporelle du nombre de lignes de la base Open DAMIR

Cette volumétrie très importante est une problématique compliquée à traiter. Il est ainsi impossible d'ouvrir un seul mois de données dans un classeur Excel et le moindre algorithme nécessite une bonne puissance de calcul (et beaucoup de temps) pour pouvoir tourner sur une année complète. Pour cette raison, nous avons choisi de réaliser nos analyses à l'aide du langage Python qui, grâce aux environnements virtuels mis à disposition par nos entreprises, est capable de traiter une grande volumétrie tout en proposant un grand nombre de bibliothèques avec des outils de traitements de données et de *machine learning* performant.

La gestion des valeurs manquantes représente un problème récurrent dans tout traitement statistique d'une base, que celle-ci soit petite ou grande. La base Open DAMIR n'échappe malheureusement pas à cette problématique. Bien au contraire, les valeurs identifiées comme manquantes au sein de la base de données peuvent représenter pour certaines variables la majorité des lignes :

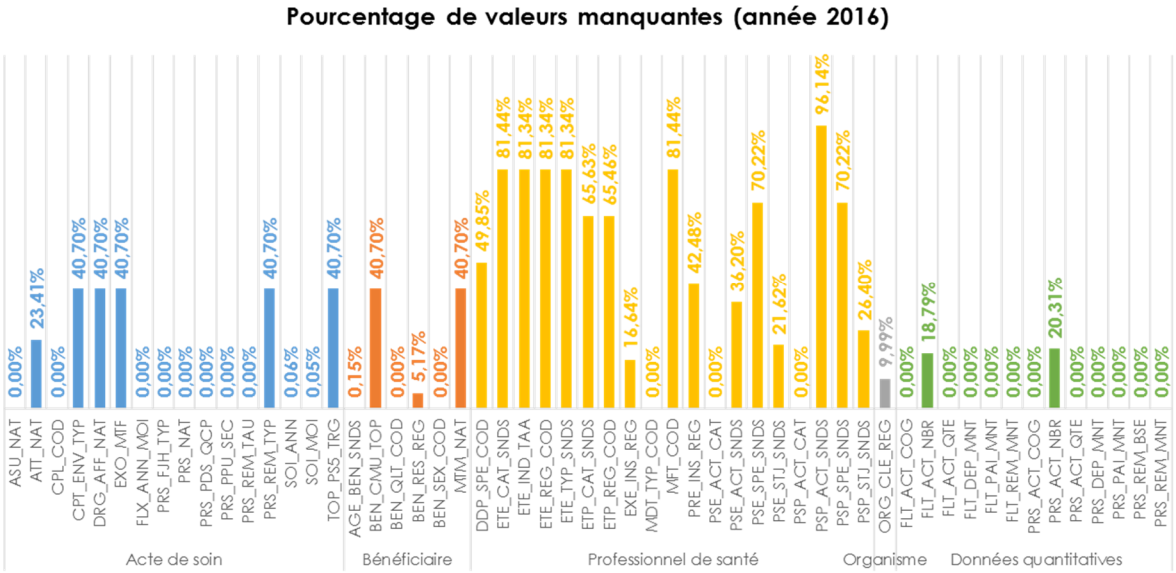


Figure 10 – Proportion de valeurs manquantes par variable dans la base Open DAMIR (2016)

Ainsi, la part des valeurs manquantes au sein de la base Open DAMIR est très disparate selon les variables avec 18 variables présentant un taux supérieur à 40% des lignes. Ce taux atteint plus de 96% dans le pire cas. Un point reste positif cependant : les variables quantitatives représentant les montants de remboursement (et par conséquent le cœur de l'information de la base de données) présentent un très faible nombre de valeurs manquantes.

Il semble par conséquent indispensable d'élaborer un modèle permettant de traiter les valeurs manquantes au sein de la base Open DAMIR avant de pouvoir l'apparier avec les données de l'assureur.

II.1.2. Analyse des valeurs manquantes

II.1.2.1. Théorie des valeurs manquantes

II.1.2.1.1. Pourquoi traiter les valeurs manquantes ?

Dans toute base de données, la problématique des valeurs manquantes et de leur traitement se pose. La base Open DAMIR ne fait pas exception. Une valeur manquante peut-être liée à une absence de réponse réelle, à une réponse inexploitable ou à un dysfonctionnement du système d'information. Dans certain cas, l'absence de valeur sera purement aléatoire ; dans d'autres, il pourra s'agir d'un manquement volontaire ; enfin la non-réponse peut être justifiée par une non applicabilité de la variable dans certaines situations. Cette variété de causes potentielles de non-réponse va potentiellement introduire un risque de biais au sein des données qui peut dans le pire des cas fausser les conclusions des analyses effectuées.

L'objectif du traitement des valeurs manquantes est d'utiliser la connaissance sur les données observées afin de compenser l'information perdue à l'aide de techniques mathématiques et statistiques. Dans la cadre de notre mémoire, le but est de déterminer pour chaque donnée manquante la valeur la plus proche de la donnée réelle. En d'autres termes, il s'agit donc d'un cas particulier de prédiction statistique.

II.1.2.1.2. Différents types de valeurs manquantes

La problématique des données manquantes et les mécanismes de non-réponse ont été théorisé en 1987 par Little et Rubin [LITTLE, R. J. A. et RUBIN, D. B. (1987)]. Ils ont proposé la typologie suivante :

- MCAR (Missing Completely At Random) : La probabilité que la donnée soit manquante est identique pour toutes les observations. Cela implique que la cause de l'absence de l'information est indépendante des données.

- MAR (Missing At Random) : La probabilité que la donnée soit manquante est liée à une ou plusieurs autres variables observées. Il s'agit du cas le plus fréquent. Cela correspond par exemple au cas d'un sondage au sein duquel les personnes interrogées les plus âgées auraient moins tendance à répondre aux questions à propos de leurs opinions politiques.

- MNAR (Missing Not At Random) : La probabilité que la donnée soit manquante dépend d'informations inconnues ; en particulier, elle peut dépendre de la valeur réelle de la variable manquante. Une illustration classique de ce phénomène est le fait que les personnes disposant d'un haut niveau de salaire indiquent moins fréquemment le montant de leur rémunération ou encore les personnes votant pour un parti politique d'extrême-droite auront davantage tendance à ne pas répondre sur leurs intentions de vote.

Nous ne détaillons pas ici le formalisme mathématique introduit par Little et Rubin. Le point important à retenir est que l'apparition de valeurs manquantes ne va pas introduire de biais dans les métriques issues des données uniquement dans le cas MCAR. A l'inverse, lorsque les données sont MNAR, il devient très compliqué d'effectuer des traitements pour compenser les valeurs manquantes car dans ce cas, la distribution des données est différente entre les données observées et les données manquantes.

II.1.2.1.3. Techniques d'imputation des valeurs manquantes

L'imputation des valeurs manquantes consiste à remplacer les données manquantes avec la meilleure estimation de ces valeurs. Il existe plusieurs modèles statistiques plus ou moins complexes permettant de réaliser cette imputation. Seules les techniques retenues sont détaillées en annexe 7. Les autres techniques sont présentées en détail, notamment dans les ouvrages de Little et Rubin [LITTLE, R. J. A. et RUBIN, D. B. (2002)], de Schafer [SCHAFER, J. L. (1997)] et de Van Buuren [VAN BUUREN, S. (2018)].

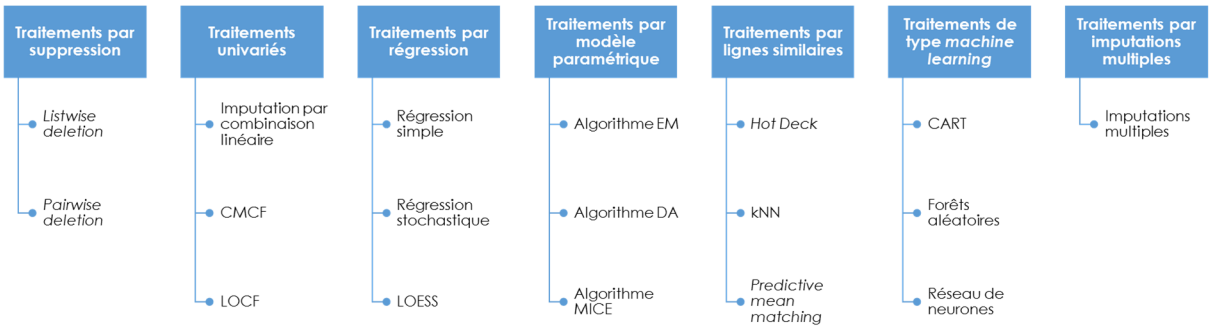


Figure 11 – Liste des principales méthodes de traitement de valeurs manquantes

II.1.2.2. Valeurs manquantes au sein d'Open DAMIR

II.1.2.2.1. Mécanisme d'apparition des valeurs manquantes

Afin de pouvoir traiter la problématique des valeurs manquantes, l'information concernant le taux de remplissage de la variable est loin d'être suffisant. Il est en effet essentiel de bien analyser la répartition des données manquantes au sein de la base afin de repérer des éventuelles corrélations entre elles et ainsi déterminer si nous nous trouvons dans un cas MCAR, MAR ou MNAR.

Comme expliqué dans au § II.1.1., la base de données Open DAMIR est issue du SNIIRAM, c'est-à-dire qu'elle est l'émanation anonymisée d'une base de données administrative dont les informations sont collectées grâce à la gestion des remboursements des dépenses de santé des Français par l'assurance maladie obligatoire. Ce point est essentiel car il nous permet de conclure de manière assez fiable que la base Open DAMIR a peu de risque de présenter des valeurs manquantes de type MNAR dès lors que les individus ne sont pas à l'origine des données collectées (comme c'est le cas lors d'un sondage ou d'une étude clinique par exemple). Nous excluons par conséquent le cas MNAR pour la construction de notre solution de traitement des valeurs manquantes.

Par ailleurs, nous présentons que certaines variables n'ont pas de sens selon la nature de la prestation de santé ce qui permettrait d'expliquer le fort pourcentage de valeurs manquantes pour certaines variables et ainsi que le fait que lorsque l'une de ces variables est manquantes, alors les autres le sont également. Cette hypothèse signifierait que le cas MCAR est également exclu et que nous nous situons donc dans le cas MAR.

Afin de confirmer cette hypothèse, nous avons étudié les *patterns*[55] des données manquantes en identifiant les variables qui présentent un comportement identique grâce au pourcentage des valeurs manquantes sur l'intégralité des données[56] :

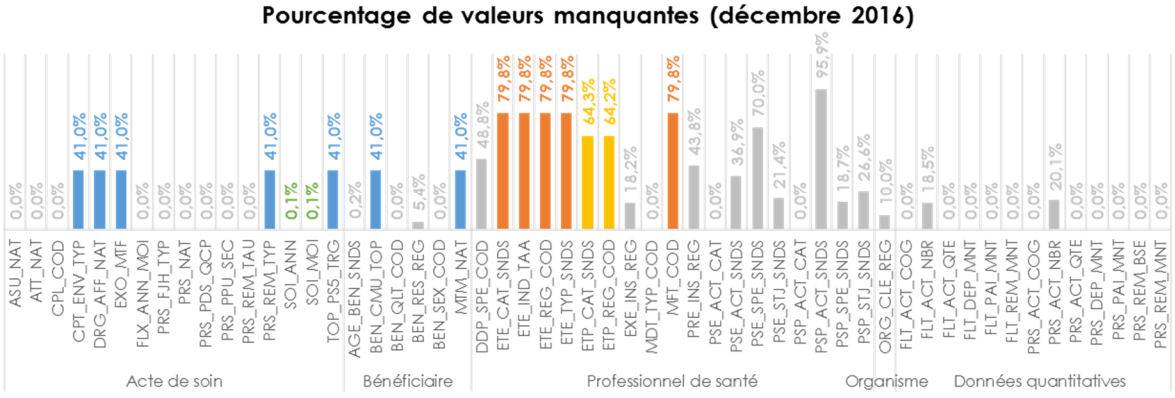


Figure 12 – Répartition des valeurs manquantes toutes les lignes de la base Open DAMIR (déc. 2016)

[55] Le concept de *pattern* de données manquantes est formellement défini en annexe 6.
[56] La fréquence d'apparition des valeurs manquantes n'est pas suffisante pour conclure que les données sont manquantes sur exactement les mêmes lignes mais la précision donne une forte suspicion de l'existence de ces *patterns*. Par ailleurs, cette suspicion a été confirmée par une analyse ligne à ligne sur des fenêtres de 100 à 4 000 lignes (non présentée ici).

Nous arrivons donc à identifier des *patterns* d'apparition de valeurs manquantes avec la mise en évidence d'au moins quatre groupes de variables présentant des données manquantes systématiquement aux mêmes lignes.

Par ailleurs, en analysant les données manquantes selon la nature de la prestation, nous observons un lien de dépendance fort, certaines variables présentant une fréquence de valeurs manquantes très différentes selon la nature de la prestation :

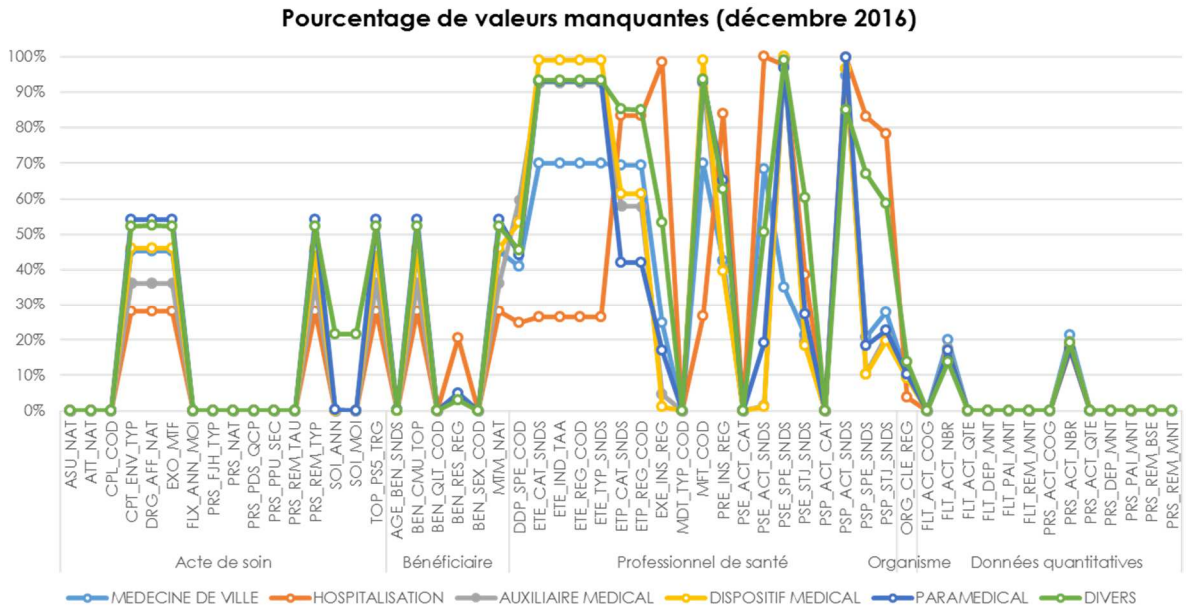


Figure 13 – Répartition selon la nature d'acte des valeurs manquantes d'Open DAMIR (décembre 2016)

Ces analyses prouvent qu'il existe bien au sein de la base Open DAMIR différents groupes de variables liées entre elles en ce qui concerne la distribution des données manquantes. Nous en concluons donc que l'hypothèse MCAR est exclue et par conséquent que la base Open DAMIR présente des données manquantes de type MAR. Cela signifie qu'il est possible d'appliquer à la base des modèles d'imputation.

II.1.2.2.2. Traitements préalables à appliquer

Avant tout traitement statistique, il est nécessaire de réaliser une analyse métier des données afin de vérifier si les données considérées sont réellement manquantes ou bien s'il faut les sortir du périmètre de l'analyse, soit parce qu'elles peuvent être considérées comme sans objet, soit parce que l'information qu'elles contiennent n'est pas pertinente.

En synthèse, nous proposons de modifier les variables présentant des données manquantes comme détaillé dans le tableau 1.

Variables	Valeurs de PRS_NAT	Traitement spécifique
ETE_CAT_SNDS ETE_REG_COD ETE_TYP_SNDS MFT_COD ETE_IND_TAA	Hospitalisation hors indemnité journalière	Analyse normale des données manquantes
	Autres valeurs	Pas d'analyse des données manquantes (sans objet)
ETP_REG_COD ETP_CAT_SNDS	Toutes les valeurs	Pas d'analyse des données manquantes
PSP_ACT_SNDS	Toutes les valeurs	Pas d'analyse des données manquantes
PSE_SPE_SNDS	Médecine de ville et acte médical en hospitalisation	Analyse normale des données manquantes
	Autres valeurs	Pas d'analyse des données manquantes (sans objet)

Tableau 1 – Détail du périmètre retenu ou exclu de l'analyse des données manquantes

La mise en place de ces mesures spécifiques permet d'améliorer sensiblement le profil des valeurs manquantes.

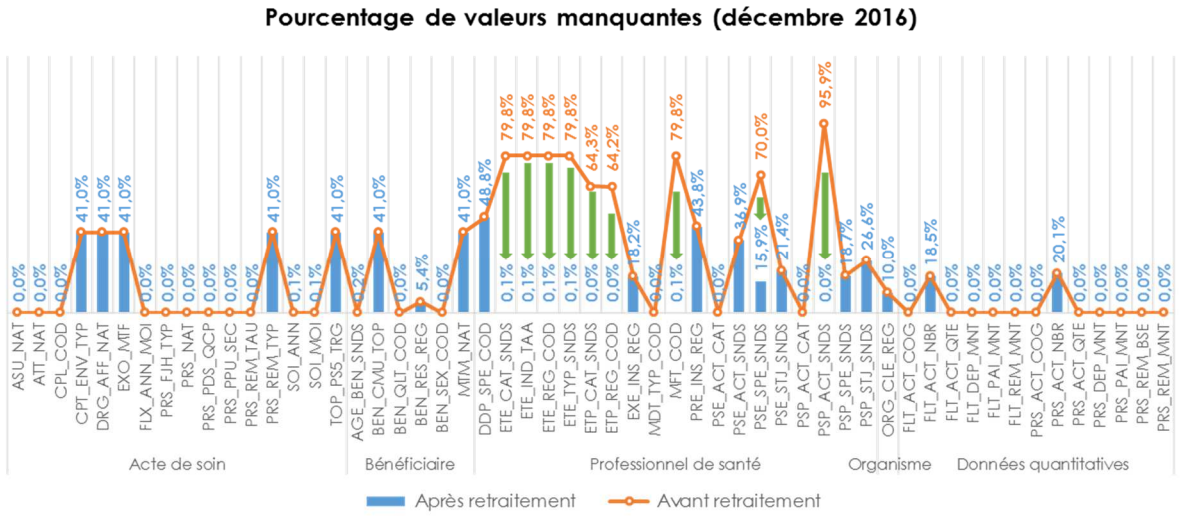


Figure 14 – Evolution de la répartition des valeurs manquantes avec le retraitement

Dans ces conditions, nous trouvons *in fine* 134 476 lignes complètes (soit 0,43% de la base initiale). Ce résultat peut paraître peu en proportion mais reste suffisant afin de créer une des bases permettant d'entraîner et de comparer les différents modèles de traitement des valeurs manquantes envisagés (cf. § II.1.3.3.).

II.1.2.2.3. Choix des méthodes d'imputation à implémenter

Grâce à ces analyses préliminaires et aux caractéristiques de la base Open DAMIR, nous avons pu sélectionner les méthodes qui nous semblent les plus pertinentes pour l'imputation des données manquantes.

Méthode	Avantages	Inconvénients	Décision
Traitements par suppression			
Listwise deletion	▶ Très simple	▶ Biais si MAR ▶ Supprime 99,57% des lignes	✓ *(pour les bases d'entrainement et de validation)*
Pairwise deletion	▶ Très simple	▶ Biais si MAR ▶ Pas adapté s'il s'agit de prédiction (et non d'estimation)	✗
Traitements univariés			
Imputation par la moyenne	▶ Très simple ▶ Traitement de base	▶ Biais si MAR	✓ *(pour les variables quantitatives)*
CMCF	▶ Très simple ▶ Traitement de base	▶ Biais si MAR	✓ *(pour les variables quantitatives)*
LOCF	▶ Très simple ▶ S'applique bien aux bases mixtes	▶ Biais si MAR ▶ Uniquement adapté pour les séries temporelles	✗
Traitements par régression			
Régression simple	▶ Simple ▶ Prise en compte de certaines corrélations	▶ Biais si MAR ▶ Peu adapté aux données qualitatives	✗
Régression stochastique	▶ Simple ▶ Bons résultats même si MAR ▶ Prises en compte de certaines corrélations ▶ Adapté à l'imputation multiple	▶ Peu adapté aux données qualitatives	✗
LOESS	▶ Plus fin que les régressions classique	▶ Peu adapté aux données qualitatives	✗

Méthode	Avantages	Inconvénients	Décision
Traitements par modèle paramétrique			
Algorithme EM	▶ Bons résultats même si MAR	▶ Nécessite un modèle multivarié pour les données ▶ Peu adapté aux bases mixtes ▶ Risque de non convergence s'il y a trop de variables	✘
Algorithme DA	▶ Bons résultats même si MAR ▶ Adapté à l'imputation multiple	▶ Nécessite un modèle multivarié pour les données ▶ Peu adapté aux bases mixtes	✘
Algorithme MICE	▶ Bons résultats même si MAR ▶ Ne nécessite pas de modèle multivarié ▶ Adapté à l'imputation multiple	▶ Nécessite une série de modèles conditionnels pour chaque variable ▶ Lourd à implémenter ▶ Peu adapté aux bases mixtes	✘
Traitements par lignes similaires			
Hot deck	▶ Simple ▶ S'applique bien aux bases mixtes	▶ Risque de biais si la base est trop petite	✘
kNN	▶ Simple ▶ Bons résultats même si MAR ▶ S'applique bien aux bases mixtes	▶ Risque de biais si la base est trop petite	✓
Predictive mean matching	▶ Simple ▶ Bons résultats même si MAR ▶ S'applique bien aux bases mixtes ▶ Adapté à l'imputation multiple	▶ Risque de biais si la base est trop petite	✘

Méthode	Avantages	Inconvénients	Décision
Traitements de type *machine learning*			
CART	▶ Simple ▶ Plus efficace si le nombre de variables est grand ▶ S'applique bien aux bases mixtes	▶ Instable ▶ Peu utilisé en pratique	✘
Random forest	▶ Simple ▶ Bon résultats même si MAR ▶ Plus efficace si le nombre de variables est grand ▶ S'applique bien aux bases mixtes ▶ Adapté à l'imputation multiple	▶ Non applicable aux petites bases de données	✔
Réseau de neurones	▶ Algorithme potentiellement très puissant ▶ Plus efficace si le nombre de variables est grand ▶ S'applique bien aux bases mixtes	▶ Complexe ▶ Rarement utilisé pour prédire des valeurs manquantes ▶ Non applicable aux petites bases de données	✔
Traitements par imputations multiples			
Imputations multiples	▶ Bons résultats même si MAR ▶ Permet une estimation de l'erreur d'imputation	▶ Très lourd à implémenter	✘[57]

Tableau 2 – Analyse comparative des différentes méthodes d'imputation

Comme le montre le tableau 2, beaucoup de méthodes d'imputation des valeurs manquantes ne sont véritablement adaptées qu'aux variables quantitatives. Or, comme vu dans le § II.1.2.2., les données manquantes de la base Open DAMIR concerne principalement les variables qualitatives. Cette contrainte va grandement restreindre les choix de méthodes qu'il est possible d'implémenter. Nous avons ainsi décidé de retenir quatre méthodes d'imputation des données manquantes :

- une combinaison d'imputation par la moyenne et d'imputation par le mode comme méthode de référence (puisqu'il s'agit d'une technique très basique) ;
- la méthode kNN ;
- la méthode *random forest* ;
- la méthode du réseau de neurones.

[57] Il aurait été intéressant d'implémenter la méthode d'imputations multiples, notamment via l'algorithme de *predictive mean matching* ou de *random forest* afin d'obtenir des résultats certainement plus pertinents et de pouvoir effectuer une analyse de l'intervalle de confiance autour de nos valeurs imputées. Néanmoins, cela aurait ajouté énormément de complexité calculatoire compte-tenu de la masse de données traitées.

II.1.3. Paramétrage des modèles d'imputation

II.1.3.1. Construction des bases d'entraînement et de validation

Les méthodes permettant de construire, de paramétrer et d'évaluer un modèle statistique (en particulier ceux utilisant des algorithmes de type *machine learning*) sont rappelées en annexe 5.

Comme vu au § II.1.2.2.2., la base Open DAMIR pour le mois de décembre 2016 contient 134 476 lignes de données complètes. Conformément à ce qui a été décrit en annexe 5, dans un premier temps, nous avons extrait ces lignes. Afin de pouvoir évaluer nos modèles d'imputation, il est nécessaire de créer au sein de cette base de données complètes des valeurs manquantes artificielles.

Cette problématique est loin d'être triviale car le mécanisme d'apparition de ces valeurs manquantes artificielles doit être similaire à celui des véritables données manquantes au sein de la base Open DAMIR. Nous avons donc élaboré un modèle spécifique permettant la création de valeur manquantes artificielles. Celui-ci est détaillé dans l'annexe 6. Au final, nous disposons donc de deux bases identiques : l'une présentant des valeurs manquantes et l'autre non.

Afin de construire les bases d'entrainement et de validation nécessaires à la méthode de validation croisée (cf. annexe 5), nous avons utilisé une fonction native de la bibliothèque *Sklearn* de Python. Il existe plusieurs fonctions permettant d'effectuer une validation croisée dans *Sklearn* (*KFold*, *Repeated KFold*, *Leave One Out*, etc.). Dans le cadre de nos travaux, nous avons retenu la fonction *Shuffle & Split* avec la base de validation représentant 10% des données complètes. Cette méthode rajoute, par rapport à la méthode classique *KFold*, un aspect stochastique en effectuant un tirage aléatoire des lignes qui seront utilisées comme jeux d'entrainement ou de validation (ce qui est très utile pour éviter un biais lorsque les données de la base sont triées par exemple).

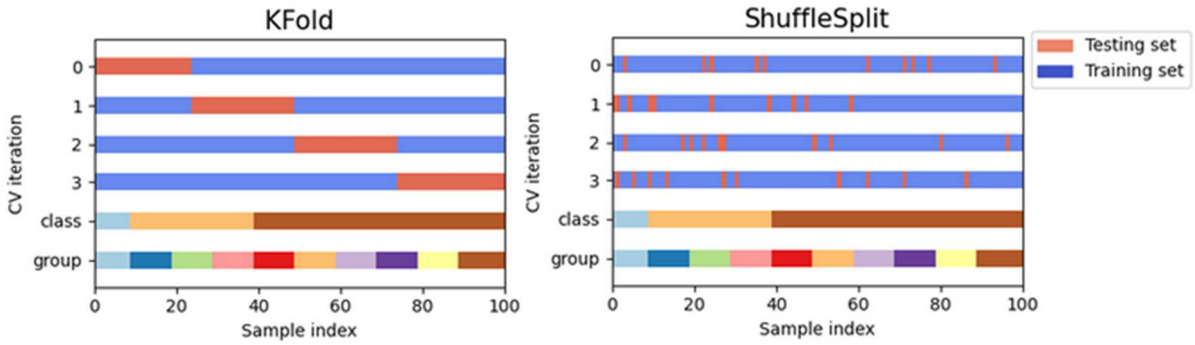


Figure 15 – Illustration des méthodes KFold et Shuffle & Split [scikit-learn.org]

II.1.3.2. Mesure de la qualité statistique

Le principe de la qualité statistique est de mesurer l'erreur entre la prédiction du modèle et la véritable valeur prise par la donnée. Cela consiste donc à évaluer la distance entre la valeur prédite $\widehat{\boldsymbol{y}}$ et la vraie valeur $\boldsymbol{y}$ (qui est connue puisque nous avons construit les bases de validation spécifiquement pour cela). Il est par conséquent indispensable de définir formellement la mesure de distance utilisée.

II.1.3.2.1. Variables quantitatives

Pour les variables quantitatives, les distances les plus utilisées sont la distance euclidienne[58] (qui donne l'erreur quadratique moyenne ou RMSE[59]) ou bien la distance de Manhattan[60] (qui donne l'erreur absolue moyenne ou MAE[61]). La RMSE utilise une mise au carrée ce qui va avoir tendance à mettre un poids plus important aux grosses erreurs. Inversement, la MAE va considérer toutes les erreurs de manière uniforme[62].

Cela dit, aucune de ces deux mesures d'erreur ne permet de comparer les résultats de diverses variables avec potentiellement un nombre différent de données manquantes. Nous proposons par conséquent d'utiliser comme mesure d'erreur pour les variables quantitatives le coefficient de détermination R^2 qui correspond à une version normalisée du RMSE. Cet indicateur est défini comme le rapport de la somme des carrés expliqués (SSE pour *Sum of Squared Explained*) sur la somme des carrés totaux (SST pour *Sum of Squared Total*). Ainsi, pour une variable quantitative $\boldsymbol{y_j}$ (avec $j \in [\![1, p]\!]$) de la base de données $\boldsymbol{Y}$ et $\widehat{\boldsymbol{y}}_{\boldsymbol{j}}$ son estimateur associé, R_j^2 s'écrit :

$$R_j^2 = \frac{SSE_j}{SST_j} = \frac{\sum_{i=1}^{n}\left(\hat{y}_{ij} - \bar{y}_{ij}\right)^2}{\sum_{i=1}^{n}\left(y_{ij} - \bar{y}_{ij}\right)^2} = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_{ij} - y_{ij}\right)^2}{\sum_{i=1}^{n}\left(y_{ij} - \bar{y}_{ij}\right)^2} = 1 - \frac{n \times RMSE_j^2}{\sum_{i=1}^{n}\left(y_{ij} - \bar{y}_{ij}\right)^2}$$

Plus le coefficient R_j^2 est proche de 100% et meilleure est l'estimation.

---

[58] La distance euclidienne $d_E(\boldsymbol{x}, \boldsymbol{y})$ entre deux vecteurs $\boldsymbol{x}(x_1, x_2, \dots, x_n)$ et $\boldsymbol{y}(y_1, y_2, \dots, y_n)$ est la racine de la somme des écarts au carré : $d_E(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$.

[59] La RMSE (pour *Root Mean Squared Error*) d'un estimateur $\widehat{\boldsymbol{Q}} = (\hat{q}_1, \hat{q}_2, \dots, \hat{q}_n)$ d'une grandeur statistique $\boldsymbol{Q} = (q_1, q_2, \dots, q_n)$ est $RMSE = \sqrt{\mathbb{E}\left[\left(\widehat{\boldsymbol{Q}} - \boldsymbol{Q}\right)^2\right]} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{q}_i - q_i)^2}{n}}$.

[60] La distance de Manhattan $d_M(\boldsymbol{x}, \boldsymbol{y})$ entre deux vecteurs $\boldsymbol{x}(x_1, x_2, \dots, x_n)$ et $\boldsymbol{y}(y_1, y_2, \dots, y_n)$ est la somme des valeurs absolues des écarts: $d_M(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n}|y_i - x_i|$.

[61] La MAE (pour *Mean Absolute Error*) d'un estimateur $\widehat{\boldsymbol{Q}} = (\hat{q}_1, \hat{q}_2, \dots, \hat{q}_n)$ d'une grandeur statistique $\boldsymbol{Q} = (q_1, q_2, \dots, q_n)$ est $MAE = \mathbb{E}\left[\left|\widehat{\boldsymbol{Q}} - \boldsymbol{Q}\right|\right] = \frac{\sum_{i=1}^{n}|\hat{q}_i - q_i|}{n}$.

[62] Le choix de la mesure d'erreur entre RMSE et MAE fait débat au sein de la communauté scientifique [WILLMOTT, C. et MATSUURA, K. (2005) vs CHAI, T. et DRAXLER, R. (2014)]

II.1.3.2.2. Variables qualitatives

Pour les variables qualitatives, il existe différentes distances permettant d'évaluer la différence entre deux chaînes de caractère comme la distance de Levenshtein[63]. Cependant, comme les valeurs qualitatives sont représentées par des nombres dans la base Open DAMIR, il est plus pertinent d'utiliser la distance de Hamming[64] : concrètement, pour une valeur estimée donnée, la distance vaudra 0 si elle a bien été estimée par le modèle et 1 sinon. Cette distance est souvent appelée *accuracy* dans les bibliothèques Python de *machine learning*.

Cependant, comme pour les variables quantitatives, nous avons besoin de comparer les performances du modèle sur plusieurs variables qui ne présentent pas le même nombre de valeurs manquantes. Nous allons donc définir un taux de prédiction T en normalisant la distance de Hamming par le nombre de données manquantes à imputer. Ainsi, pour une variable qualitative $\boldsymbol{y_j}$ (avec $j \in [\![1, p]\!]$) de la base de données $\boldsymbol{Y}$ présentant n_j^{miss} valeurs manquantes et $\widehat{\boldsymbol{y}}_{\boldsymbol{j}}$ son estimateur associé, T_j s'écrit :

$$T_j = 1 - \frac{d_H(\widehat{\boldsymbol{y}}_{\boldsymbol{j}}, \boldsymbol{y_j})}{n_j^{miss}} = 1 - \frac{\sum_{i=1}^{n} \mathbf{1}_{\{\hat{y}_{ij} \neq y_{ij}\}}}{n - \sum_{i=1}^{n} r_{ij}}$$

Plus le coefficient T_j est proche de 100% et meilleure est l'estimation.

II.1.3.2.3. Performance globale du modèle

La qualité statistique globale $\mathcal{Q}$ du modèle sera alors une moyenne de la qualité statistique de chaque variable $\boldsymbol{y_j}$ (R_j^2 ou T_j selon le cas), pondérée par le nombre de valeurs manquantes n_j^{miss} :

$$\mathcal{Q} = \frac{\sum_{\boldsymbol{y_j}\ \text{quanti}} \left(R_j^2 \times n_j^{miss}\right) + \sum_{\boldsymbol{y_j}\ \text{quali}} \left(T_j \times n_j^{miss}\right)}{\sum_{j=1}^{p} n_j^{miss}}$$

Ce faisant, nous cherchons à optimiser le nombre total de données manquantes correctement imputées par le modèle sans y intégrer d'ordre d'importance d'une variable par rapport à une autre[65].

[63] La distance de Levenshtein $d_L(\mathcal{S}, \mathcal{T})$ entre deux chaînes de caractères $\mathcal{S}$ et $\mathcal{T}$ correspond au nombre d'opérations élémentaires (ajout, suppression ou substitution de caractère) à appliquer sur $\mathcal{S}$ afin d'obtenir $\mathcal{T}$.

[64] La distance de Hamming $d_H(\boldsymbol{x}, \boldsymbol{y})$ entre deux vecteurs $\boldsymbol{x}(x_1, x_2, \dots, x_n)$ et $\boldsymbol{y}(y_1, y_2, \dots, y_n)$ est le nombre de cas où $x_i \neq y_i$ pour $i \in [\![1, n]\!]$: $d_H(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} \mathbf{1}_{\{x_i \neq y_i\}}$.

[65] Il s'agit ici d'un choix arbitraire. Nous aurions en effet tout aussi bien pu choisir de privilégier certaines variables comme celles relatives au profil du bénéficiaire.

II.1.3.3. Implémentation des modèles

II.1.3.3.1. Imputation par la moyenne et par le mode

L'imputation par la moyenne (pour les variables quantitatives) et par le mode (pour les variables qualitatives) est une méthode particulièrement simple à mettre en œuvre puisqu'elle ne nécessite aucun paramétrage.

D'un point de vue programmation, nous avons tout simplement utilisé les fonctions *mean* et *mode* disponibles au sein de la bibliothèque *Pandas*.

II.1.3.3.2. Imputation par kNN

L'imputation par la méthode des k plus proches voisins est à peine plus compliquée puisque le seul[66] hyperparamètre à optimiser est le nombre k.

D'un point de vue programmation, nous avons utilisés les fonctions *KNeighborsRegressor* (pour l'imputation des variables quantitatives) et *KNeighborsClassifier* (pour l'imputation des variables qualitatives) de la bibliothèque *Sklearn*. En appliquant différents paramétrages sur les bases d'entraînement et de test obtenues par validation croisée, nous obtenons les résultats ci-dessous (sur la base de 769 377 valeurs imputées réparties sur 28 variables) :

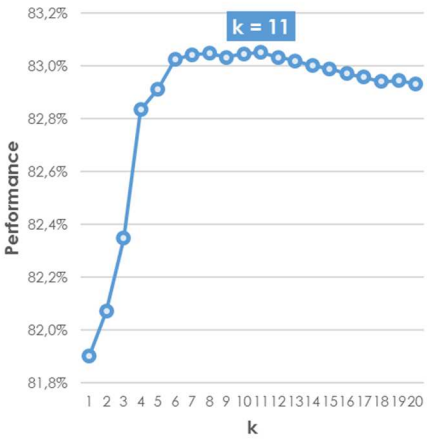


Figure 16 – Performance de l'algorithme kNN

Ces résultats illustrent parfaitement le phénomène de surapprentissage qui apparaît lorsque $k > 11$. Nous choisissons donc logiquement $k = 11$.

[66] En réalité, il existe d'autres hyperparamètres mais qui servent soit à optimiser le temps de calcul, soit à adapter la méthode en termes de distance ou d'algorithme d'optimisation et qui n'ont par conséquent pas été étudiés dans le cadre de ce mémoire.

II.1.3.3.3. Imputation par random forest

L'imputation par forêts aléatoires est plus compliquée à paramétrer puisque l'algorithme propose de nombreux hyperparamètres. Pour notre analyse, nous en avons retenu quatre :

- n_{tree} : nombre d'arbres au sein de la forêt ;
- max_{depth} : nombre maximum de niveaux de l'arbre ;
- max_{feature} : nombre de variables considérées pour séparer un nœud en deux ;
- min_{leaf} : nombre minimum de valeurs au sein d'un nœud pour le considérer comme terminal.

Au niveau de la programmation, nous avons utilisés les fonctions *RandomForestRegressor* et *RandomForestClassifier* de la bibliothèque *Sklearn*. En appliquant différents paramétrages[67] sur les mêmes bases que pour la méthode kNN, nous obtenons les résultats ci-dessous :

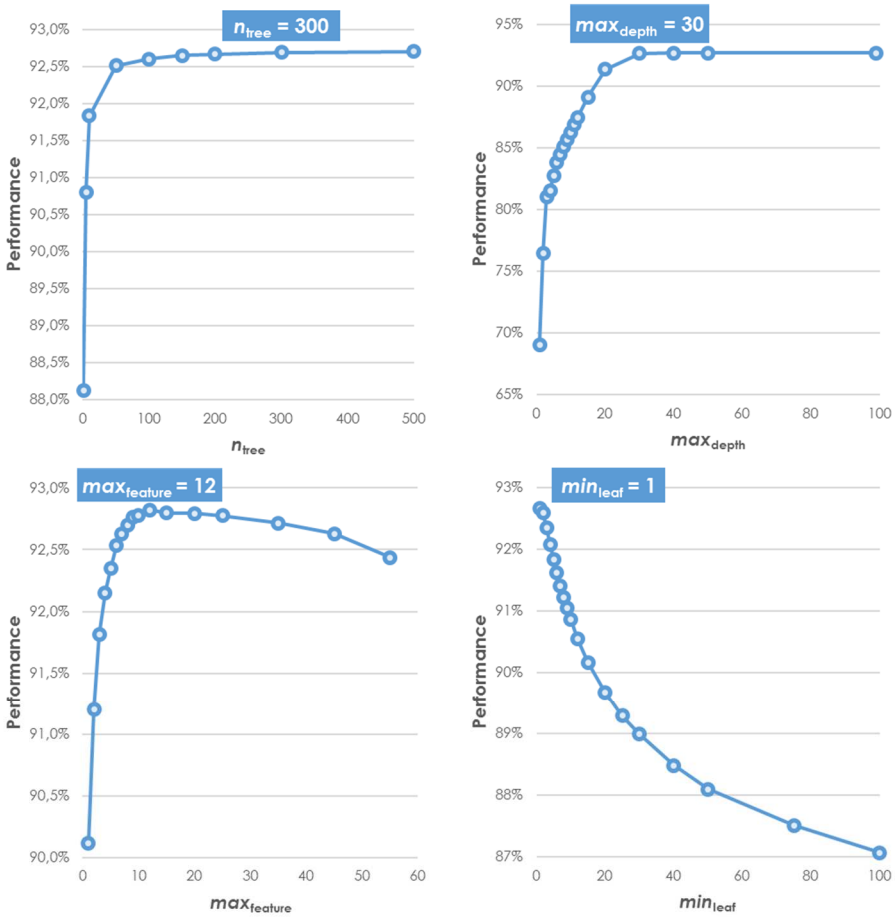


Figure 17 – Performance de l'algorithme random forest

[67] En toute rigueur, il aurait fallu rechercher le paramétrage qui maximise la performance de l'imputation au niveau global, c'est-à-dire sur l'espace en 4 dimensions des hyperparamètres. Pour des raisons de simplifications évidentes, nous nous sommes restreints à la recherche d'un maximum local en optimisant individuellement chacun des hyperparamètres.

De manière évidente, les deux premiers hyperparamètres n_{tree} et max_{depth} augmentent la performance de l'imputation au prix d'une complexité accrue. Nous choisirons donc pour ceux-ci la plus petite valeur au-delà de laquelle le gain en performance est minime. De manière symétrique, l'hyperparamètre min_{leaf} vient réduire la complexité de l'algorithme en diminuant sa performance. Mais contrairement à n_{tree} et max_{depth}, nous n'observons pas de plateau de performance qui pourrait permettre de gagner en efficacité sans perdre trop en performance. Finalement, seul l'hyperparamètre max_{feature} va présenter un phénomène de surapprentissage. En conclusion, le paramétrage de l'imputation des données manquantes par forets aléatoires est le suivant :

- $n_{\text{tree}} = 300$;
- $max_{\text{depth}} = 30$;
- $max_{\text{feature}} = 12$;
- $min_{\text{leaf}} = 1$.

II.1.3.3.4. Imputation par réseau de neurones

L'imputation par réseau de neurones est très complexe à paramétrer. En effet, il convient tout d'abord de sélectionner le type de réseau (propagation avant, rétro-propagation, élagage…), la fonction de seuil utilisée et surtout l'architecture des couches cachées[68]. Comme vu en annexe 7, nous avons décidé de nous restreindre aux réseaux de neurones à propagation avant (*perceptron*). Les hyperparamètres à optimiser sont ainsi :

- N_{layer} : nombre de couches cachées ;
- $k_i, i \in [\![1, N_{\text{layer}}]\!]$: nombre de neurones appartenant à la couche i ;
- f_s : fonction de seuil ;
- max_{iter} : nombre maximum d'itérations de l'algorithme d'optimisation (afin de stopper l'algorithme lorsque celui-ci ne converge pas).

Le nombre de combinaisons possibles d'un réseau de neurones étant infini, nous avons limité notre recherche d'optimisation a des architectures avec $N_{\text{layer}} \in \{1,2,3\}$ telles que :

- $N_{\text{layer}} = 1$: pas de contrainte ;
- $N_{\text{layer}} = 2 : \begin{cases} k_1 = k_2 \\ k_1 = 2k_2 \\ 2k_1 = k_2 \end{cases}$;
- $N_{\text{layer}} = 2 : k_1 = k_2 = k_3$

[68] Seules les couches cachées sont prises en compte dans le paramétrage du modèle puisque la couche d'entrée dépend du nombre de variables en entrée et de même pour la sortie. Il ne s'agit donc pas d'hyperparamètres à optimiser.

En termes de code, nous avons utilisés les fonctions *MLPRegressor* et *MLPClassifier* de la bibliothèque *Sklearn*[69]. Nous obtenons alors les résultats suivants :

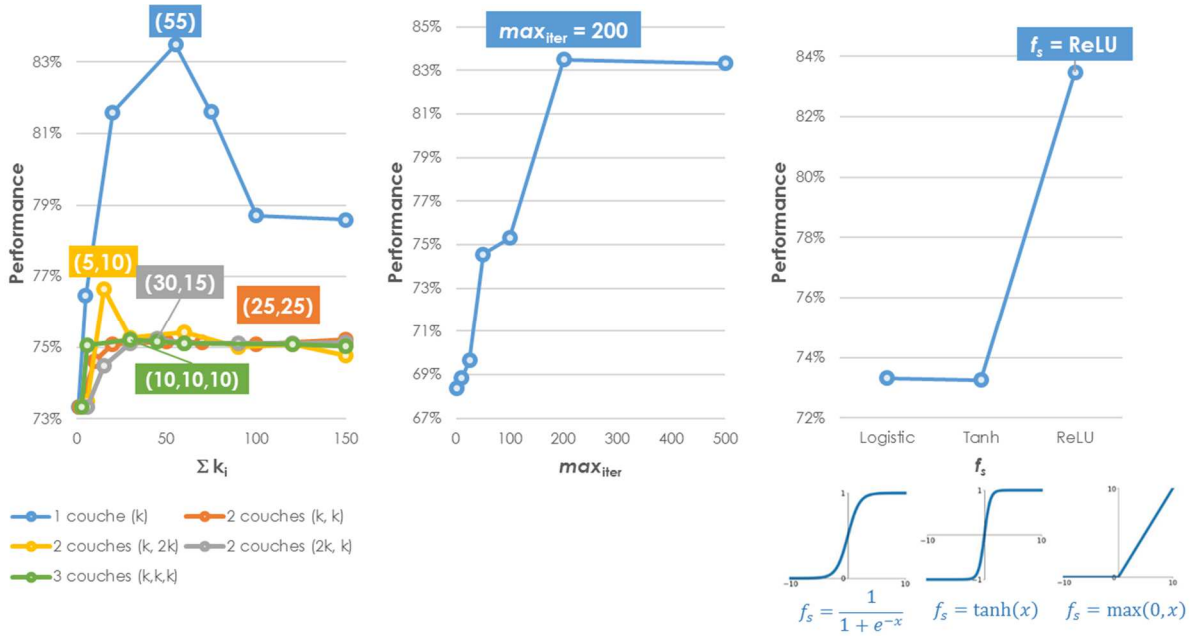


Figure 18 – Performance de l'algorithme réseau de neurones

Ces résultats[70] démontrent une nette supériorité[71] du réseau neurones à une seule couche cachée avec une fonction de seuil de type ReLU. L'hyperparamètre max_{iter} améliore logiquement la performance de l'imputation au détriment de la vitesse de l'algorithme. Finalement, nous adoptons le paramétrage suivant :

- $N_{\text{layer}} = 1$;
- $k_1 = 55$;
- $f_s = ReLU = x \longmapsto \max(0, x)$;
- $max_{\text{iter}} = 200$.

[69] A noter que la bibliothèque *Keras* propose des solutions de conception de réseau de neurones en Python beaucoup plus complètes et efficaces.

[70] Contrairement aux algorithmes précédents, les données ont été *scalées* en amont, c'est-à-dire que les variables quantitatives ont été centrées et réduites tandis que les variables qualitatives ont été ramenées à un codage entre 0 et $n_{\text{catégories}} - 1$. Ce traitement est nécessaire dans le cadre d'un réseau de neurones afin d'éviter de surpondérer des variables uniquement parce qu'elles contiennent de grandes valeurs. La technique du *scaling* est davantage développée au § II.3.1.2.

[71] Il est intéressant de noter que si en termes de performance des résultats, le réseau de neurones à une seule couche cachée est bien supérieur aux autres, il est également bien plus lent et a beaucoup plus de difficultés à converger que les réseaux avec davantage de couches de neurones.

II.1.3.4. Choix du modèle retenu

Comme expliqué en annexe 5, l'évaluation finale des performances de nos modèles d'imputations des données manquantes s'effectue sur une base de données différentes de celle utilisée pour le paramétrage : le mois de juin 2016.

Comme pour le § II.1.3.1., il est nécessaire de disposer d'une base de données complètes découpée entre une base d'entraînement (représentant 90% de la base des données complètes) et une base d'évaluation (représentant donc 10% de la base des données complètes) sur laquelle sont créées des valeurs manquantes artificielles selon le modèle construit en annexe 6. La mesure de la performance est logiquement identique à celle définie dans le § II.1.3.2.

Dans ces conditions, nous obtenons les résultats suivants :

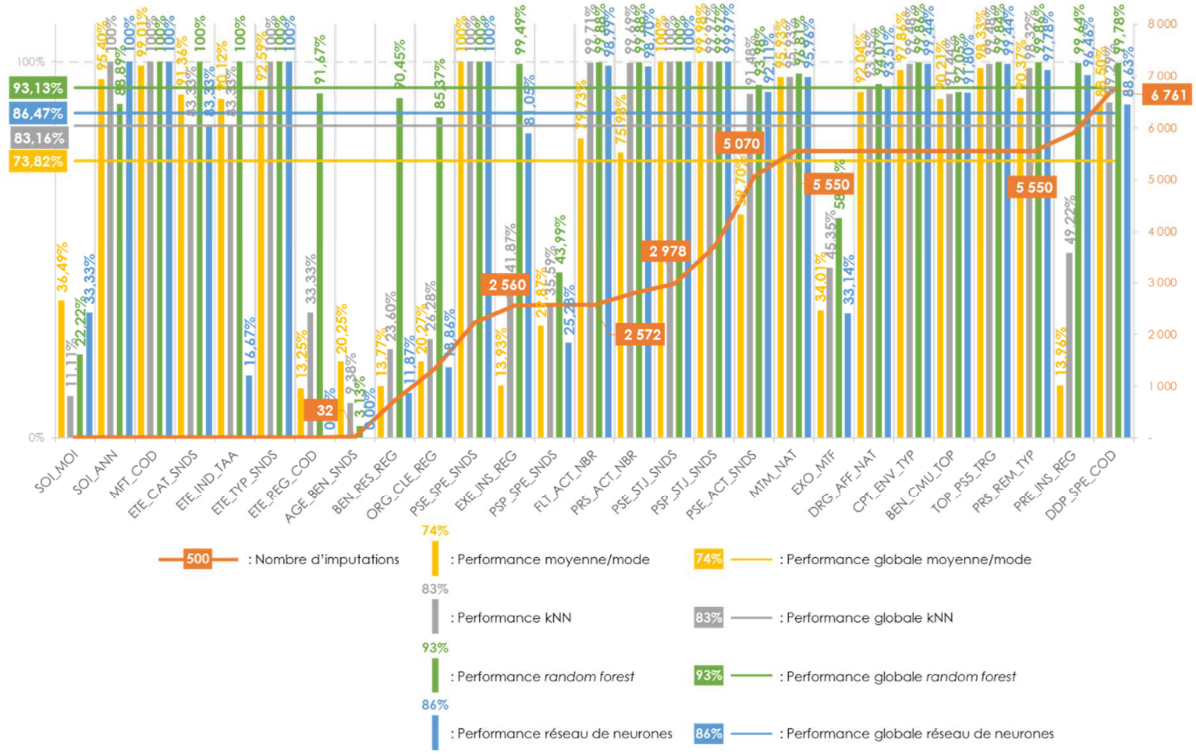


Figure 19 – Evaluation de la performance des quatre modèles d'imputation des valeurs manquantes

Le modèle des forêts aléatoires ressort comme l'algorithme comme le plus performant au global, suivi par le modèle de réseau de neurones. En regardant au niveau de chaque variable, le modèle des forêts aléatoires reste le plus performant (à l'exception des variables ***SOI_MOI***, ***SOI_ANN*** et ***AGE_BEN_SNDS***). En conclusion, nous choisissons le modèle des forêts aléatoires pour imputer les valeurs manquantes sur l'intégralité des variables[72] de la base Open DAMIR.

[72] Il aurait été possible d'adapter le modèle d'imputation à chaque variable ou encore jouer sur l'ordre d'imputation des variables au lieu de toutes les imputer simultanément avec le même modèle. Néanmoins, pour des raisons de temps et de simplicité, ces méthodes n'ont pas été explorées.

II.1.4. Mise-en-œuvre du modèle

A présent que notre modèle d'imputation des valeurs manquantes a été entièrement conçu, il reste à l'appliquer sur les données réelles. Sur les bases des 24 mois de données des années 2015 et 2016 et en appliquant les mêmes méthodes que précédemment, nous avons construit une base d'apprentissage de 1 159 550 lignes pour notre algorithme de forêts aléatoires.

Notre modèle d'imputation appliqué sur les 713 146 868 lignes des années 2015 et 2016 permet ainsi de retrouver les valeurs probables de 4,5 milliards de données manquantes (ce qui représente près de 11,5% des 39,2 milliards de données traitées). Il est malheureusement impossible de mesurer la qualité de ces imputations dès lors que les véritables valeurs ne sont pas connues. Cela étant dit, les excellents résultats constatés sur la base d'évaluation (cf. § II.1.3.4.) permettent de rester confiant quant à la performance de notre modèle d'imputation des valeurs manquantes.

Néanmoins, ces résultats doivent être relativisés car, lorsqu'est appliqué un modèle d'imputation de valeurs manquantes sur un grand volume, il existe un risque de créer artificiellement une structure au sein des données, ce qui peut remettre en question l'apport des traitements effectués pour le modèle final. En outre, certains algorithmes de *machine learning* utilisés pour prédire les décès dans le § II.3, en particulier celui des forêts aléatoires, fonctionnent bien en présence de valeurs manquantes et peuvent même se nourrir de cette information.

Par ailleurs, une autre limite de l'imputation par l'utilisation du *random forest* est que, même si l'algorithme affiche une haute précision, il perd une grande partie de l'information autour de la donnée imputée puisqu'une seule valeur est tirée par l'algorithme. Ainsi, si la donnée imputée peut prendre la valeur « A » ou « B » de manière plus ou moins équiprobable, le modèle dira que la valeur vaut « A » et va perdre l'information sur la probabilité que la valeur soit « B ». Etant donné que le modèle développé s'intéresse aux événements rares, c'est-à-dire aux queues de distribution, ce point pourrait être problématique. Heureusement, cette limite est compensée par le grand nombre de valeurs manquantes à imputer qui permet de reconstituer la distribution de probabilité perdue par l'algorithme d'imputation.

Pour ces raisons, il est indispensable de bien tester l'apport de l'imputation des données manquantes que nous avons réalisée ici au sein du modèle final (cf. § II.3.4.).

II.2. CONSTITUTION DE LA BASE DE DONNEES D'ETUDE

Il s'agit désormais de constituer la base qui va nous servir pour réaliser notre modèle de prédiction des décès. Pour cela, nous sélectionnons tout d'abord les données de l'assureur (cf. § II.2.1.) qui permettront de réaliser la prédiction de notre cible, dont la définition exacte est très liée à la façon dont les données sont structurées et qui sera donc précisée à cette occasion (cf. § II.2.3.). Puis, ces données seront d'appariées avec celles issues de l'*open data* de santé (cf. § II.2.2.) pour aboutir à une table unique qui servira d'entrée au modèle.

II.2.1. Récupération des données de l'assureur

II.2.1.1. Présentation du principe retenu

Par rapport à la cible, le principe est de retenir les informations sur l'assuré qui servent habituellement au calcul des provisions en santé et en prévoyance. De plus, l'objectif étant d'associer les données de l'assureur avec celles issues de l'*open data* de santé, nous avons également récupéré des informations sur les prestations santé versées par l'assureur. Enfin, de manière à pouvoir tester les résultats, les données récupérées dans les bases de l'assureur doivent contenir également les sinistres prévoyance.

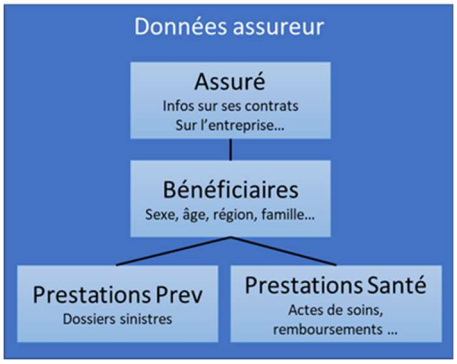


Figure 20 – Modèle simplifié des données de l'assureur qui seront utilisées

II.2.1.2. Réalisation du traitement de récupération

II.2.1.2.1. Choix du niveau de détail

Pour essayer d'avoir un bon pouvoir prédictif, nous nous plaçons dans un premier à la maille de l'individu[73], ce qui permettra ensuite d'ajouter des informations de sources externes plus facilement. En se plaçant à une maille plus large comme celle de l'entreprise ou du contrat, les données individuelles risquent d'être trop agrégées pour identifier les signaux faibles.

[73] L'analyse de la conformité RGPD du traitement est abordée dans le § II.2.1.3.

Nous nous sommes ensuite restreints aux individus assurés à la fois en santé et en prévoyance car les données disponibles permettent d'avoir sous cette contrainte un volume qui reste très important (218 mille individus).

Afin de conserver les informations portant sur les bénéficiaires[74] santé et pas uniquement sur les ouvrants droits[75], nous avons transformé les informations relatives aux bénéficiaires sous la forme d'une cellule familiale (situation familiale, nombre d'enfants, nombre d'hommes et de femmes dans le foyer), sur laquelle nous avons agrégé les montants de consommation de soin.

II.2.1.2.2. Choix de la fenêtre temporelle

La population assurée n'est évidemment pas un groupe fermé, et il nous faut une période temporelle fixe pour avoir des données comparables entre elles. Pour cette étude, nous nous sommes alors basés principalement sur les données d'un exercice N, en prenant l'hypothèse d'une étude qui serait réalisée au 01/01/N+1, et qui serait reproductible chaque année.

Ce choix impacte le traitement des données sur les différents points suivants :

- **Les périodes d'affiliation :** pour traiter le problème des périodes d'affiliations différentes suivants les assurés, et afin d'avoir des données comparables, nous avons choisi de récupérer le nombre de jours d'affiliation de l'ouvrant droit pendant l'exercice N. Nous gardons les personnes qui sont toujours affiliées en fin d'année N, mais le nombre de jours d'affiliation dans l'année pourra servir de pondération, notamment par rapport à leur consommation de soin. La population d'affiliation peut également changer en cours d'exercice : nous choisissons de retenir la dernière en date du 31/12/N.

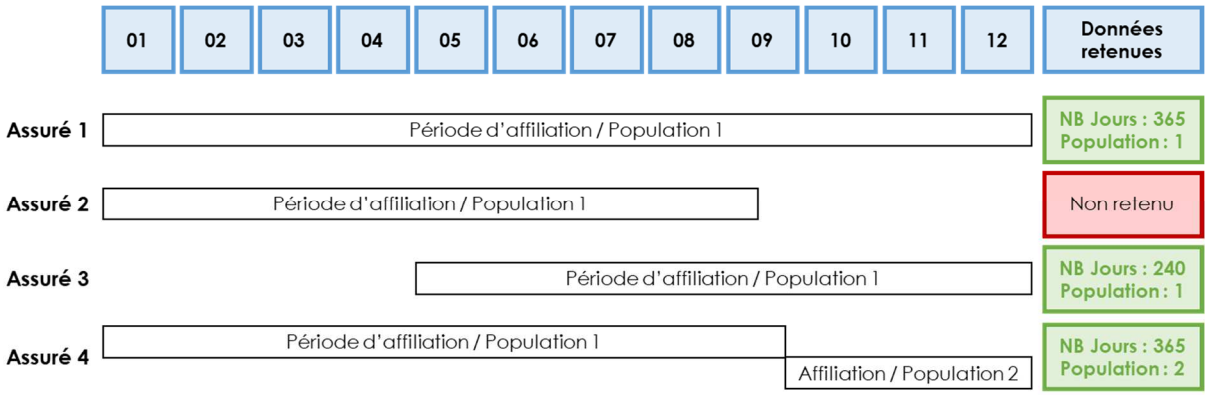


Figure 21 – Traitement des périodes d'affiliations dans la préparation des données

[74] Le bénéficiaire correspond à la personne qui bénéficie des prestations, qu'il s'agisse de l'ouvrant droit ou de ses ayants droits.
[75] L'ouvrant droit correspond au salarié couvert par un contrat collectif. Son affiliation lui permet d'être assuré et également d'« ouvrir les droits » à son conjoint, ses enfants, voire parfois ses ascendants.

- **Les informations complémentaires sur les ouvrants droits :** les différentes données ont été filtrées de manière à être représentatives de l'année N, notamment les données relatives à leur lieu de résidence, à leur emploi (Code NAF et CCN de l'entreprise). Plus exactement, c'est la situation connue au 31/12/N qui a été retenue, afin d'être cohérent avec le principe d'une étude au 01/01

- **Les prestations santé :** les prestations santé de l'exercice N ont été récupérées à vision fin exercice N, de manière à se placer à l'identique du cas pratique d'une étude faite au 01/01. Les données obtenues sont classées par famille d'actes, ce qui nous a permis de récupérer le détail du nombre d'actes, des montants de dépense réelle et du remboursement effectué. Nous avons choisi de mettre en évidence également le nombre de dépassement d'honoraire[76] et le montant correspondant pour chaque famille d'acte. De la même manière que pour les autres informations, nous nous plaçons dans la situation d'une étude qui aurait lieu au 01/01, date à laquelle nous aurions connaissance des événements comptables de l'année mais pas l'intégralité des dépenses liés aux actes de soin de cette même année[77].

- **Les prestations prévoyance :** la problématique est similaire aux prestations santé. Nous avons alors récupéré la présence de dossiers de prestations, ainsi que les montants de prestations par risque. Une autre information importante pour l'étude est le fait de savoir si le dossier est toujours en cours en fin d'année N. Enfin, ces prestations nous ont également fourni la cible à atteindre : la survenance d'un décès en exercice N+1. Les prestations en dessous de la franchise ne sont pas représentées ici car non disponibles. Cela serait une donnée intéressante à ajouter, en étudiant notamment les informations présentes dans la DSN[78], mais ce sujet n'a pas pu être traité dans le cadre de ce mémoire.

Dans l'absolu, au lieu de se limiter à la situation connue à une date donnée, il aurait été envisageable de récupérer l'évolution de cette situation tout au long de l'exercice et de travailler sur la série temporelle constituée par la consommation de soins des individus. Néanmoins, procéder ainsi aurait posé des problèmes au niveau de la conformité de l'étude (une telle accumulation d'information engendrerait des possibilités de recoupement et de réidentification). De plus, le lien avec les données Open DAMIR aurait été moins naturel, car il aurait été impossible de reconstituer des chronologies d'actes de soins cohérentes à cause des agrégations réalisées. Nous avons donc écarté cette solution.

L'avantage de l'étude annuelle, est qu'elle peut se rapprocher d'un processus opérationnel de calcul des provisions d'inventaire. De plus, en déclinant cette méthode sur plusieurs années, il est possible de mettre en évidence les disparités (ou leur absence) entre les facteurs de prédiction des différents exercices[79].

[76] Le dépassement d'honoraires correspond à la partie du montant facturé de l'acte de soin qui dépasse le tarif conventionné de la Sécurité Sociale pour cet acte.
[77] La notion d'acte de soin regroupant toutes les dépenses s'étalant dans le temps par rapport à cet acte, sur une période qui peut être longue suivant sa nature (par ex : radio puis kinésithérapie consécutive à une fracture).
[78] La Déclaration Sociale Nominative est constituée de fichiers normés mensuels adressés par les employeurs aux organismes et administrations concernées contenant les données nécessaires à la gestion de la protection sociale des salariés.
[79] Cependant, étendre la méthode à d'autres années nécessiterait de vérifier la stationnarité des données dans le temps. Le modèle à mettre en place dans le cas contraire serait complètement différent.

II.2.1.2.3. Cas des données historiques

Le contenu des données est identique pour les années N-1 et N-2 ainsi que pour les données de l'exercice étudié. Cependant, à ce stade et pour éviter de rentrer dans le cadre réglementé par le RGPD (cf. § I.3.2.), nous ne disposons plus des identifiants des personnes physiques (pseudonymisation), et ne pouvons enrichir nos données directement. Pour exploiter tout de même ces informations, nous choisissons de les agréger sur les champs qui nous serviront à l'associer à nos données de l'exercice étudié (agrégation).

Les variables qui permettent d'apparier les données historiques (agrégées) avec la base de l'exercice étudié (à la maille individu) sont choisies de manière à avoir suffisamment de volume sur chaque agrégation pour éviter les cas limites, tout en restant précises au niveau des informations jugées susceptibles d'avoir une importance dans l'étude de la cible. Ainsi, les données que nous avons identifiées sont :

- le sexe ;
- l'âge (par tranche) ;
- le régime[80] ;
- la localisation géographique (région et pays) ;
- les informations sur la cellule familiale (situation familiale, nombre d'enfants, nombre d'hommes et de femmes dans le foyer).

Sur ces modalités, les différentes données quantitatives sont agrégées par type d'acte :

- le nombres d'actes ;
- le montant de la dépense réelle ;
- le montant du remboursement de l'assurance maladie obligatoire ;
- le montant du dépassement d'honoraires ;
- le montant remboursé par l'assurance maladie complémentaire.

II.2.1.3. Traitement des exigences de conformité et de confidentialité

Afin de limiter le risque de réidentification, nous avons limité les données présentes dans l'extraction.

- **Données à caractère personnel :** le tableau 3 détaille le traitement de ces données.

Donnée à caractère personnel	Anonymisation réalisée
Nom	Suppression
Prénom	Suppression
NIR	Inaccessible

[80] Les données contiennent aussi bien l'information du grand régime (général, agricole, etc.) que celle du régime local (Alsace-Moselle) le cas échéant.

Donnée à caractère personnel	Anonymisation réalisée
Date de naissance	Transformation en âge
Adresse - Libellé de voie	Suppression
Adresse - Code postal	Conservation du département
Adresse -Nom commune	Suppression
Email	Suppression
Téléphone	Suppression
Profession	Suppression[81]
Numéro compte bancaire	Inaccessible
IBAN	Inaccessible
Clé interne inter-applicative	Suppression[82]

Tableau 3 – Anonymisation réalisée sur les différentes données à caractère personnel

- **Données de santé :** nous nous sommes limités à ce niveau à un nombre d'actes et à des montants de dépenses par famille d'acte. Les données à caractère personnel étant largement anonymisées à cette étape du traitement, ce niveau d'agrégation de l'information ne pose pas de problème vis à vis des contraintes réglementaires.

- **Données médicales :** aucune donnée médicale n'était présente dans les données extraites.

- **Données complémentaires :** nous avons limité les données supplémentaires avec lesquelles nous aurions pu enrichir notre table afin d'éviter des problématiques de risque de corrélation ou d'inférence (cf. § I.3.2.).

Les auteurs du présent mémoire ayant tous les deux des employeurs différents, la confidentialité des données internes pouvait poser problème. Ce problème a été résolu en utilisant les données d'un seul des deux employeurs et en faisant en sorte que seul le salarié de celui-ci puisse accéder aux données. Ces sujets ont été adressés en utilisant les processus internes Malakoff Humanis sur l'étude de risque et de conformité concernant toute utilisation des données :

- Remplissage d'une fiche décrivant le cas d'usage.

- Soumission de la fiche au relais DPO et échanges pour complétion.

- Retour du relai DPO suite à consultation des instances sur acceptation ou sur la définition des actions de sécurisation supplémentaires à mettre en place.

[81] L'information du code NAF de l'entreprise employeur et du collège de l'assuré est conservée.
[82] Nous n'avons plus de clé inter-applicative après récupération des données, c'est pourquoi nous avons dû réaliser une association par agrégation pour les données historiques (cf. § II.2.1.2.3.).

II.2.2. Appariement avec les données externes

II.2.2.1. Description de la problématique

Une des difficultés d'utilisation de données externes en *open data* est de pouvoir les associer avec les informations détenues par ailleurs ou les données internes. En effet, la façon dont ces données externes sont codifiées est la plupart du temps difficilement conciliable avec une codification interne (excepté pour les informations normées ou qui disposent d'une codification de référence, comme les codes postaux). Pour les données de la base Open DAMIR, c'est le cas notamment pour la nature de la prestation qui va nécessiter d'étudier une correspondance avec la nomenclature de l'assureur.

De plus, le fait que les données d'*open data* soient par nature anonymisées restreint de fait les possibilités d'association fine. Dans le cas présent, les données de la base Open DAMIR sont très agrégées : une ligne représente un grand nombre de personnes, ce qui fait que l'appariement ne pourra se faire qu'à un niveau plus large.

II.2.2.2. Réalisation

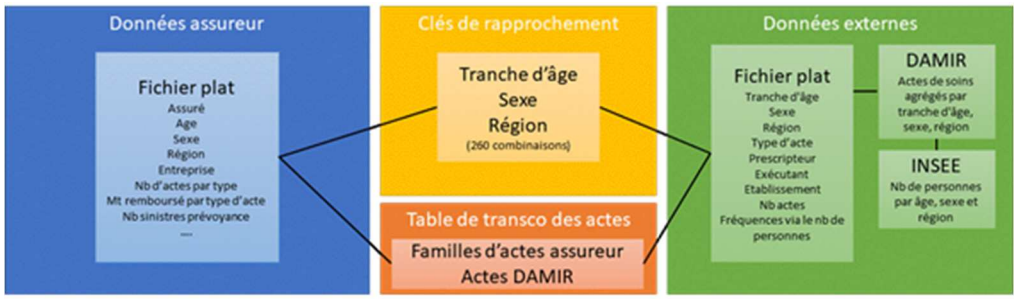


Figure 22 – Modèle de l'appariement des données provenant des différentes sources

Les variables utilisées pour réaliser l'appariement des données sont les suivantes :

- **âge :** il s'agit de faire correspondre à l'âge de l'assuré à la tranche d'âge DAMIR ;
- **région :** récupération simple de la région à partir d'un département ;
- **sexe :** simple transcodification (H = 1, F = 2) pour permettre la jointure ;
- **famille et nombre d'actes :** la correspondance était ici peu évidente, il a fallu reclasser tous les actes décrits dans la base DAMIR (plus de 800 libellés d'actes différents) dans les familles d'actes dont nous disposons côté assureur (14 familles d'actes). Par ailleurs, certains actes côté Open DAMIR ne sont pas à comptabiliser dans le nombre d'actes pour être cohérent avec les données en provenance de l'assureur (par exemple, le « complément spécialiste », qui correspond à un acte côté Open DAMIR et non côté assureur) : la table de transcodification construite pour réaliser l'association doit également contenir cette information qui permettrait de reconstituer le nombre d'actes.

Pour cela, nous sommes repartis de la table de correspondance qui avait été réalisée pour l'analyse de données manquantes afin de classer les natures d'actes de la base Open DAMIR dans les familles d'actes de l'assureur. Puis en face de chaque acte, selon sa description, nous avons ajouté s'il s'agissait effectivement d'un « acte » à compter dans le dénombrement ou non.

Cette méthode a permis de restituer dans une table l'ensemble des données assureur, à la maille de l'assuré, puis des données agrégées Open DAMIR correspondant à sa classe d'âge, son sexe et la région de résidence, en ayant des colonnes comparables par famille d'actes sur d'une part, les nombres d'actes, les montants de dépense et les dépassements de l'assuré, et d'autre part, sur les fréquences et les coûts moyens constatés sur cette population.

II.2.2.3. Feature engineering[83]

Après obtention d'une table contenant les données appariées, nous constatons qu'étant donné la méthode d'association des données internes et des données publiques, nous disposons de beaucoup d'information redondantes : sans compter les familles d'actes, il n'y a que 260 combinaisons différentes pour l'association des données, ce qui fait que les données ajoutées provenant d'Open DAMIR auront peu de pouvoir prédictif en l'état.

Nous avons donc défini d'autres variables afin d'intégrer ces données dans des variables plus individualisée, ainsi que pour se rapprocher, à notre sens, de ce qui pourrait être des variables significatives vis-à-vis de notre cible. En particulier :

- à partir des dépenses par famille d'acte de l'individu et des dépenses moyennes, nous avons mis en évidence l'écart par rapport à la moyenne en termes de consommation de soin de chaque individu[84] ;
- à partir des fréquences et des montants de dépassement d'honoraire, nous avons également ajouté l'écart à la moyenne sur le sujet des dépassements (fréquence et montant).

Afin d'obtenir les moyennes évoquées ici, nous avons utilisé les données de l'INSEE sur la même maille que celle de l'appariement de la base Open DAMIR (âge, région, sexe), afin de récupérer le nombre de personnes total correspondant aux actes présents dans cette base. A cette étape, nous réalisons également la transformation des variables qualitatives via une binarisation de chacune des modalités en autant de colonnes[85].

[83] Le terme *feature engineering* regroupe tout le processus de transformation des données d'entrées de manière à en extraire des propriétés (ou *features*) qui permettront, via leur aspect métier, d'améliorer les performances des modèles de *machine learning*

[84] Les données Open DAMIR et les données de santé détenues par l'assureur ne sont pas exactement sur le même plan : sur certaines garanties de l'assureur, l'assurance maladie obligatoire n'intervient pas et inversement, ce qui est pris en charge à 100% par la Sécurité Sociale n'est pas visible par l'assureur. Mais la mise en évidence de cet écart reste intéressante pour critère de comparaison sur toute la base.

[85] Cette méthode est appelée *One Hot Encoding*. Une colonne avec n modalités est transformée en n colonnes booléennes.

II.2.3. Précision des modalités de la cible de prédiction

En synthèse, la base de données qui servira d'entrée aux modèles de prédiction a été construite ainsi :

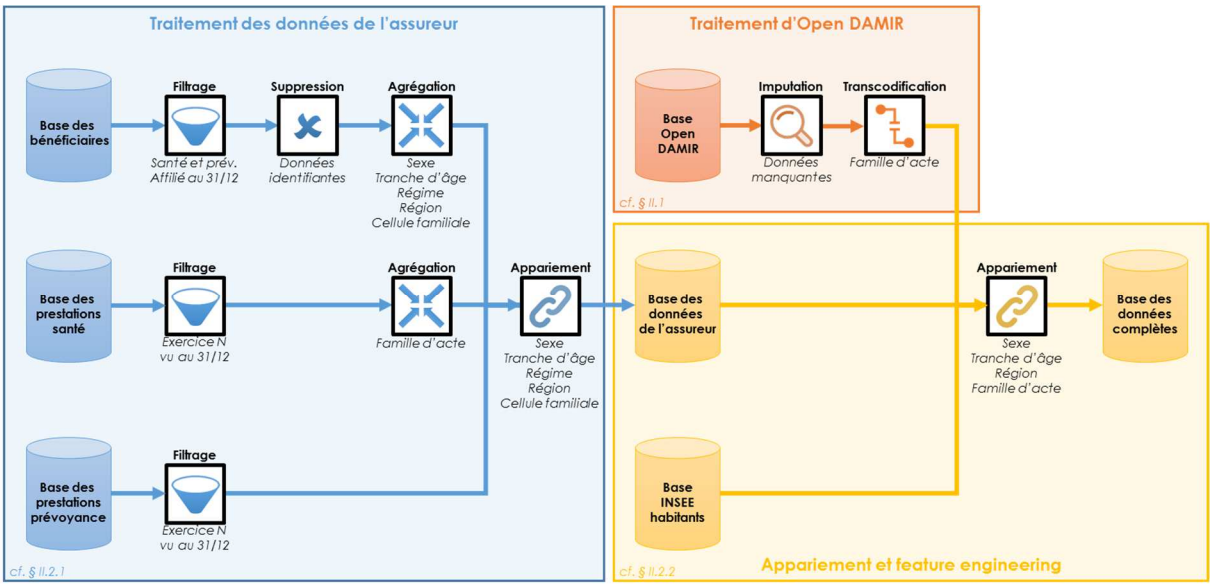


Figure 23 – Construction de la base d'étude

Tous ces choix au niveau de la récupération des données, de leur mise en forme et de la façon de les rapprocher, a des conséquences sur la prédiction.

Nous pouvons donc définir plus précisément notre cible en l'énonçant de la façon suivante :

- en se plaçant au 01/01 de l'année N,
- pour un assuré encore affilié à cette date,
- suivant les informations connues à cette date (sur lui mais aussi de portée plus générale),

anticipons-nous la survenance d'un décès au cours de l'exercice N ?

II.3. PREDICTION DES DECES

II.3.1. Première analyse de la base d'étude

II.3.1.1. Analyse préalable de données

Nous disposons maintenant de notre base d'étude, dont l'intégralité des données est détaillée en annexe 8. Après quelques vérifications d'usages sur le contenu de la base et la production d'indicateurs, présentés en annexe 9, nous décidons de conserver ces données en l'état.

En effet, de la même manière qu'abordé lors du traitement des valeurs aberrantes de la base DAMIR (cf. annexe 4), les données qui s'éloignent un peu trop de la moyenne pourraient constituer malgré tout autant d'informations qui pourraient permettre d'affiner les prédictions sachant qu'il est compliqué de les considérer comme erronées. C'est d'autant plus vrai dans un cas de prédiction d'un événement rare.

Il n'y a ainsi pas de problématique de censure sur les données de notre base d'étude, puisqu'essentiellement basé sur le périmètre des données de l'assureur. Par ailleurs, le phénomène de troncature est quasiment inexistant sur notre cible telle que définie (le cas des déclarations tardives étant traité par le fait d'étudier une année éloignée – 2016 – et la fraude étant plus que marginale en décès) et reste très limité sur l'ensemble de nos données.

II.3.1.2. Feature Scaling

Nous traitons au sein de notre base d'étude (ou « *dataset* ») des données numériques ayant des ordres de grandeurs très différents, à juste titre étant donné leur signification, mais cela peut perturber certains algorithmes de *machine learning*. Si cela affecte principalement leur performance, cela peut dans le pire des cas les empêcher de converger vers la solution optimale (celle-ci consistant à trouver, pour de nombreux algorithmes, le minimum d'une certaine fonction de coût). Dans des circonstances où la puissance de calcul est limitée, une étape de mise à l'échelle (*feature scaling* [scikit-learn.org]) paraît indispensable. Par ailleurs, cette étape améliore la stabilité[86] du modèle[87] et, selon la façon dont fonctionne l'algorithme utilisé, cela évite de donner plus de poids aux colonnes dont la valeur est plus importante.

[86] Stabilité par rapport à son caractère stochastique, c'est-à-dire que cela diminuera la variance des résultats obtenus lors de différentes exécutions du modèle.
[87] Par exemple, dans une descente de gradient, le coefficient d'apprentissage ne dépend pas de l'échelle des variables explicatives et pourrait donc faire de « grands pas » là où il faudrait en faire des « petits » (cf. annexe 7).

Nous faisons ici le choix d'une normalisation plutôt qu'une standardisation[88] car cette dernière suppose que les données suivent une loi normale. L'algorithme *MinMaxScaler* qui sera utilisé pour réaliser cette normalisation consiste, pour chaque colonne contenant des montants, à appliquer la transformation suivante :

$$x_i^{scaled} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Cela permet de ramener toutes les valeurs numériques entre 0 et 1, tout en conservant la forme de la distribution d'origine des différentes données. Toute l'information est conservée, ainsi que l'importance de valeurs aberrantes (contrairement à d'autres procédés de *feature scaling*).

II.3.2. Prédiction d'un événement rare

II.3.2.1. Rééchantillonnage des données

II.3.2.1.1. Principe

Comme l'événement à prédire est un événement rare[89], la distribution entre les deux classes (survie ou décès dans l'année) est très déséquilibrée. La classe recherchée est largement minoritaire, ce qui va impacter l'utilisation des modèles de *machine learning*. La plupart des algorithmes sont plutôt adaptés à des données dont la distribution est équilibrée et ont un meilleur pouvoir prédictif dans cette configuration. Une des méthodes pour pallier ce problème est de procéder à un rééchantillonnage, c'est-à-dire de sous-échantillonner la classe majoritaire et/ou de sur-échantillonner la classe minoritaire[90].

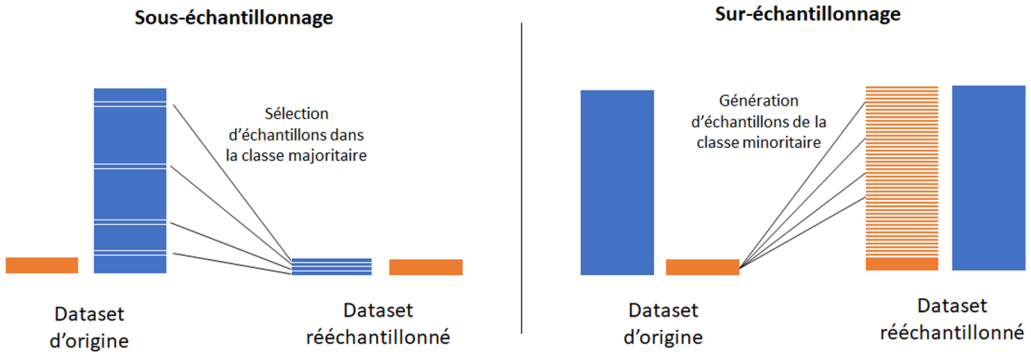


Figure 24 – Principe de fonctionnement du sous et sur-échantillonnage

[88] La standardisation consiste à appliquer une transformation de type $x_i^{scaled} = \frac{x_i - \bar{x}_i}{\sigma_{x_i}}$.

[89] Il n'y a pas de seuil précis à partir duquel il est considéré qu'un jeu de donnée est déséquilibré, mais la problématique est généralement abordée lorsque la classe minoritaire représente autour de 1% des données

[90] Il est généralement préférable de ne pas rééchantillonner les données de sorte à les altérer le moins possible. Néanmoins, ici, toutes les tentatives d'apprentissage directement sur les données d'origine se sont soldées par un échec. Dans ce cas précis, le rééchantillonnage a permis de résoudre le problème.

II.3.2.1.2. Notion de frontière de décision

L'objectif du modèle qui va suivre ce prétraitement est de trouver une fonction f entre la variable cible Y (qui vaut 0 pour survie ou 1 pour décès) et les n variables explicatives (les différentes données présentes dans la base d'étude).

Il est possible de représenter algébriquement cette problématique en considérant un espace à n dimensions dans lequel les différents individus sont positionnés et la fonction f tenterait de délimiter les deux espaces occupés respectivement par les individus de la classe 0 et ceux de la classe 1. La fonction f définie par un modèle correspond à ce qui se nomme sa frontière de décision. Il est plus simple de visualiser ceci dans une configuration où $n = 2$ (ou, comme ci-dessus, en prenant les deux premières composantes principales d'une ACP[91]).

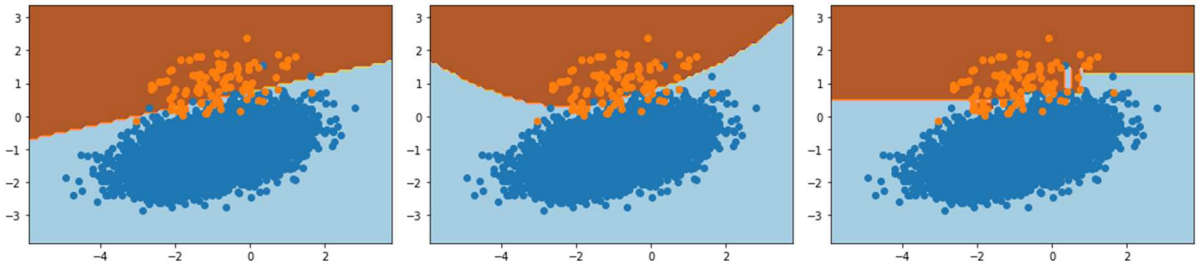
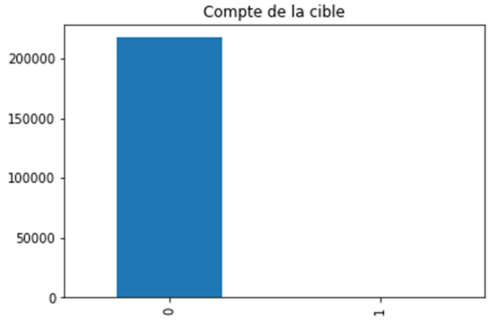


Figure 25 – Exemples de frontières de décisions (dans l'ordre : linéaire, quadratique et discontinue)

Les algorithmes de rééchantillonnage que nous allons évaluer sur nos données utilisent cette notion de frontière de façon à améliorer sa définition (parfois dégradée par l'aspect déséquilibré des données) pour tenter d'améliorer la performance des modèles.

II.3.2.2. Choix de la métrique

En présence d'un déséquilibre dans le *dataset*, il faut adapter la métrique qui afin d'évaluer la pertinence du modèle. Dans notre cas, nous obtenons la répartition suivante :



Classe 0 (Survie) : 217393

Classe 1 (Décès) : 324

Proportion Classe 0 / Classe 1 : 670.97 : 1

Figure 26 – Compte d'échantillon de chaque classe

[91] L'Analyse en Composantes Principales consiste à transformer des variables corrélées entre elles en nouvelles variables décorrélées les unes des autres. Elle permet de réduire la redondance et le nombre de variables en ne conservant que les n axes principaux (ici $n = 2$ pour la représentation graphique)

II.3.2.2.1. Cas d'une prédiction directe

Lorsque nous tentons de prédire directement le résultat de notre cible (sur le choix binaire 0 pour survie et 1 pour décès), nous pouvons nous appuyer sur une matrice de confusion, montrant les prédictions correctes et les différents types de prédictions incorrectes. Elle permet de visualiser rapidement les Vrais Positifs (v_p), Faux Positifs (f_p), Vrais Négatifs (v_n) et Faux Négatifs (f_n).

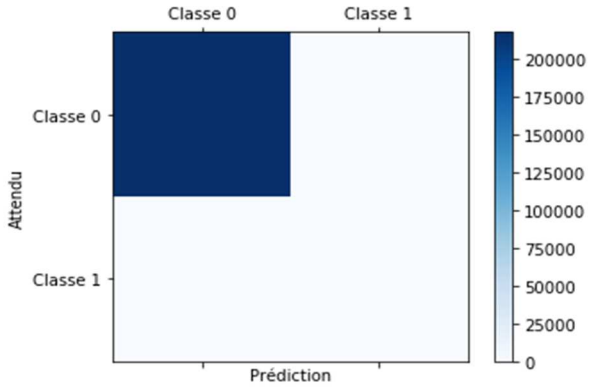


Figure 27 – Matrice de confusion pour le modèle qui prédit toujours « 0 »

Le déséquilibre de nos données la rend peu pertinente puisqu'une lecture rapide de celle-ci pour un modèle qui prédit toujours la survie pourrait nous amener à une mauvaise conclusion. Cet élément va être important pour le choix d'une métrique qui sera adaptée à notre cas.

Métrique	Formule	Signification
Accuracy	$\frac{v_p + v_n}{v_p + v_n + f_p + f_n}$	Il s'agit du taux de prédictions justes. En calculant le score d'un modèle qui prédit toujours 0, nous obtenons alors un score satisfaisant : $Accuracy = 99.85\%$ alors que le modèle n'a évidemment aucun intérêt. Cette mesure est donc peu adaptée à notre cas du fait du déséquilibre de nos données ($p \ll n$).
Précision (*Positive Predictive Value*)	$\frac{v_p}{v_p + f_p}$	Il s'agit du nombre de vrais positifs divisé par le nombre de prévisions positives. Cela permet de mesurer l'exactitude du modèle. Une faible précision implique un nombre élevé de faux positifs. Avec le même modèle que ci-dessus, nous trouvons Précision = 0.
Sensitivité (*Recall*)	$\frac{v_p}{v_p + f_n}$	Il s'agit du taux de vrais positifs (nombre de vrais positifs divisé par le nombre de valeurs positives dans l'échantillon de test). Elle permet de mesurer la complétude[92] du modèle. Une sensitivité basse signifie un nombre élevé de faux négatifs. Ici aussi, cette mesure vaut 0 pour le modèle qui prédit toujours « 0 ».

[92] « Complétude » dans le sens « Ai-je bien identifié tous les résultats positifs ? ».

Métrique	Formule	Signification
F1 Score	$2\frac{\text{Précision} \times \text{Sensitivité}}{\text{Précision} + \text{Sensitivité}}$	La moyenne harmonique de la précision et de la sensitivité. Il est possible également d'utiliser un F_β Score, avec β facteur de pondération entre la précision et la sensitivité, pour lequel l'utilisateur choisira suivant ce qu'il souhaite privilégier entre ces deux mesures. (non défini pour le modèle qui donne toujours « 0 »)
Spécificité	$\frac{v_n}{v_n + f_p}$	Il s'agit du taux de vrais négatifs (nombre de vrais négatifs divisé par le nombre de valeurs négatives dans l'échantillon de test). Elle mesure la capacité du modèle à éviter les faux positifs. Elle attribue le score de 100% au modèle qui donne toujours « 0 ».
Moyenne géométrique (*G-mean*)	$\sqrt{\text{Sensitivité} \times \text{Spécificité}}$	C'est la racine du produit des sensitivités calculées pour chacune des classes. Le but est de maximiser de façon équilibrée la précision sur les deux classes. Lorsqu'il s'agit de classification binaire, comme c'est le cas ici, il s'agit du produit de la sensitivité et de la spécificité (la spécificité pour une des classes correspondant à la sensitivité vis-à-vis de l'autre classe). Le résultat atteint par le modèle qui prédit toujours « 0 » est 0.

Tableau 4 – Définition de différentes métriques et commentaire sur leur cas d'utilisation

Il n'y a évidemment pas de mesure qui serait plus significative que les autres et qui serait à retenir dans l'absolu. Comme nous le voyons dans le tableau 4, il convient de choisir en fonction des objectifs recherchés : faut-il chercher à maximiser les vrais positifs (précision) ? à minimiser les faux positifs (spécificité) ? à trouver un bon compromis entre les deux (moyenne géométrique, F-Score) ? Au-delà de l'optimisation du modèle, il faut en tout cas avoir conscience de la réalité de son pouvoir prédictif, de ses forces et de ses limites.

II.3.2.2.2. Cas de la prédiction d'une probabilité

Dans notre cas, plutôt que prédire directement un résultat, il semble plutôt intéressant d'analyser la probabilité d'appartenance à la classe 1 (décès). En effet, par la nature même de notre cible, nous savons qu'elle comporte une part d'aléa qui ne sera pas présente dans le *dataset*. Viser une prédiction directe qui permettrait d'ajuster le provisionnement de ces sinistres paraît illusoire, voire une erreur de méthodologie [HARRELL, F. E. (2017)].

Les métriques vues précédemment s'appliquent une fois un seuil donné, qui, pour un problème de classification binaire, correspond par défaut à la condition :

$$\mathbb{P}[Y = 1|\boldsymbol{X}] > \mathbb{P}[Y = 0|\boldsymbol{X}] \Leftrightarrow \mathbb{P}[Y = 1|\boldsymbol{X}] > 0{,}5$$

Appliquer ce seuil (ou tout autre seuil permettant de prédire un résultat plutôt qu'une probabilité) revient à faire l'approximation suivante :

$$\mathbb{P}[Y = 1|\boldsymbol{X}] > 0{,}5 \Rightarrow Y = 1$$

Au moment où cette opération est réalisée, nous sortons donc du domaine des statistiques pour entrer dans celui de la prise de décision (selon la problématique traitée, les conséquences d'un faux positif ou d'un faux négatif, etc.).

Certaines mesures permettent de conserver la prédiction sous la forme d'une probabilité tout en illustrant ce phénomène de seuil, en particulier les courbes ROC et PR[93].

Courbe ROC

La courbe ROC (*Receiver Operating Characteristic*) prend le taux de faux positifs en abscisse et le taux de vrais positifs en ordonnée. Il est possible de l'interpréter comme une courbe permettant un arbitrage sur la valeur du seuil : pour tel pourcentage de vrais positifs (y), combien de faux positifs (x) seraient obtenus.

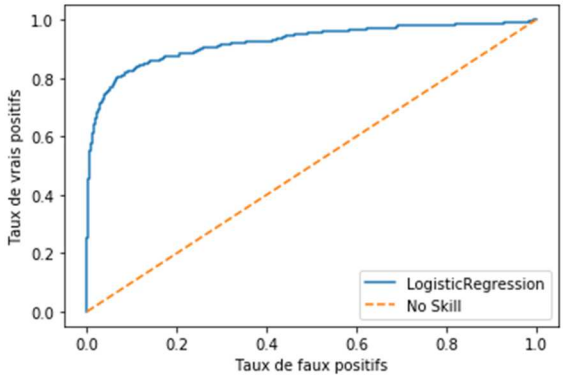


Figure 28 – Exemple de courbe ROC[94]

Courbe PR

Il s'agit du même principe que la courbe ROC, mais en faisant le graphe entre la sensitivité (ou « *recall* ») en abscisses et la précision en ordonnées.

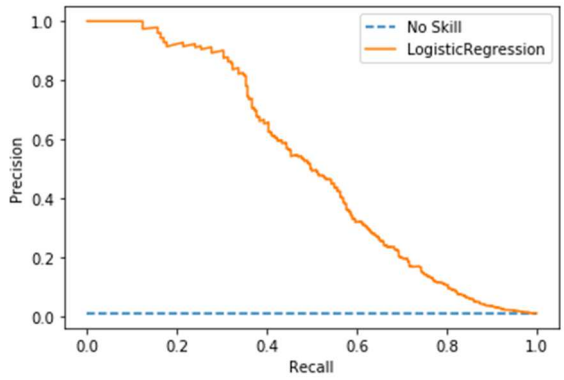


Figure 29 – Exemple de courbe PR[95]

Dans cette courbe, un modèle parfait se traduit par un point aux coordonnées (1,1). Au contraire, un modèle sans aucun pouvoir prédictif va être représenté par une ligne

[93] D'autres méthodes existent pour mesurer la performance du modèle, comme la courbe de *lift* ou une courbe de calibration, mais elles ne permettent pas d'élaborer facilement un score.
[94] Courbe obtenue par une régression logistique appliquée sur les données exemples des méthodes d'échantillonnage. La régression logistique est abordée en annexe 7.
[95] Courbe obtenue par une régression logistique appliquée sur les données exemples des méthodes d'échantillonnage

horizontale à une ordonnée correspondant à la proportion de cas de la classe minoritaire (très proche de 0 ici). Elle est particulièrement adaptée dans notre cas puisqu'elle focalise le calcul de la performance sur la classe minoritaire.

Scores ROC AUC[96] et PR AUC

Afin de classer nos modèles, il faut encore pouvoir déterminer à partir de ces courbes lesquelles correspondent aux meilleurs modèles. Un moyen simple est de se ramener au calcul de l'aire sous la courbe (métrique AUC). Le score ROC AUC d'un modèle s'étend entre 0,5 (aucun pouvoir prédictif) et 1 (prévision parfaite).

Cela dit, lorsque le déséquilibre entre les deux classes est très prononcé, cette mesure peut être trompeuse, dans le sens ou quelques bonnes prédictions peuvent augmenter drastiquement le score et ce, malgré de nombreux faux positifs. A contrario, le score PR AUC donne une mesure intéressante pour évaluer nos modèles, l'objectif étant de se rapprocher le plus de 1.

II.3.2.3. Choix du modèle

Nous avons sélectionné plusieurs modèles[97] de *machine learning* qui seront à même de classer nos individus dans les deux catégories de notre cible (survie ou décès) et de restituer une prédiction sous la forme d'une probabilité[98]. Parmi ceux-ci :

- l'algorithme des k plus proches voisins ;
- les réseaux de neurones ;
- les arbres de décisions et les forêts aléatoires ;
- la régression logistique ;
- les machines à vecteurs de supports ;
- l'analyse discriminante ;
- le boosting.

Etant donné le volume important de données ainsi que la combinatoire des méthodes de rééchantillonnage et des modèles à évaluer, chacun disposant de leurs hyperparamètres, il est nécessaire de définir une méthode qui permettra d'alléger les temps de calcul en procédant à des éliminations par étapes.

Ainsi, nous choisissons d'établir dans un premier temps une sélection de modèles qui semblent bien se comporter avec les données sources. Puis, nous évaluons les différentes méthodes de rééchantillonnage avec ces mêmes modèles afin d'en sélectionner une.
Enfin, sur la base des données rééchantillonnées par la méthode retenue, nous finalisons la recherche du meilleur modèle et de ses hyperparamètres.

[96] *Area Under the Curve*
[97] Une description de ces modèles est présente en annexe 7.
[98] Parmi les possibilités non développées ici, un modèle additif généralisé (GAM) semblerait particulièrement adapté à la problématique traitée mais les étapes de sa mise en place pourraient faire l'objet d'un mémoire à part entière.

La métrique utilisée pour déterminer le score des modèles est ajustée à chaque étape en fonction des traitements appliqués sur les données : lorsque le jeu d'entrainement a subi un rééchantillonnage, le déséquilibre entre les deux classes est réduit, ce qui permet d'utiliser le score ROC AUC.

Les trois étapes se décrivent de cette façon :

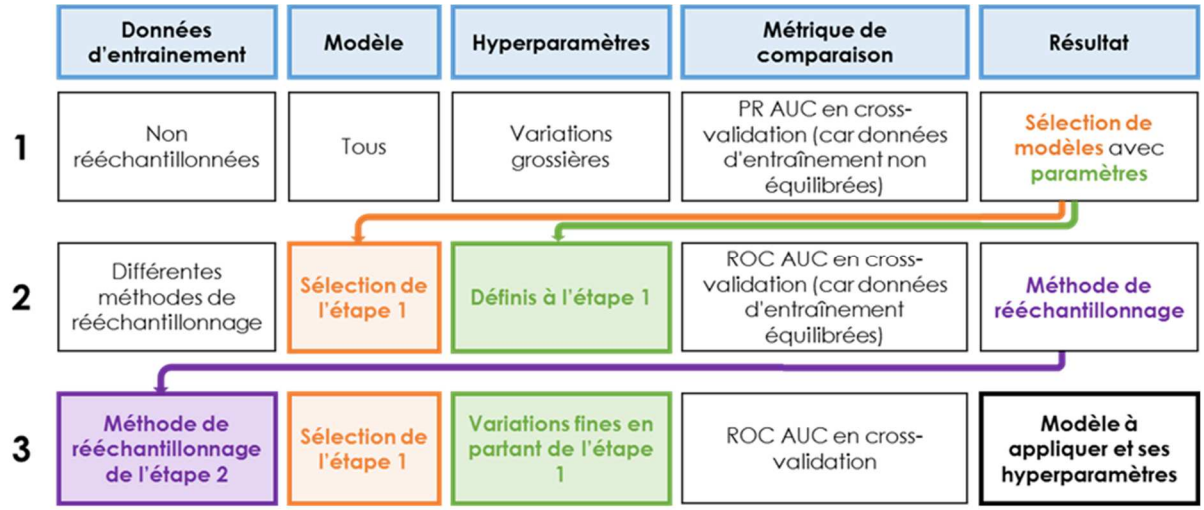


Figure 30 – Etapes de sélection du modèle à appliquer

Ces étapes se dérouleront en utilisant deux tiers des données disponibles comme *dataset* d'entrainement. Le tiers restant servira à éprouver le modèle et la pertinence des résultats. Les résultats de ces étapes et leur interprétation sont détaillés dans la partie suivante.

II.3.3. Exploitation des données

II.3.3.1. Résultats de la sélection du modèle

II.3.3.1.1. Etape 1

Lors de cette première étape, l'objectif est de réaliser une sous sélection de modèles qui semblent les plus adaptés à notre problématique. En plus du résultat du modèle dans la métrique choisie, nous regardons également l'écart type de cette mesure obtenue lors du lancement d'une validation croisée, (cf. annexe 5) puisque la stabilité du résultat va être importante au regard du caractère stochastique de ces algorithmes.

Modèle	Moyenne du score (PR AUC)	Ecart type du score	Décision
Random forest	0,157	0,005	✓
Régression logistique	0,148	0,016	✓
Analyse discriminante	0,138	0,030	✓
Arbre de décision	0,116	0,029	✗
Gradient boosting	0,097	0,010	✓
Réseaux de neurones	0,005	0,003	✗
KNN	0,003	0,002	✗
Machine à vecteurs de support	*Non mesuré*	*Non mesuré*	✗

Tableau 5 – Résultat de l'étape 1 et liste des modèles retenus

Ces résultats du tableau 5 montrent les points suivants :

- L'algorithme des machines à vecteur de support n'a pas pu faire l'objet d'une mesure, ne convergeant pas après 24h de traitement. C'est en effet un des inconvénients majeurs déjà évoqué et qui s'est avéré bloquant.
- Nous écartons le modèle des arbres de décision malgré des résultats satisfaisants, étant donné que l'algorithme *random forest* produit un bien meilleur score de façon plus stable, tout en étant basé sur le même principe intrinsèque.
- Nous choisissons au contraire de conserver le *gradient boosting* puisque c'est un algorithme qui est susceptible de donner de bien meilleurs résultats après rééchantillonnage du *dataset* d'entrainement

II.3.3.1.2. Etape 2

Nous testons au cours de cette étape tout une série d'algorithmes de rééchantillonnage et nous mesurons leur efficacité à travers les résultats du modèle de *machine learning* qui suit.

Les chaines de traitements évaluées sont donc composées de trois algorithmes successifs :

1. Sur-échantillonnage
2. Sous-échantillonnage
3. Modèle de *machine learning*

Le modèle est entrainé via cette chaîne de traitement (et donc sur des données rééchantillonnées) mais le score est ensuite calculé sur les données d'entrainement d'origine, sans quoi les scores ne seraient pas comparables car partant de bases différentes.

Nous constatons dans un premier temps, via le tableau 6, que sur la totalité des chaines de traitements testées, il semble plus efficace de sur-échantillonner la classe minoritaire avant de sous-échantillonner la classe majoritaire.

Ordre des traitements	Moyenne du score (ROC AUC)	Ecart type du score	Décision
Sur-échantillonnage suivi d'un sous-échantillonnage	0,67	0,024	✔
Sous-échantillonnage suivi d'un sur-échantillonnage	0,64	0,028	✘

Tableau 6 – Résultat de l'étape 2 suivant l'ordre des algorithmes de rééchantillonnage

Au global sur les différents rééchantillonnages testés, les différents modèles de *machine learning* sélectionnés à l'étape 1 donnent alors un classement, visible dans le tableau 7, légèrement différent de ce que nous avions vu précédemment puisque le modèle *random forest* reste le plus adapté juste devant la méthode *gradient boosting* qui semble effectivement plus performante après rééchantillonnage des données.

Modèle	Moyenne du score (ROC AUC)	Ecart type du score	Décision
Random forest	0,71	0,020	✔
Gradient boosting	0,70	0,031	✘
Régression logistique	0,64	0,030	✘
Analyse discriminante	0,55	0,018	✘

Tableau 7 – Résultat de l'étape 2 par modèle utilisé

Au final, en testant les différentes combinaisons des algorithmes de rééchantillonnages présentés en annexe 10, les meilleurs résultats sont présentés dans le tableau 8.

Sur-échantillonnage	Sous-échantillonnage	Moyenne du score (ROC AUC)	Ecart type du score	Décision
BorderlineSMOTE	***TomekLinks***	0,81	0,028	✓
SVMSMOTE	***EditedNN***	0,79	0,035	✗
SVMSMOTE	***TomekLinks***	0,78	0,030	✗
BorderlineSMOTE	***EditedNN***	0,78	0,042	✗
BorderlineSMOTE	***OneSidedSelection***	0,77	0,039	✗

Tableau 8 – Résultat de l'étape 2 par algorithme de rééchantillonnage

II.3.3.1.3. Etape 3

Suite aux résultats obtenus, nous avons maintenant défini la chaine de traitements, composée des algorithmes de rééchantillonnage puis du modèle de *machine learning* qui va réaliser la prédiction.

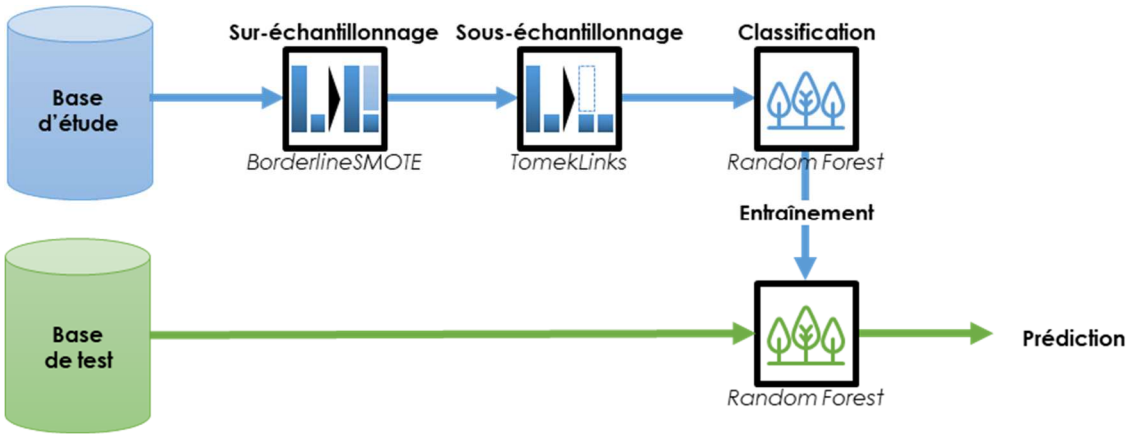


Figure 31 – Modèle de prédiction construit

L'objectif est maintenant de faire varier les hyperparamètres de ces algorithmes afin d'en fixer la valeur optimale pour l'entrainement de notre modèle. Nous appliquons ici la même méthode que celle détaillée au § II.1.3. D'un point de vue programmation, nous avons utilité la fonctionnalité *Pipeline* de la bibliothèque *Sklearn* afin de regrouper les trois modèles dans une seule chaîne de traitement :

- les algorithmes de rééchantillonnage *BorderlineSMOTE* et *TomekLinks* venant de la bibliothèque *Imblearn*, spécialisée dans les traitements de rééchantillonnage ;

- le modèle prédictif utilisé est *RandomForestClassifier*, de la bibliothèque *Sklearn*.

La fonctionnalité *GridSearchCV* permet de réaliser une validation croisée via *RepeatedStratifiedKFold*, tout en faisant varier les hyperparamètres sur toute la chaine de traitement :

- *BorderlineSMOTE* : le nombre k de la recherche de k plus proches voisins ;
- *TomekLinks* : aucun hyperparamètre ;
- *RandomForestClassifier* : les hyperparamètres décrits au § II.1.3.3.3.

GridSearchCV exécute toutes les combinaisons possibles d'hyperparamètres, classe les résultats puis permet de finaliser l'entrainement du modèle ayant obtenu les meilleurs résultats, ce qui donne l'optimisation présentée en tableau 9.

Borderline SMOTE	*RandomForest Classifier*
$k = 3$	$n_{\text{tree}} = 500$ $max_{\text{depth}} = 10$ $max_{\text{feature}} = 21$ $min_{\text{leaf}} = 1$

Tableau 9 – Hyperparamètres optimaux en résultat de la validation croisée de l'étape 3

II.3.3.2. Entrainement du modèle sélectionné

Avec les hyperparamètres déterminés à l'étape 3, nous réalisons finalement l'entrainement du modèle. Sur les données d'entrainement, nous obtenons des résultats qui semblent probants :

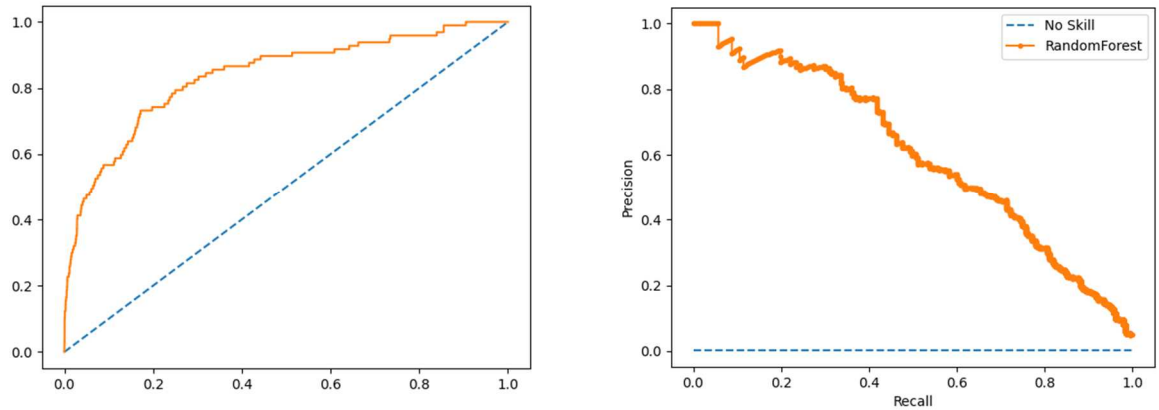


Figure 32 – Courbe ROC (à gauche) et PR (à droite) : ROC AUC = 0,86, PR AUC = 0,61

Nous avons tenté d'éviter le surapprentissage via la validation croisée, mais à cette étape, le bon comportement de l'algorithme sur les données de test est encore très incertain.

II.3.3.3. Réalisation de la prédiction

Le modèle ainsi entraîné sert donc ensuite à réaliser la prédiction sur les données de test et produit les résultats suivants :

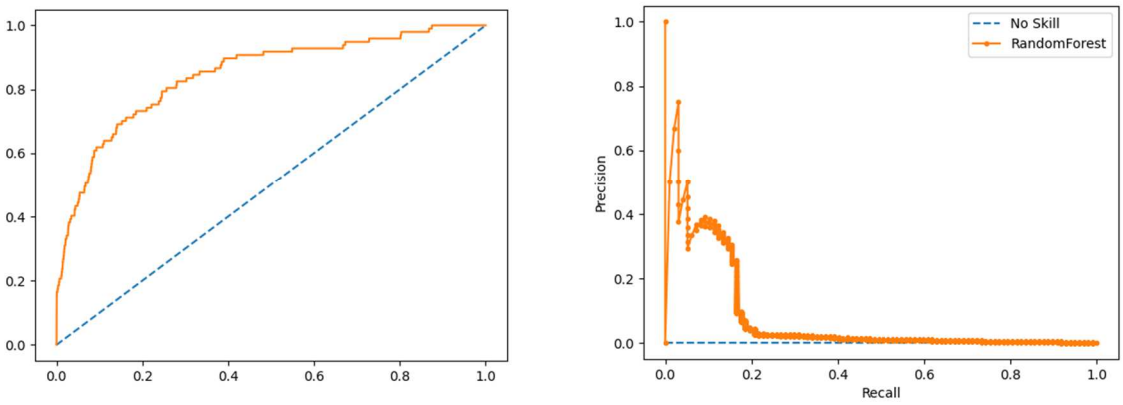


Figure 33 – Courbe ROC (à gauche) et PR (à droite) : ROC AUC = 0,845, PR AUC = 0,074

Le score PR semble décevant au premier abord, mais en regardant de plus près les prédictions réalisées sous la forme de probabilité d'appartenance aux deux classes (survie ou décès), il s'avère que le modèle attribue en moyenne une probabilité de décès plus de vingt fois supérieure aux individus appartenant effectivement à la classe décès (cf. tableau 10).

Résultat réel	Probabilité de décès moyenne prédite par le modèle
Survie	**0,21%**
Décès	**4,70%**

Tableau 10 – Probabilité de décès moyenne par classe d'individus prédite par le modèle retenu

Cela signifie que les données sur la consommation de soin dont nous disposons, qualifiées et retravaillées, apportent bien de l'information pertinente mais ne suffisent pas à réaliser une prédiction précise des décès (qui peuvent être liés à d'autres causes que la santé et qui demeurent des événements intrinsèquement aléatoires).

II.3.4. Exploitation des résultats

En l'absence de certaines données, notamment des niveaux de garanties ainsi que des salaires (dont la présence dans le jeu de données d'étude aurait complexifié la conformité au RGPD du traitement), nous ne sommes pas en mesure de comparer le provisionnement théorique induit par notre modèle afin de le comparer au provisionnement effectué par l'assureur dans ces comptes ou au montant de sinistres réellement survenus. En revanche,

il est possible de comparer les résultats de notre modèle avec les méthodes conventionnelles d'évaluation des probabilités de décès dans l'année.

Pour effectuer cette comparaison, nous utilisons :

- les décès réels qui ont été finalement constatés sur l'année de notre cible ;
- le même modèle mais avec l'ajout des données Open DAMIR d'origine, avant imputation des données manquantes;
- le même modèle mais sans l'ajout des données Open DAMIR (afin de mesurer l'apport d'information lié à l'*open data* de santé) ;
- la méthode de provisionnement classique, utilisant les tables réglementaires en cas de décès TH/TF 00-02.

Classe d'individus	Résultat réel	Probabilité de décès moyenne prédites			
		Modèle complet *(avec Open DAMIR et imputation)*	Modèle sans imputation *(avec Open DAMIR sans imputation)*	Modèle sans Open DAMIR	Tables de mortalité
Survie	**0%**	**0,21%**	**0,24%**	**0,79%**	**0,27%**
Décès	**100%**	**4,70%**	**4,68%**	**9,07%**	**0,52%**
Toute classe	**0,15%**	**0,22%**	**0,24%**	**0,80%**	**0,27%**

Tableau 11 – Comparatif des prédictions des différents modèles par rapport aux résultats réels en probabilités moyenne par classe

Le tableau 11 montre que le modèle développé présente d'excellents résultats par rapport aux autres méthodes. Tout d'abord, il est le plus proche de la probabilité de décès globale constatée sur les données réelles (+7 points contre +12 points avec la table réglementaire), ce qui est le plus important pour son usage en tant que méthode de provisionnement.

Par ailleurs, ces résultats montrent aussi que le modèle mis en œuvre est également très performant pour identifier la population au sein du portefeuille de bénéficiaire qui présente le plus grand risque de décès. En effet, en regardant le rapport entre les probabilités moyenne de décès et celles de survie (qui représente la puissance de classification de l'algorithme), nous obtenons les résultats du tableau 12.

Modèle complet *(avec Open DAMIR et imputation)*	Modèle sans imputation *(avec Open DAMIR sans imputation)*	Modèle sans Open DAMIR	Tables de mortalité
22,4	**19,5**	**11,5**	**1,9**

Tableau 12 – Comparatif du pouvoir de classification des différents modèles

Ce résultat est très intéressant car il prouve que le modèle développé a le potentiel de segmenter la population des assurés d'une manière beaucoup plus fine que la maille d'agrégation utilisée (sexe, tranche d'âge, région, régime et cellule familiale). Ainsi, le modèle développé pourrait en théorie permettre d'augmenter la prime des individus qu'il

identifie comme à risque. Or, il s'agit exactement de ce qu'il ne faut pas faire d'après le cadre éthique et réglementaire que nous avons défini dans la partie I : une telle décision pourrait en effet mettre à mal la mutualisation du risque décès entre les assurés.

Au niveau des individus, les résultats sont présentés dans le tableau 13.

Classe d'individus	Nombre de décès annuel (en espérance)				
	Résultat réel	Modèle complet *(avec Open DAMIR et imputation)*	Modèle sans imputation *(avec Open DAMIR sans imputation)*	Modèle sans Open DAMIR	Tables de mortalité
Survie	**0**	**139**	**154**	**515**	**173**
Décès	**97**	**5**	**5**	**9**	**1**
Total	**97**	**144**	**159**	**524**	**174**

Tableau 13 – Comparatif des prédictions des différents modèles par rapport aux résultats réels en nombre de décès annuels

Le modèle que nous proposons est donc celui qui se rapproche le plus de la cible et améliore la performance de la prédiction des décès de plus de 30 points par rapport à l'utilisation de la table réglementaire, comme on le voit dans le tableau 14.

Erreur	Nombre de décès annuel (en espérance)			
	Modèle complet *(avec Open DAMIR et imputation)*	Modèle sans imputation *(avec Open DAMIR sans imputation)*	Modèle sans Open DAMIR	Tables de mortalité
Absolue	**+47**	**+62**	**+427**	**+77**
Relative	**+48,5%**	**+64,0%**	**+440,2%**	**+79,4%**

Tableau 14 – Comparatif des erreurs de prédictions des différents modèles

Dans ce cas précis, en traduisant sous la forme de provisionnement, cela donnerait une provision inférieure à celui de la table réglementaire, ce qui génèrerait autant de résultat disponible pour l'assureur. Inversement, si nous étions dans un cas de surmortalité par rapport à la table réglementaire, le modèle développé aurait abouti à un plus grand provisionnement et donc aurait permis de limiter des mali de sinistralité.

Enfin, ces résultats démontrent par la même occasion que l'utilisation des données Open DAMIR a constitué un réel gain pour la qualité de la prédiction puisque le même modèle sans l'apport de l'*open data* ne donne que des résultats assez médiocres. De même, l'imputation des données manquantes fonctionne puisqu'elle a permis d'améliorer sensiblement les performances du modèle.

CONCLUSION

Dans la construction du modèle proposé, le point fondamental a été la définition précise de la cible, c'est-à-dire de ce qui allait être précisément demandé au modèle de classification et sous quelles circonstances la prédiction serait exprimée. L'étude et le traitement préalable des données, quel que soit leur source (*open data* ou assureur) était également un point crucial, que ce soit en imputant les valeurs manquantes, en les travaillant de manière à faire sortir des informations pertinentes ou en prenant en compte le caractère rare de l'événement à prédire.

En apportant des réponses appropriées à ces sujets, notamment via les algorithmes de *machine learning*, les techniques de rééchantillonnage et le choix des métriques d'évaluation, nous avons pu obtenir un résultat satisfaisant en aboutissant au final à un provisionnement plus juste que celui des tables réglementaires.

Néanmoins, ce résultat n'est pas exempt de critiques. En particulier, il faudrait éprouver le modèle proposé sur plusieurs années pour vérifier sa fiabilité avant de pouvoir s'en servir comme méthode de provisionnement en conditions réelles. En effet, le modèle en question a été entrainé sur des données N pour estimer un provisionnement N+1. L'entrainement serait probablement à renouveler chaque année, pour prendre en compte l'évolution du portefeuille assuré. S'il devait servir véritablement au moment de constituer les provisions, les bases permettant de valider son bon réentrainement ne seraient pas encore disponibles. Il faudrait alors envisager un processus de *backtesting* afin de valider le modèle *a posteriori*.

Par ailleurs, le caractère relativement non explicable de l'algorithme de *machine learning* utilisé (*random forest)* est un vrai frein pour son utilisation opérationnelle[99] : l'absence d'informations sur la manière dont chacun des critères fait varier la prédiction restreint les possibilités d'analyses, d'études d'impacts et de projections.

Il est en revanche intéressant de noter que sans l'apport des données de l'Open DAMIR, le modèle développé est très éloigné de la réalité. C'est l'ajout de ces données qui a permis à notre modèle d'avoir un résultat satisfaisant et qui valide notre hypothèse de départ sur l'intérêt de l'*open data*, même utilisé sous de fortes contraintes pour assurer l'anonymisation des données et la sauvegarde de la mutualisation entre les assurés. Nous avons, de la même manière, réussi à démontrer l'apport de l'imputation des données manquantes puisque celle-ci permet d'améliorer encore l'efficacité de notre modèle.

[99] Il convient de nuancer cette affirmation car il est désormais possible de mesurer l'importance de chacune des variables explicatives sur la prédiction via une analyse *Ceteris Paribus*, c'est-à-dire en voyant comment évoluent les résultats du modèle selon l'évolution des variables une par une. Cependant, cela produit parfois des résultats qui restent difficilement interprétables, avec notamment des discontinuités dans le profil de certaines variables.

CONCLUSION

CONCLUSION GENERALE

Les techniques d'intelligence artificielle et de *big data* ne devraient pas être aveuglément mises en place comme un nouvel *el dorado* : ce n'est pas parce que la *data science* donne accès à de nouveaux traitements de la donnée que ces innovations sont intrinsèquement préférables. En particulier, elles peuvent se révéler intrusives pour la vie privée, biaisées ou erronées par rapport à ce qu'elles chercher à modéliser et porter potentiellement en germe la remise en cause du principe de mutualisation en assurance.

Pour autant, ce serait probablement une erreur de refuser les avancées technologiques du seul fait de l'existence de ces risques et de se priver de ces nouvelles façons d'extraire de l'information qui pourrait être bénéfique pour la collectivité. Il faut néanmoins les utiliser de manière raisonnée, en cultivant un cadre éthique et responsable de l'utilisation de l'intelligence artificielle en assurance. Afin d'illustrer ce qui pourrait constituer ce cadre, nous avons élaboré au sein du présent mémoire un cas d'usage d'algorithme de *machine learning* appliqué aux données issues de l'*open data* de santé afin d'améliorer la prédiction des risques de décès pour le provisionnement en prévoyance.

Le modèle mis en œuvre s'appuie sur Open DAMIR, une base issue de l'anonymisation des données du SNIIRAM sur le remboursement des actes de soin par l'assurance maladie obligatoire. Il a tout d'abord fallu retravailler les données afin de traiter la problématique des valeurs manquantes. Cette étape est indispensable dès lors que la présence de ces valeurs manquantes ne sont pas complètement dues au hasard et peuvent donc introduire un biais important sur les résultats du modèle. Cela a été en outre l'occasion de tester certains algorithmes de *machine learning*. Le plus performant d'entre eux pour réaliser l'imputation des valeurs manquantes s'est révélé être celui des forêts aléatoires.

Les données de l'*open data* de santé retravaillées ont ensuite été appariées avec les données de l'assureur – à une maille très agrégée afin de limiter le risque de réidentification et de conserver l'effet de mutualisation – pour former la base servant d'entrée au modèle. Néanmoins, pour prédire des événements rares comme les décès, il est nécessaire d'utiliser des techniques de rééchantillonnage. Le choix du modèle final nécessite donc d'optimiser en même temps l'algorithme prédictif et celui de préparation des données. Au final, la technique des forêts aléatoires a de nouveau été retenue.

Le modèle ainsi développé donne d'excellents résultats sur notre jeu de test puisqu'il a presque divisé par deux l'erreur entre prédiction et décès réellement constatés par rapport à la table réglementaire. Dans le même temps, nous avons prouvé que l'*open data* de santé constitue un apport réel pour la performance du modèle, le même algorithme sans ces données affichant de piètres performances. Ces résultats demandent à être confirmés en généralisant sur d'autres jeux de données mais n'en demeurent pas moins prometteurs.

Cependant, à travers le modèle prédictif réalisé, c'est davantage la démarche que le résultat final qui nous semble intéressante. En effet, nous avons cherché à prouver au travers de ce mémoire qu'il était possible d'utiliser les innovations permises par le *big data* et l'intelligence artificielle de manière efficace tout en respectant les contraintes éthiques

et réglementaires liées à la protection des données à caractère personnel, les potentiels biais dissimulés au sein des algorithmes et la sauvegarde de la mutualisation du risque assurantiel.

En conclusion, nous souhaitons réfléchir à la question de savoir si la *data science* remplacera l'actuariat dans l'avenir. Cette interrogation paraît légitime au vu des formidables avancées permises par cette nouvelle matière dans le traitement de la donnée. Nous pensons ainsi qu'il est indispensable que les actuaires s'arment de ces nouvelles techniques car elles peuvent proposer des améliorations substantielles à leurs modèles de prédiction du risque. Cela étant dit, afin d'assurer une utilisation éthique de l'intelligence artificielle en assurance telle que nous l'avons définie, l'actuaire nous semble indispensable en tant qu'expert du secteur soumis à un code de déontologie. Celui-ci doit rester au centre de l'analyse des données, de la conception des modèles et de la prise de décision qui en résulte pour vérifier que les risques inhérents à ces nouvelles technologies soient encadrés et que la démutualisation n'ait effectivement pas lieu.

BIBLIOGRAPHIE

Publication et ouvrages scientifiques

ANTONIO, K. et al. (2020) – *When stakes are high: balancing accuracy and transparency with Model-Agnostic Interpretable Data-driven suRRogates*

Disponible sur : <https://arxiv.org/pdf/2007.06894.pdf> (consulté en décembre 2020)

BATISTA, G. E. A. P. A. et al. (2004) – A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data – *ACM SIGKDD Explorations Newsletter*, vol. 6

Disponible sur : <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.58.7757&rep=rep1&type=pdf> (consulté en août 2020)

BENNETT, J. et LANNING, S. (2007) – The Netflix prize – *KDD Cup and Workshop in conjunction with KDD*

Disponible sur : <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.6998> (consulté en août 2020)

BRAND, J. P. L. (1999) – *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets* – Thèse de doctorat : Université de Rotterdam – 224 p. – p. 15-24

Disponible sur : <https://core.ac.uk/download/pdf/18508128.pdf> (consulté en août 2020)

BREIMAN, L. (2001) – Random forests – *Machine Learning*, vol. 45, n° 1 – p. 5-32

Disponible sur : <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf> (consulté en août 2020)

BREIMAN, L. et al. (1984) – *Classification and Regression Trees* – Wadsworth & Brooks – 368 p. – ISBN: 978-0412048418

CHAI, T. et DRAXLER, R. (2014) – Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature – *Geoscientific Model Development*, vol. 7, n° 3 – p. 1247-1250

Disponible sur : <https://pdfs.semanticscholar.org/11c9/aefb2fa45b9fd3292454ff8de134cfd1c6b1.pdf> (consulté en août 2020)

CHARPENTIER, A. et al. (2015) – Segmentation et mutualisation : les deux faces d'une même pièce ? – *Risques*, vol. 103 – p. 19-23

Disponible sur : <https://f.hypotheses.org/wp-content/blogs.dir/253/files/2015/10/Risques-Charpentier-Denuit-Elie.pdf> (consulté en février 2021)

CHAW, N. V. et al. (2002) – SMOTE: Synthetic Minority Over-sampling Technique – *Journal of Artificial Intelligence Research*, vol. 16 – p. 321-357

Disponible sur : <http://scholar.google.fr/scholar_url?url=https://www.jair.org/index.php/jair/article/%20download/10302/24590&hl=fr&sa=X&scisig=AAGBfm0zNdcfXdPynWxoQ3FsFum2KdF9ow&nossl=1&oi=scholarr> (consulté en août 2020)

CHE, Z. et al. (2018) – Recurrent neural networks for multivariate time series with missing values – *Scientific Reports*, vol. 8, n° 6085

Disponible sur : <https://www.nature.com/articles/s41598-018-24271-9.pdf> (consulté en août 2020)

DE MONTJOYE, Y.-A. et al. (2013) – Unique in the Crowd: The privacy bounds of human mobility – *Scientific Reports*, vol. 3, n° 1373

Disponible sur : <https://www.nature.com/articles/srep01376> (consulté en août 2020)

FRANCOEUR, D. (2010) – Machines à vecteurs de support : une introduction – *CaMUS*, vol. 1 – p. 7–25 – Université de Sherbrooke

Disponible sur : <http://camus.math.usherbrooke.ca/revue/revue1/article2.pdf> (consulté en novembre 2020)

KING, G. et al. (2001) – Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation – *American Political Science Review*, vol. 95, n° 1 – p. 49–69

Disponible sur : <https://gking.harvard.edu/files/gking/files/evil.pdf> (consulté en août 2020)

KONISHI, S. et KITAGAWA, G. (2008) – *Information Criteria and Statistical Modeling* – Springer – 273 p. – ISBN: 978-0387718866 – p. 1-8

LITTLE, R. J. A. et RUBIN, D. B. (1987) – *Statistical analysis with missing data* – 1ère éd. – John Wiley & Sons

LITTLE, R. J. A. et RUBIN, D. B. (2002) – *Statistical analysis with missing data* – 2ème éd. – John Wiley & Sons – 408 p. – ISBN: 978-1118625880

MACHANAVAJJHALA, A. et al. (2007), l-Diversity: Privacy beyond k-anonymity – *ACM Transactions on Knowledge Discovery from Data*, vol. 1, n°1, art. 3 – 52 p. – p. 17

Disponible sur : <https://desfontain.es/PDFs/PhD/LDiversityPrivacyBeyondKAnonymity.pdf> (consulté en août 2020)

MC CULLOCH, W. et PITTS, W. (1943) – A Logical Calculus of Ideas Immanent in Nervous Activity – *Bulletin of Mathematical Biophysics*, vol. 5 – p. 115-133

MC LACHLAN, G. J. et al. (2004) – *Analyzing Microarray Gene Expression Data* – John Wiley & Sons – 368 p. – ISBN: 978-0471226161

MEKONTSO FOTSING, A. C. (2018) – *L'Open DAMIR : apport à la maîtrise des dépenses de santé* – Mémoire d'actuariat : ISFA – 151 p.

MINSKY, M. L. (1967) – *Computation: Finite and Infinite Machines* – Prentice Hall – 317 p. – ISBN: 978-0131655638

O'NEIL, C. (2016) – *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* – 209 p. – ISBN: 978-0451497338

Disponible sur : <http://governance40.com/wp-content/uploads/2019/03/Weapons-of-Math-Destruction-Cathy-ONeil.pdf> (consulté en août 2020)

O'NEIL, C. (2020) – Interview par Nicolas Martin – *La Méthode scientifique*, émis. du 30 avril 2020 – France Culture

Disponible sur : <https://www.franceculture.fr/emissions/la-methode-scientifique/cathy-oneil-pour-une-ethique-des-algorithmes-0> (consulté en août 2020)

RAKOTOMALALA, R. (2020a) – *Pratique de l'Analyse Discriminante Linéaire* – Université Lumière Lyon 2 – 281 p.

Disponible sur : <http://eric.univ-lyon2.fr/~ricco/cours/cours/Pratique_Analyse_Discriminante_Lineaire. pdf> (consulté en octobre 2020)

RAKOTOMALALA, R. (2020b) – *Gradient Boosting* : *Technique ensembliste pour l'analyse prédictive* – Université Lumière Lyon 2 – 28 p.

Disponible sur : <http://eric.univ-lyon2.fr/~ricco/cours/slides/gradient_boosting.pdf> (consulté en octobre 2020)

ROCHER, L. et al. (2019) – Estimating the success of re-identifications in incomplete datasets using generative models – *Nature Communications*, vol. 10, art. 3069

Disponible sur : <https://www.nature.com/articles/s41467-019-10933-3.pdf> (consulté en août 2020)

SCHAFER, J. L. (1997) – *Analysis of Incomplete Multivariate Data* – Chapman & Hall / CRC – 514 p. – ISBN: 978-0367803025

SCHAFER, J. L. et GRAHAM, J.W. (2002) – Missing Data: Our View of the State of the Art – *Psychological Methods*, vol. 7, n° 2 – p. 147–177

Disponible sur : <http://sta559s11.pbworks.com/f/schafer_graham_MissingDataPsychMethods02.pdf> (consulté en août 2020)

SCHWENDER, H. (2012) – Imputing Missing Genotypes with Weighted k Nearest Neighbors – *Journal of Toxicology and Environmental Health, part. A*, vol. 75 – p.438-446

Disponible sur : <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.861.4155&rep=rep1&type=_pdf> (consulté en août 2020)

SWEENEY, L. (2002) – k-anonymity: a model for protecting privacy – *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* vol. 10, n°5 – p. 557-570

Disponible sur : <https://epic.org/privacy/reidentification/Sweeney_Article.pdf> (consulté en août 2020)

SUTTON, O. (2012) – *Introduction to k Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction* – 10 p.

Disponible sur : <http://www.math.le.ac.uk/people/ag153/homepage/KNN/OliverKNN_Talk.pdf> (consulté en août 2020)

TROYANSKAYA, O. et al. (2001) – Missing value estimation methods for dna microarrays – *Bioinformatics*, vol. 17, n° 6 – p. 520-525

Disponible sur : <https://www.researchgate.net/publication/220263062_Missing_Value_Estimation_Methods_for_DNA_Microarrays> (consulté en août 2020)

TURING, A. M. (1950) – Computing Machinery and Intelligence – *Mind*, vol. 49 – p. 433-460

Disponible sur : <https://www.csee.umbc.edu/courses/471/papers/turing.pdf> (consulté en août 2020)

VAN BUUREN, S. (2018) – *Flexible Imputation of Missing Data* – 2nde éd. – Chapman & Hall / CRC – 326 p. – ISBN: 978-1439868249

Disponible sur : <http://pzs.dstu.dp.ua/DataMining/preprocessing/bibl/fimd.pdf> (consulté en août 2020)

WILLMOTT, C. et MATSUURA, K. (2005) – Advantages of the mean absolute error (MAE) over root mean square error (RMSE) in assessing average model performance – *Climate research*, vol. 30, n° 1, p. 79-82

Articles

ANDERSON, N. (2009) – "Anonymized" data really isn't—and here's why not – *Ars Technica*

Disponible sur : <https://arstechnica.com/%20tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin> (consulté en août 2020)

BARLYN, S. (2018) – Strap on the Fitbit: John Hancock to sell only interactive life insurance – *Reuters*

Disponible sur : <https://www.reuters.com/article/us-manulife-financi-john-hancock-lifeins/strap-on-the-fitbit-john-hancock-to-sell-only-interactive-life-insurance-idUSKCN1LZ1WL> (consulté en août 2020)

FRANCOIS, M. (2018) – Avec ses suggestions d'amis, Facebook favorise la mise en relation des partisans de Daech – *Le Figaro*

Disponible sur : <https://www.lefigaro.fr/secteur/high-tech/2018/05/11/32001-20180511ARTFIG00120-avec-ses-suggestions-d-amis-facebook-favorise-la-mise-en-relation-des-partisans-de-daech.php> (consulté en août 2020)

HARRELL, F. E. (2017) – Damage Caused by Classification Accuracy and Other Discontinuous Improper Accuracy Scoring Rules – *Statistical Thinking*

Disponible sur : <https://www.fharrell.com/post/class-damage> (consulté en août 2020)

HERN, A. (2017) – 'Anonymous' browsing data can be easily exposed, researchers reveal – *The Guardian*

Disponible sur : <https://www.theguardian.com/technology/2017/aug/01/data-browsing-habits-brokers> (consulté en août 2020)

LANEY, D. (2001) – *3D Data Management: Controlling Data Volume, Velocity, and Variety* – META Group.

Disponible sur : <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf> (consulté en août 2020)

LE GOFF, E. (2018) – Les 10 risques les plus redoutés par les entreprises (baromètre Allianz – *L'Argus de l'assurance*

Disponible sur : <https://www.argusdelassurance.com/gestion-des-risques/grands-risques/les-10-risques-les-plus-redoutes-par-les-entreprises-barometre-allianz.125752> (consulté en novembre 2020)

PERRIN, G. (2019) – Données personnelles : le groupe québécois Desjardins victime d'une fuite massive – *L'Argus de l'assurance*

Disponible sur : <https://www.argusdelassurance.com/tech/donnees-personnelles-le-groupe-quebecois-desjardins-victime-d-une-fuite-massive.149295> (consulté en août 2020)

PORTE, D. (2019) – L'intelligence artificielle au secours de la police – *Police Scientifique*

Disponible sur : <https://www.police-scientifique.com/articles-recents/lintelligence-artificielle-au-secours-de-la-police> (consulté en août 2020)

REESE, H. (2016) – Why Microsoft's 'Tay' AI bot went wrong – *Tech Republic*

Disponible sur : <https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong> (consulté en août 2020)

TANWAR, S. (2019) – Introduction to machine learning and deep learning – *Medium*

Disponible sur : <https://medium.com/@sanchittanwar75/introduction-to-machine-learning-and-deep-learning-bd25b792e488> (consulté en août 2020)

TASSET, M. (2019) – Le volume de données mondial sera multiplié par 45 entre 2020 et 2035 – *Journal du Net*

Disponible sur : <https://www.journaldunet.com/solutions/dsi/1424245-le-volume-de-donnees-mondial-sera-multiplie-par-45-entre-2020-et-2035-selon-statista> (consulté en août 2020)

VON PRESSESTELLE, G. (2020) – LfDI Baden-Württemberg verhängt Bußgeld gegen AOK Baden-Württemberg – Wirksamer Datenschutz erfordert regelmäßige Kontrolle und Anpassung – *LfDI Baden-Württemberg*

Disponible sur : < https://www.baden-wuerttemberg.datenschutz.de/lfdi-baden-wuerttemberg-verhaengt-bussgeld-gegen-aok-baden-wuerttemberg-wirksamer-datenschutz-erfordert-regelmaessige-kontrolle-und-anpassung> (consulté en novembre 2020)

WHITTAKER, Z. (2020) – Cyber insurer Chubb had data stolen in Maze ransomware attack – *Techcrunch*

Disponible sur : <https://techcrunch.com/2020/03/26/chubb-insurance-breach-ransomware/?guccounter=1> (consulté en novembre 2020)

YATES, J. (2017) – J'ai testé les algorithmes de Facebook et ça a rapidement dégénéré – *Radio-Canada*

Disponible sur : <https://ici.radio-canada.ca/nouvelle/1029916/experience-facebook-algorithmes-bulle-desinformation> (consulté en août 2020)

Textes juridiques et réglementaires

CIRCULAIRE – *Circulaire du 26 mai 2011* relative à la création du portail unique des informations publiques de l'Etat « data.gouv.fr » par la mission « Etalab » et l'application des dispositions régissant le droit de réutilisation des informations publiques

Disponible sur : <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000024072788&cate gorieLien=id> (consulté en août 2020)

CNIL (2015) – *Délibération n°2015-255 du 16 juillet 2015* refusant la mise en œuvre par la société JCDecaux d'un traitement automatisé de données à caractère personnel ayant pour finalité de tester une méthodologie d'estimation quantitative des flux piétons sur la dalle de La Défense

Disponible sur : <https://www.legifrance.gouv.fr/affichCnil.do?oldAction=rechExpCnil&id=CNILTEXT000 031159401> (consulté en août 2020)

CODE – *Code des relations entre le public et l'administration* – Livre III (codification issue de la loi n° 78-753 du 17 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal, dite « Loi CADA »)

Disponible sur : < https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000031366350/LEGISCTA000031367685?codeTitle=relation#LEGISCTA000031367685> (consulté en août 2020)

CODE – *Code de la sécurité sociale* – Art. L161-28-1

Disponible sur : <https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=2213F5E897954B4755D8BE1F1222F01D.tplgfr27s_2?idArticle=LEGIARTI000038886948&cidTexte=LEGITEXT000006073189&categorieLien=id&dateTexte=> (consulté en août 2020)

CODE – *Code de la santé publique* – Art. L.1111-8

Disponible sur : <https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000033862549&cidTexte=LEGITEXT000006072665&dateTexte=20180401> (consulté en août 2020)

CODE – *Code de la santé publique* – Art. L.1110-4

Disponible sur : <https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=7D336D5AA99F278787567C1D3EF0E117.tplgfr27s_2?idArticle=LEGIARTI000036515027&cidTexte=LEGITEXT000006072665&categorieLien=id&dateTexte=> (consulté en août 2020)

CODE – *Code de la santé publique* – Art. L.1461-1 et suivants

Disponible sur : <https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000038886868> (consulté en novembre 2020)

CODE – *Code de la santé publique* – Art. L.1462-1 et suivants

Disponible sur : <https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006072665/LEGISCTA000031923858/#LEGISCTA000031923861> (consulté en novembre 2020)

DECRET – *n° 2016-1871 du 26 décembre 2016* relatif au traitement de données à caractère personnel dénommé « système national des données de santé »

Disponible sur : < https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033702840&dateTexte=20200903> (consulté en août 2020)

JORF – *n° 0193 du 22 août 2014* – p. 13 972 – texte n°89

Disponible sur : <https://www.legifrance.gouv.fr/affichTexte.do;jsessionid=7D336D5AA99F278787567C1D3EF0E117.tplgfr27s_2?cidTexte=JORFTEXT000029388087&dateTexte=&oldAction=rechJO&categorieLien=id&idJO=JORFCONT000029387119> (consulté en août 2020)

LOI – *n° 78-17 du 6 janvier 1978* relative à l'informatique, aux fichiers et aux libertés, dite « Loi Informatique et Libertés »

Disponible sur : <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000886460> (consulté en août 2020)

LOI – *n° 2016-41 du 26 janvier 2016* de modernisation de notre système de santé

Disponible sur : <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000031912641&categorieLien=id#JORFARTI000031914480> (consulté en août 2020)

REGLEMENT (UE) – *Règlement délégué 2015/35 de la Commission du 10 octobre 2014* complétant la directive 2009/138/CE du Parlement européen et du Conseil sur l'accès aux activités de l'assurance et de la réassurance et leur exercice (Solvabilité II)

Disponible sur : <https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32015R0035&from= FR> (consulté en août 2020)

REGLEMENT (UE) – *Règlement 2016/679 du Parlement européen et du Conseil du 27 avril 2016* relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE dit « Règlement Général sur la Protection des Données »

Disponible sur : <https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:32016R0679> (consulté en août 2020)

Sites internet

CNIL – *Définition : Finalité d'un traitement* – www.cnil.fr

Disponible sur : <https://www.cnil.fr/fr/definition/finalite-dun-traitement> (consulté en octobre 2020)

DATA.GOUV.FR – *Open DAMIR : base complète sur les dépenses d'assurance maladie inter régimes* – www.data.gouv.fr

Disponible sur : <https://www.data.gouv.fr/fr/datasets/open-damir-base-complete-sur-les-depenses-dassurance-maladie-inter-regimes/# > (consulté en août 2020)

DREES – *Mise en œuvre du système national des données de santé et nouveau cadre d'accès aux données de santé* – drees.solidarites-sante.gouv.fr

Disponible sur : <https://drees.solidarites-sante.gouv.fr/etudes-et-statistiques/acces-aux-donnees-de-sante/article/mise-en-oeuvre-du-systeme-national-des-donnees-de-sante-et-nouveau-cadre-d> (consulté en octobre 2020)

IMBALANCED-LEARN – *Tomek Links* – imbalanced-learn.readthedocs.io

Disponible sur : <https://imbalanced-learn.readthedocs.io/en/stable/under_sampling.html#tomek-s-links> (consulté en octobre 2020)

INDS – *Les composantes du SNDS* – www.indsante.fr

Disponible sur : <https://www.indsante.fr/fr/les-composantes-du-snds> (consulté en octobre 2020)

OPEN KNOWLEDGE FOUNDATION – *Global Open Data Index* – okfn.org

Disponible sur : <https://index.okfn.org/place> (consulté en août 2020)

OPEN KNOWLEDGE FOUNDATION – *What is open?* – okfn.org

Disponible sur : <https://okfn.org/opendata> (consulté en octobre 2020)

SCIKIT-LEARN – *Cross-validation: evaluating estimator performance* – scikit-learn.org

Disponible sur : <https://scikit-learn.org/stable/modules/cross_validation.html> (consulté en août 2020)

SCIKIT-LEARN – *Importance of Feature Scaling* – scikit-learn.org

Disponible sur : <https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance. html> (consulté en août 2020)

SHAP – *Documentation* – shap.readthedocs.io

Disponible sur : <https://shap.readthedocs.io/en/latest> (consulté en décembre 2020)

SNDS – *Processus d'accès aux données* – www.snds.gouv.fr

Disponible sur : <https://www.snds.gouv.fr/SNDS/Processus-d-acces-aux-donnees> (consulté en août 2020)

THE PUBLIC VOICE – *Universal Guidelines for Artificial Intelligence* – thepublicvoice.org

Disponible sur : <https://thepublicvoice.org/ai-universal-guidelines> (consulté en août 2020)

TRANSPORT.DATA.GOUV.FR – *Qu'est-ce que le Point d'Accès National ?* – doc.transport.data.gouv.fr

Disponible sur : <https://doc.transport.data.gouv.fr/guide-du-pan> (consulté en octobre 2020)

WIKIPÉDIA – *Fonction de hachage cryptographique* – wikipedia.org

Disponible sur : <https://fr.wikipedia.org/wiki/Fonction_de_hachage_cryptographique> (consulté en août 2020)

Autres publications

ACPR (2018) – *Intelligence artificielle : enjeux pour le secteur financier* – 37 p.

Disponible sur : <https://acpr.banque-france.fr/sites/default/files/medias/documents/2018_12_20_intelligence_artificielle_fr_0.pdf> (consulté en août 2020)

BRAS, P.-L. (2013) – *Rapport sur la gouvernance et l'utilisation des données de santé* – DREES dans le cadre de la mission confiée par le Ministre des Affaires sociales et de la santé – 128 p. – p. 27

CCSF (2020) – *Recommandation du Comité consultatif du secteur financier sur la déshérence de l'épargne retraite supplémentaire*

Disponible sur : <https://www.ccsfin.fr/sites/default/files/reco_ccsf_desherence_retraites_supplementaire_21012020_def.pdf> (consulté en août 2020)

CESIN (2019) – *Baromètre de la cyber-sécurité des entreprises* – 51 p.

Disponible sur : <https://www.opinion-way.com/fr/component/edocman/?task=document.viewdoc&id=2019&Itemid=0> (consulté en novembre 2020)

CNAM (2015) – *Le système national d'information interrégimes de l'Assurance Maladie* – Direction de la Stratégie, des Études et des Statistiques – 26 p.

Disponible sur : <https://www.ameli.fr/fileadmin/user_upload/documents/Presentation_du_Sniiram.pdf> (consulté en août 2020)

CNIL (2017) – *Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle* – Synthèse du débat public animé par la CNIL dans le cadre de la mission de réflexion éthique confiée par la loi pour une république numérique – 80 p.

Disponible sur : <https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_garder_la_main_web.pdf> (consulté en août 2020)

COMMISSION DE REGULATION DE L'ENERGIE (2017) – *Rapport du comité d'études relatif aux données dont disposent les gestionnaires de réseaux et d'infrastructures d'énergie* – 116 p.

Disponible sur : <https://www.cre.fr/Documents/Publications/Rapports-thematiques/Rapport-donnees-gestionnaires-de-reseaux-et-d-infrastructures-d-energie> (consulté en octobre 2020)

DICK, P. K. (1956) – *Rapport minoritaire*

GEFFRAY, E. (2016) – Quelle protection des données personnelles dans l'univers de la robotique ? – *Dalloz IP/IT : Robots, intelligence artificielle et droit* – p. 295

GRÄFE, S. (2019) – *Les enjeux et stratégies dans la transformation digitale de l'assurance*, 2ème éd. – Groupe XERFI – 200 p.

GROUPE DE TRAVAIL « ART. 29 » SUR LA PROTECTION DES DONNÉES (2014) – Avis 05/2014 (WP216) sur les techniques d'anonymisation – 42 p.

Disponible sur : <https://www.cnil.fr/sites/default/files/atoms/files/wp216_fr.pdf> (consulté en août 2020)

GROUPE DE TRAVAIL « ART. 29 » SUR LA PROTECTION DES DONNÉES (2018) – Avis 17/FR (WP251rev.01) sur les lignes directrices relatives à la prise de décision individuelle automatisée et au profilage aux fins du règlement (UE) 2016/679 – 43 p. – p. 28

Disponible sur : <https://www.cnil.fr/sites/default/files/atoms/files/wp251_profilage-fr.pdf> (consulté en novembre 2020)

INSTITUT DES ACTUAIRES (2014) – Code de déontologie – 7 p.

Disponible sur : < https://www.institutdesactuaires.com/global/gene/link.php?news_link=2014134730_deontologie.pdf&fg=1> (consulté en août 2020)

LA FABRIQUE D'ASSURANCE (2019) – *Livre blanc : Intelligence artificielle et éthique dans le secteur de l'assurance* – 131 p. – ISBN : 978-2955828533

MALAKOFF HUMANIS (2018) – *Politique de gestion des données de santé et médicales* – Définition Donnée Médicale

OHM, P. (2009) – Broken promises of privacy: responding to the surprising failure of anonymization – *UCLA Law Review*, vol. 57 – 77 p.

Disponible sur : <https://www.uclalawreview.org/pdf/57-6-3.pdf> (consulté en août 2020)

PING AN INSURANCE (2019) – *Interim Report* – 164 p. – p. 20

Disponible sur : <http://www.pingan.com/app_upload/images/info/upload/d645e2ea-b15f-4616-a311-7704dd 34667a.pdf> (consulté en décembre 2020)

SCOR (2018) – *The impact of artificial intelligence on the (re)insurance sector* – 36 p. – ISSN: 1638-3133

Disponible sur : <https://www.scor.com/sites/default/files/focus_scor-artificial_intelligence.pdf> (consulté en août 2020)

SIMMONS&SIMMONS (2017) – *Expertise juridique sur l'intérêt public dans le contexte des données de santé* – Rapport réalisé à la demande de l'INDS – 50 p.

Disponible sur : <https://www.indsante.fr/sites/default/files/Documents_publics/rapport_dexpertise_juridique_sur_l_evaluation_de_linteret_public.pdf> (consulté en août 2020)

VILLANI, C. (2018) – *AI for Humanity : Donner un sens à l'intelligence artificielle* – Mission parlementaire – 233 p. – ISBN : 978-2111457089

Disponible sur : <https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf> (consulté en août 2020)

ANNEXES

ANNEXE 1 : PROCEDURE D'ACCES SNDS

La procédure d'évaluation de la demande d'accès au SNDS est décrite dans la documentation de la façon suivante :

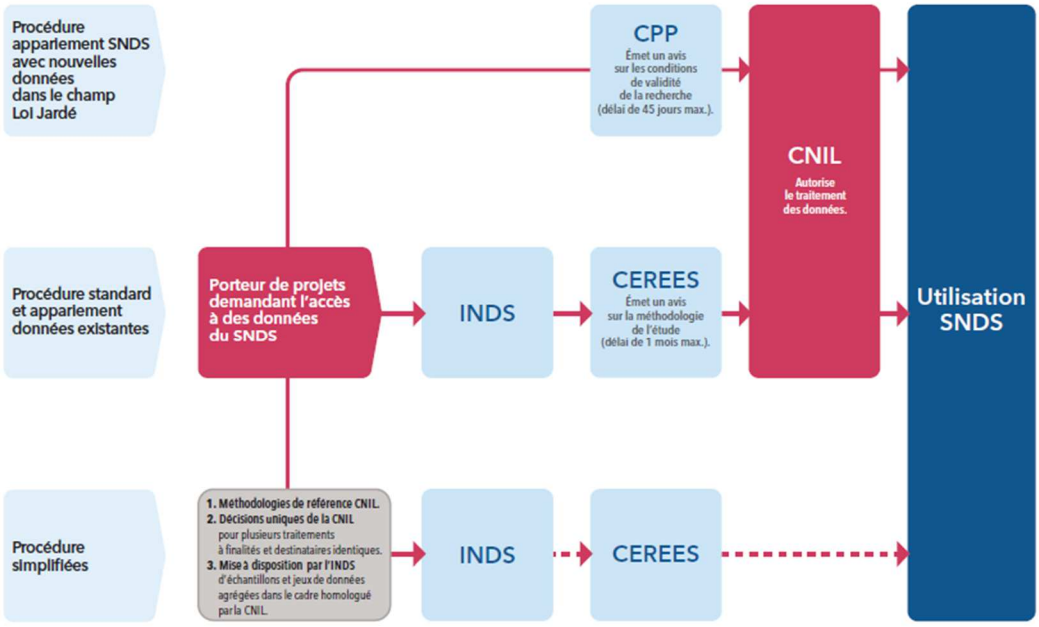


Figure 34 – Schématisation du circuit des demandes d'accès au SNDS [snds.gouv.fr]

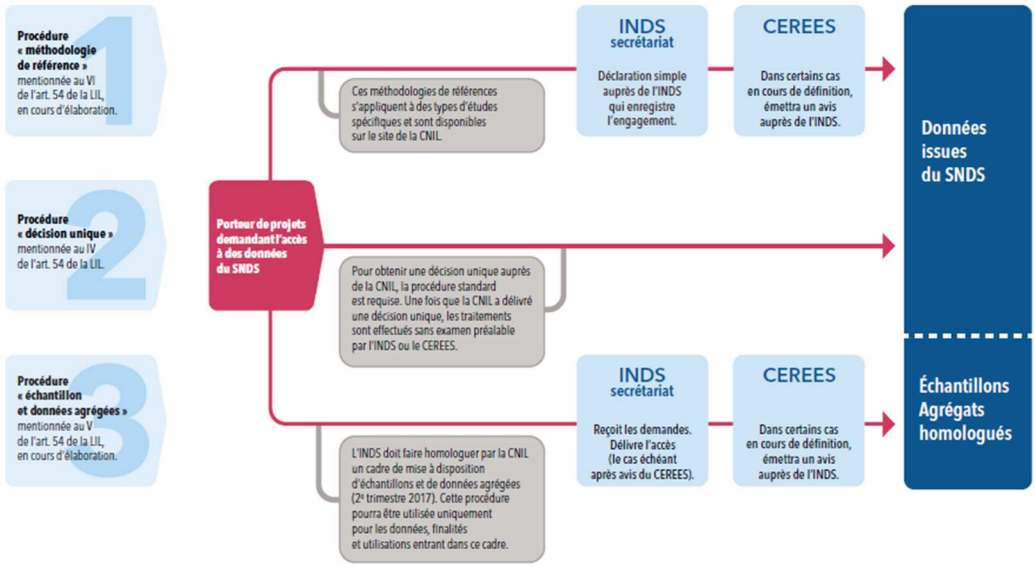


Figure 35 – Schématisation des trois procédures simplifiées selon les usages et les données [snds.gouv.fr]

ANNEXE 2 : TECHNIQUES DE REDUCTION DU RISQUE DE REIDENTIFICATION

Pseudonymisation

La pseudonymisation est définie par le RGPD comme « *Le traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable* » [REGLEMENT UE (2016)].

Il ne s'agit pas *stricto sensu* d'une technique d'anonymisation [REGLEMENT UE (2016)]. Néanmoins, il s'agit du premier pas nécessaire à l'anonymisation des données. Elle a pour objectif de supprimer les données directement identifiantes au sein de la base.

Il existe plusieurs techniques permettant d'obtenir ce résultat qui sont réparties selon trois catégories :

- la dé-identification a pour but de simplement retirer les données directement identifiantes ou de les remplacer par un code (incrémenteur ou aléatoire) ;
- la tokenisation consiste à remplacer un identifiant (typiquement le numéro de sécurité sociale) par un numéro unique moins riche en information :
- le chiffrement permet de rendre inintelligibles les données d'une base si la clé de déchiffrement n'est pas possédée.

Parmi les techniques de chiffrement classiques, nous pouvons citer le hachage et le salage. La fonction de hachage (ou *hash function*) est une fonction non réversible qui permet à partir d'une entrée de longueur variable d'obtenir une sortie de longueur fixe servant à identifier de manière fiable et unique la donnée initiale[100].

[100] Afin d'atteindre cet objectif, les fonctions de hachage vont présenter différentes caractéristiques spécifiques comme le fait qu'il soit impossible pour une valeur de hachage donnée de trouver la donnée d'entrée correspondante, le fait que deux données d'entrée distinctes donnent nécessairement deux valeurs de hachage différentes ou encore le fait qu'un très léger changement dans la donnée d'entrée donne un changement important dans la valeur de hachage.

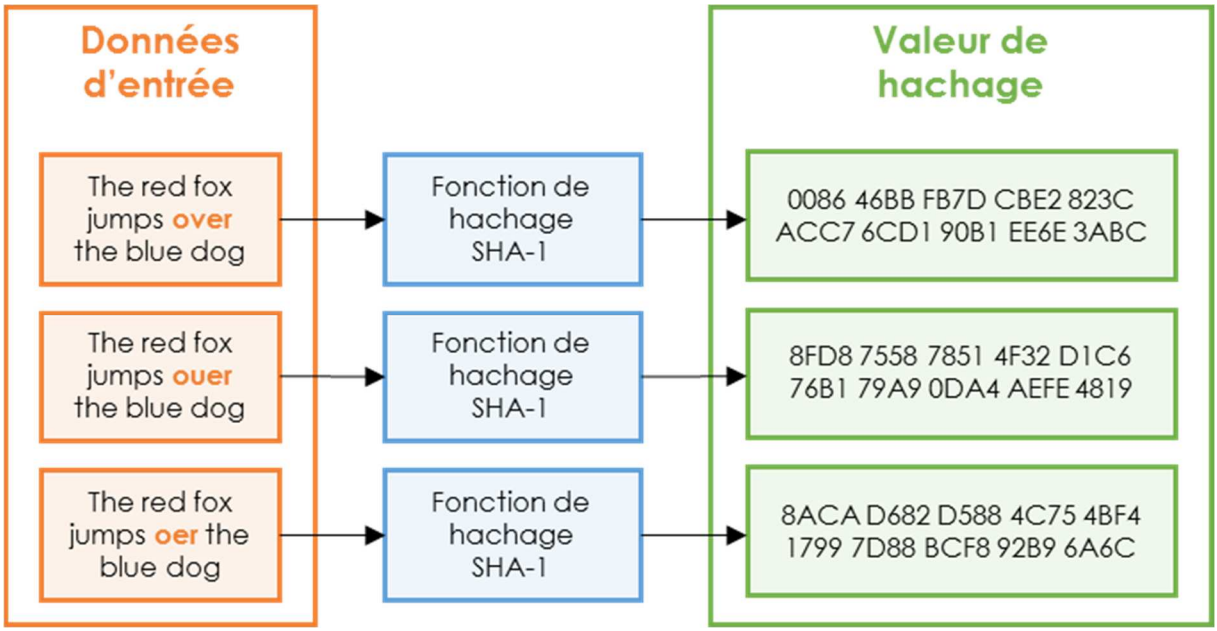


Figure 36 – Illustration de l'effet d'une fonction de hachage [wikipedia.org]

Le salage est une technique particulière de hachage consistant à rajouter un attribut aléatoire (le sel) à la valeur qui est hachée dans le but d'empêcher que deux données d'entrée identiques aboutissent à la même valeur de hachage.

Comme évoqué précédemment, la pseudonymisation ne suffit pas au regard de la réglementation pour considérer qu'une base de données a été anonymisée. En effet, de nombreux cas ont prouvé ces dernières années qu'il était quasiment toujours possible de réidentifier les individus au sein de bases pseudonymisées.

Un cas qui a fait date est celui du Massachusetts Group Insurance Commission, qui a mis à disposition de tous dans les années 1990 une base « anonymisée » contenant toutes les visites à l'hôpital des employés de l'état du Massachusetts [ANDERSON, N. (2009)]. Avec uniquement le code postal, la date de naissance et le sexe, Latanya Sweeney, enseignant-chercheur à Harvard, a pu identifier de manière certaine le gouverneur au sein de ce jeu de données [SWEENEY, L. (2002)], obtenant ainsi toutes ses informations médicales. Elle a démontré par la suite que ces trois informations suffisaient à identifier 87% de la population américaine avec très peu d'erreurs.

De même, il a été démontré qu'il était possible de réidentifier un individu dans une base de données médicales en connaissant son sexe, code postal et sa date de naissance, dans une base de données téléphoniques sur la base de quatre points de géolocalisation [DE MONTJOYE, Y.-A. et al. (2013)]. Autre exemple : en 2016, des journalistes ont réussi à identifier des personnalités politiques au sein d'une base de données « anonyme » de trois millions de citoyens allemands contenant les historiques de recherche internet [HERN, A. (2017)] (et ainsi toutes les informations sensibles qui peuvent y être rattachées comme les informations médicales, les croyances religieuses ou les orientations sexuelles).

Suppression

La méthode de suppression totale ou partielle (par troncature) consiste à supprimer définitivement et totalement des informations de la base de données. Il est bien sûr possible de supprimer les colonnes de variables non nécessaires mais la suppression s'opère essentiellement sur les lignes de la base de données. Comme pour la pseudonymisation, il ne s'agit pas à proprement parler d'une technique d'anonymisation.

L'idée sous-jacente est qu'en supprimant une grande partie d'une base de données pseudonymisée (sans donnée directement identifiante donc), il est possible d'aboutir à un échantillon ne représentant que 10% ou même 1% de la base initiale, ce qui permettrait d'assurer l'anonymisation. En effet, dans ce cas, même si quelqu'un parvenait à réidentifier un individu, il ne pourrait pas être certain que cette réidentification soit correcte puisque la personne potentiellement identifiée a de fortes chances de ne pas faire partie de la base de données.

C'est en suivant cette logique que Netflix a publié 10% de sa base de données utilisateurs [BENNETT, J. et LANNING, S. (2007)] pour lancer un concours : si une personne prétendait alors avoir réussi à identifier quelqu'un dans cette base, l'entreprise pourrait toujours prétendre que le fait qu'un profil puisse correspondre à une personne ne signifiait pas qu'il y avait réellement réidentification.

Cela étant dit, il a été prouvé que la vraisemblance de réidentification correcte d'un individu particulier même au sein d'une base de données pseudonymisée fortement incomplète peut être estimée de manière fiable [ROCHER, L. et al. (2019)] et que, en croisant avec des données sociodémographiques et d'autres types d'études, il est possible de réidentifier une personne à partir d'une base représentant un échantillonnage de 1% de la population avec très peu d'erreurs. Une base de données incomplètes ne peut donc être considérée *a priori* comme anonyme et l'échantillonnage n'est pas une méthode suffisante pour éviter le risque de réidentification.

Changement aléatoire

La méthode de changement aléatoire (ou *randomization*) consiste à altérer la véracité des données afin d'affaiblir le lien entre les données et l'individu. En effet, si les données sont suffisamment incertaines, elles ne peuvent alors plus être rattachées à un individu. Le principe général est ainsi de bruiter la donnée au niveau individuel (réduisant ainsi le risque de réidentification) tout en conservant le profil au niveau de la population globale (pour conserver l'intégrité statistique des données).

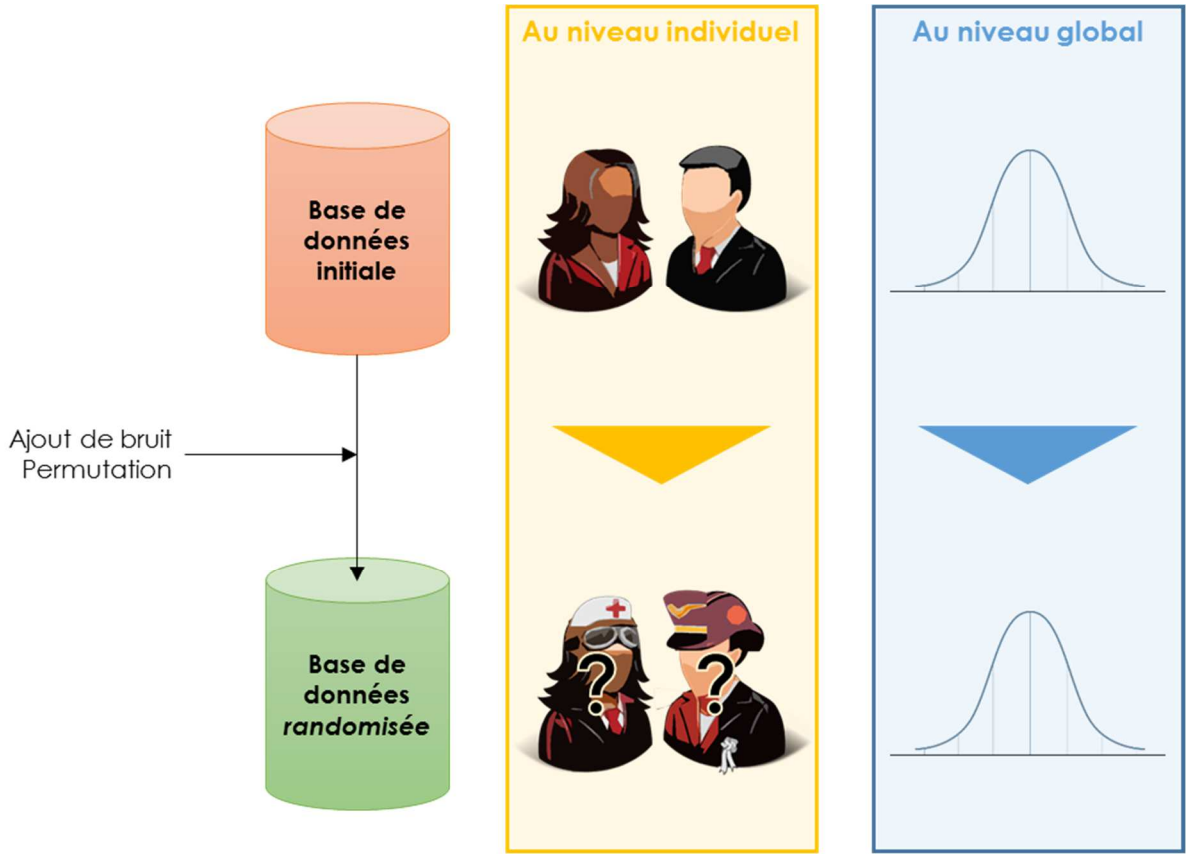


Figure 37 – Illustration de l'effet des changements aléatoires

Cette méthode utilise plusieurs techniques dont l'ajout de bruit ou la substitution :

- l'ajout de bruit consiste à modifier les attributs de la base de données tout en conservant la distribution au niveau de la population ;
- la permutation a pour objet de mélanger certaines données entre certains individus.

Dans les deux cas, la distribution statistique est bien conservée au niveau global mais la donnée n'est plus fiable au niveau de l'individu. Cependant, ces méthodes introduisent un risque de déformation des corrélations potentiellement présentes au sein des données.

Généralisation ou agrégation

La méthode de généralisation ou d'agrégation consiste à regrouper plusieurs individus au sein d'une même ligne dans la base de données en diluant les informations afin de rendre l'identification impossible au niveau d'un individu. Par exemple, cela consistera à remplacer une ville par une région, une profession par un secteur d'activité ou encore une date de naissance par une tranche d'âge. Ainsi, une ligne au sein de la base correspondra à un nombre inconnu d'individus (mais toujours supérieur à un).

De l'avis même de la CNIL, il s'agit de la seule façon réellement fiable d'assurer l'anonymisation d'une base de donnée [CNIL (2015)]. Concrètement, l'agrégation est systématiquement utilisée dans le cas de restitution de résultat d'enquête ou d'étude, en particulier lorsque ces données sont transmises à des tiers.

	Données à caractère personnel			Données médicales
	Code postal	**Age**	**Nationalité**	**Diagnostic**
1	1305*	≤ 40	*	Maladie cardiaque
4	1305*	≤ 40	*	Infection virale
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Maladie cardiaque
7	1485*	> 40	*	Infection virale
8	1485*	> 40	*	Infection virale
2	1306*	≤ 40	*	Maladie cardiaque
3	1306*	≤ 40	*	Infection virale
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Figure 38 – Illustration de la généralisation en une table 3-diverse [MACHANAVAJJHALA, A. et al. (2007)]

La généralisation utilise plusieurs techniques dont :

- **Le *k*-anonymat** consiste à ce que chaque ligne de la base de donnée ne corresponde plus à un unique individu mais à un regroupement de k individus groupés ensemble. Le groupement est rendu possible en généralisant les données (une date de naissance devient un mois de naissance, une grandeur quantitative devient un intervalle, etc.) de sorte que les k individus partagent le même attribut.

- **La *l*-diversité** permet d'étendre le concept du k-anonymat en vérifiant que les données sensibles pour chaque groupe d'individus disposent de l valeurs différentes. Ainsi, dans le tableau ci-dessus, il n'est pas possible de connaître la condition médicale d'un individu de 31 ans avec le code postal 13056 car la table est 3-diverse.

- **La *t*-proximité** est un raffinement de la l-diversité qui consiste à compléter au sein de chaque groupe d'individus les valeurs des données sensibles de sorte que la distribution de ces valeurs au sein de chaque groupe corresponde à la distribution de la base initiale.

ANNEXE 3 : SIMULATION DE L'EFFET DE DEMUTUALISATION

Afin d'illustrer l'effet macro-économique que pourrait avoir une démutualisation du risque induite par les innovations du *big data* et de l'IA, nous avons réalisé une simulation codée en R.

Introduction et notations

La charge totale de sinistre S_i d'un portefeuille i d'assurés s'écrit ainsi :

$$S_i = \sum_{k=1}^{N_i} Y_k^{(i)}$$

avec :

- N_i le nombre de sinistres de i ;
- $Y_k^{(i)}$ le coût du sinistre k de i.

La prime pure π_i correspondante pour l'assureur i sera, en fonction de ses choix et de l'information qu'il a à sa disposition :

- s'il mutualise le risque global sur l'ensemble de ses assurés : $\pi_i = \mathbb{E}[S_i]$;
- dans le cas d'une segmentation en fonction d'un facteur de risque Ω (pas toujours observable) : $\pi_i = \mathbb{E}[S_i|\Omega = \omega]$;
- dans le cas d'une segmentation en fonction des paramètres X, observables et permettant de se rapprocher de facteur de risque réel de son portefeuille : $\pi_i = \mathbb{E}[S_i|X = x]$.

Dans ce dernier cas, la prime dépend fortement du modèle et des paramètres choisis X sur la fréquence et l'ampleur des sinistres, qui peuvent être spécifiques au portefeuille de l'assureur. En supposant une indépendance approximative entre N_i et $Y_k^{(i)}$, nous obtenons :

$$\pi_i(x) = \mathbb{E}[S_i|X = x] \approx \mathbb{E}[N_i|X = x] \times \mathbb{E}\left[Y_k^{(i)}\middle|X = x\right]$$

Hypothèses de modélisation

Soit un marché de 10 000 individus et 3 assureurs couvrant le même risque. Les individus sont classés par niveau de risque (de 0 : le moins risqué, à 5 : le plus grand risque) considéré comme connu suivant la répartition suivante :

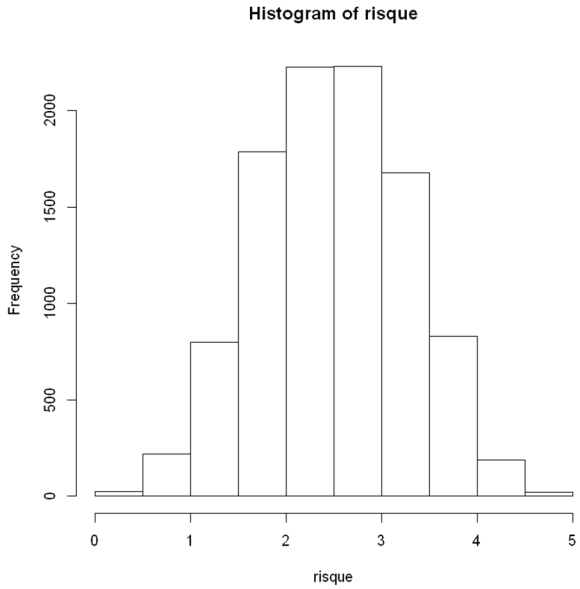


Figure 39 – Distribution des individus couverts par niveau de risque

La charge de sinistre pour chaque individu est ensuite générée aléatoirement, fonction du risque et d'un résidu normal. Le schéma suivant place chaque individu suivant la charge de sinistre (*loss*) et son niveau de risque.

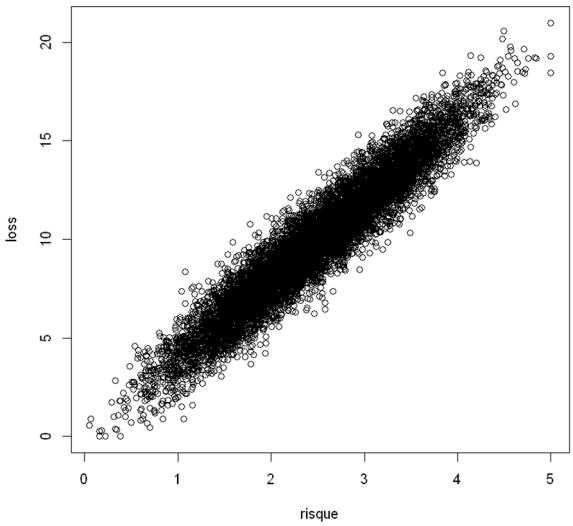


Figure 40 – Charge de sinistre par niveau de risque

Les 3 assureurs adoptent des logiques de segmentation différentes :

1. Aucune segmentation
2. Une segmentation partielle
3. Une tarification segmentée à l'extrême

Cas 1 : absence de segmentation

Sans segmentation, la prime est identique pour tous. Cela donne, en ne considérant que la prime pure :

	Assuré	Assureur
Perte	$\mathbb{E}[\boldsymbol{S}]$	$\boldsymbol{S} - \mathbb{E}[\boldsymbol{S}]$
Espérance de perte	$\mathbb{E}[\boldsymbol{S}]$	0
Variance	0	$\boldsymbol{Var}[\boldsymbol{S}]$

Ici, toute la variance est portée par l'assureur, le risque est mutualisé entre tous les assurés :

- A gauche, le nuage de points des individus, avec le niveau de la prime affichée en rouge
- A droite, la distribution de la perte pour l'assureur

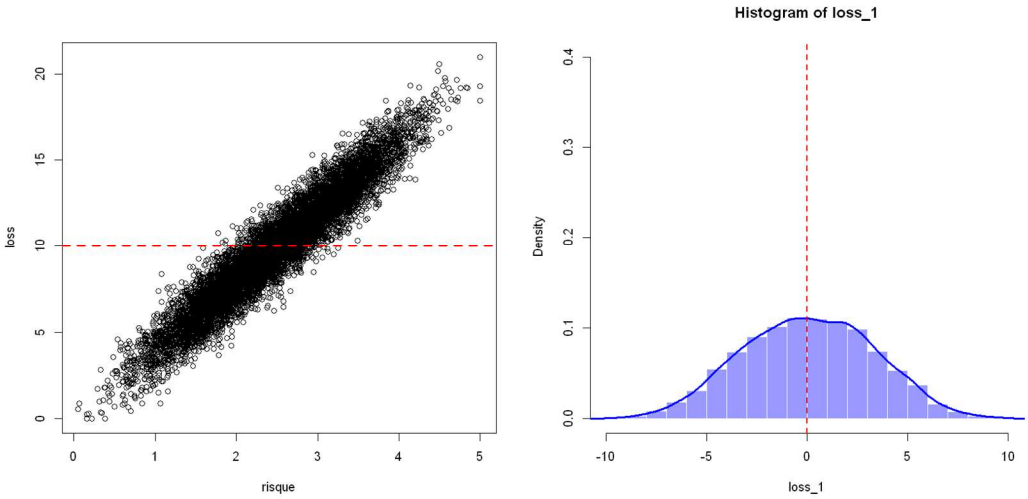


Figure 41 – Cas 1 : sans segmentation

Cas 2 : segmentation poussée à l'extrême

En considérant que le facteur de risque Ω est observable et bien connu de l'assureur et qu'il a donc ajusté sa tarification en fonction de chaque assuré, le tableau est le suivant :

	Assuré	Assureur
Perte	$\mathbb{E}[\boldsymbol{S}\|\boldsymbol{\Omega}]$	$\boldsymbol{S} - \mathbb{E}[\boldsymbol{S}\|\boldsymbol{\Omega}]$
Espérance de perte	$\mathbb{E}[\boldsymbol{S}]$	0
Variance	$\boldsymbol{Var}\left[\mathbb{E}[\boldsymbol{S}\|\boldsymbol{\Omega}]\right]$	$\boldsymbol{Var}\left[\boldsymbol{S} - \mathbb{E}[\boldsymbol{S}\|\boldsymbol{\Omega}]\right]$ $= \mathbb{E}\left[\boldsymbol{Var}[\boldsymbol{S}\|\boldsymbol{\Omega}]\right]$

La variance $Var[S]$ est répartie entre l'assuré et l'assureur.

L'assureur connaît exactement le niveau de risque de chaque individu et l'utilise pour individualiser la prime, ce qui donne le résultat suivant :

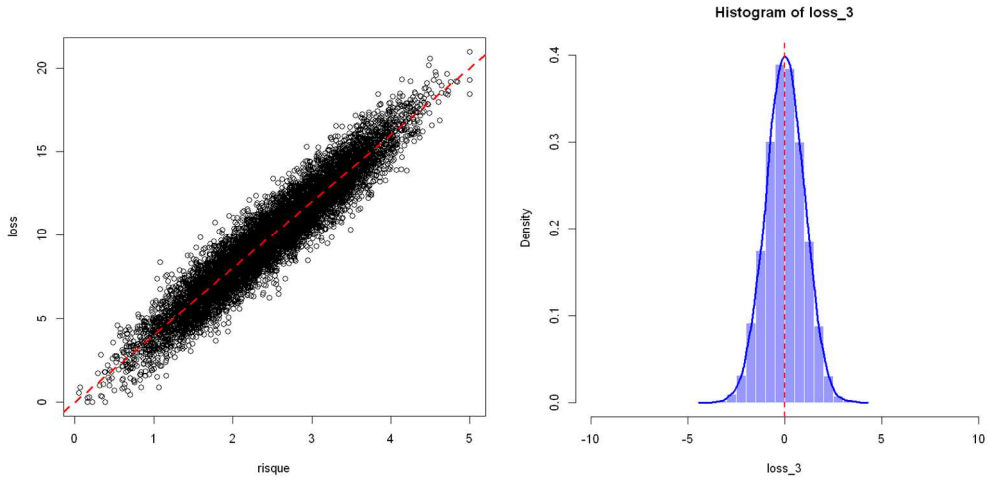


Figure 42 – Cas 2 : segmentation (théorique) à l'individu

Nous observons donc clairement une variance bien plus faible pour l'assureur.

Cas 3 : segmentation partielle

Si les facteurs observables ne permettent que d'avoir une vision partielle du risque, l'assureur va pouvoir segmenter son portefeuille par une classification.

	Assuré	Assureur
Perte	$\mathbb{E}[\boldsymbol{S}\|\boldsymbol{X}]$	$\boldsymbol{S} - \mathbb{E}[\boldsymbol{S}\|\boldsymbol{X}]$
Espérance de perte	$\mathbb{E}[\boldsymbol{S}]$	0
Variance	$\boldsymbol{Var}\left[\mathbb{E}[\boldsymbol{S}\|\boldsymbol{X}]\right]$	$\boldsymbol{Var}\left[\boldsymbol{S} - \mathbb{E}[\boldsymbol{S}\|\boldsymbol{X}]\right]$ $= \mathbb{E}\left[\boldsymbol{Var}[\boldsymbol{S}\|\boldsymbol{X}]\right]$

Avec *a priori* une partie de la variance plus importante portée par l'assureur du fait de la mutualisation restante, conséquence de la connaissance imparfaite des facteurs de risques et du choix de segmentation :

$$\mathbb{E}\left[Var[S|X]\right] = \mathbb{E}\left[\mathbb{E}[Var[S|\Omega]|X]\right] + \mathbb{E}\left[Var[\mathbb{E}[S|\Omega]|X]\right] = \underbrace{\mathbb{E}\left[Var[S|\Omega]\right]}_{(1)} + \underbrace{\mathbb{E}\left[Var[\mathbb{E}[S|\Omega]|X]\right]}_{(2)}$$

Le terme (1) correspond à la variance résiduelle du cas 2 avec une segmentation parfaite et le terme (2) représente la mutualisation issue de la segmentation partielle.

Considérons un assureur qui a l'information pour savoir si un assuré présente un niveau de risque supérieur ou inférieur à la moyenne. Il segmente donc son portefeuille selon :

- un tarif unique pour les individus au risque > 2,5 (prime pure sur ce périmètre) ;
- un second tarif unique pour les autres individus.

Ces deux primes sont représentées par le tracé rouge sur le schéma de gauche.

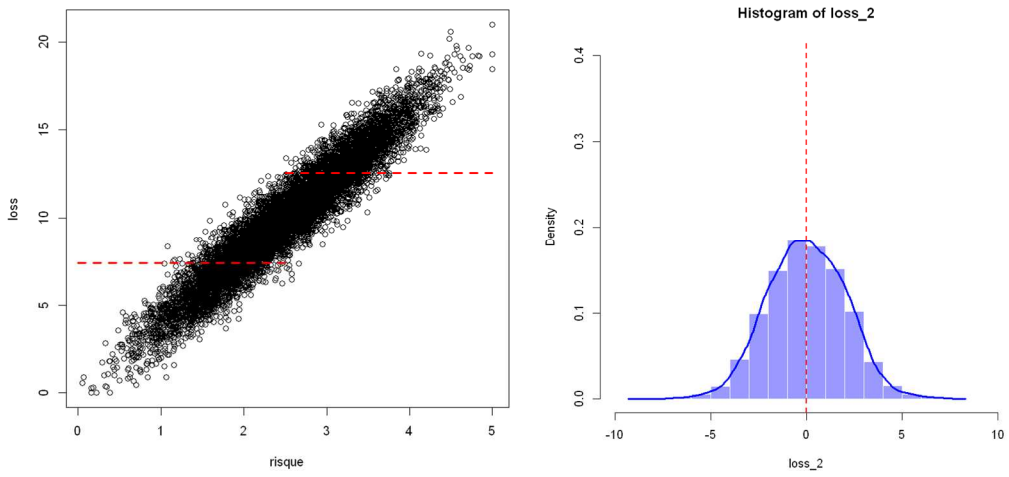


Figure 43 – Cas 3 : segmentation en deux tarifs

La variance de la perte de l'assureur se situe à un niveau entre le cas 1 et le cas 3.

Evolution du marché

En considérant que les assurés vont systématiquement choisir l'assureur qui leur propose la prime la moins élevée (choix rationnel), la répartition parmi les trois assureurs est la suivante :

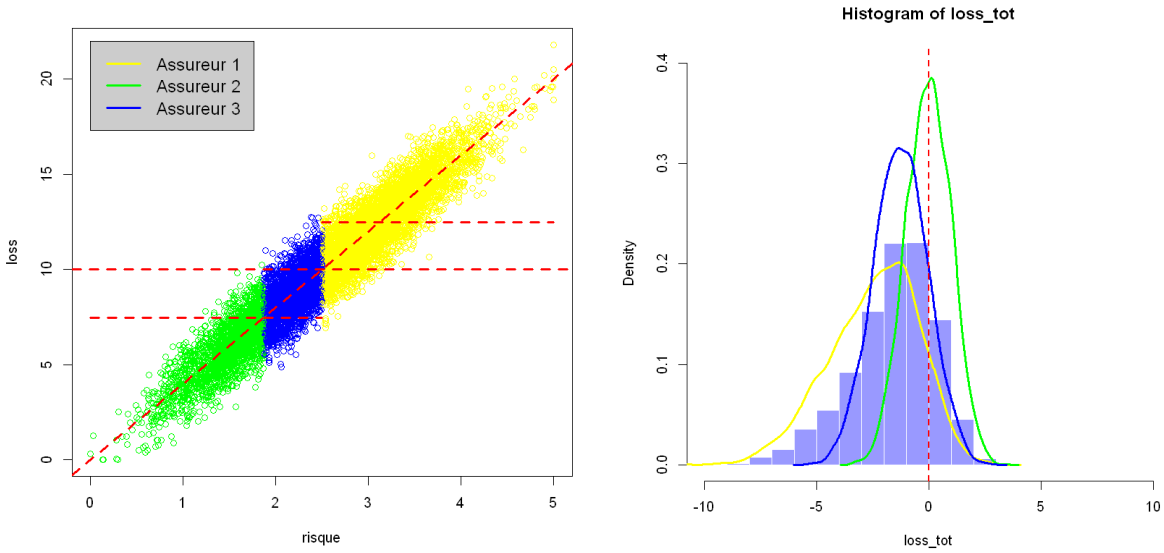


Figure 44 – Evolution du marché dans le cas d'un choix rationnel effectué par les assurés

Sur la figure de droite, dont l'histogramme et les densités représentent respectivement la perte tous assureurs confondus et celles pour chaque assureur, il est clair que le marché est déséquilibré. Il subit une destruction de valeur globale, qui est intégralement portée par les assureurs 1 et 3. Ce déséquilibre pousserait ces derniers à s'aligner sur la segmentation de l'assureur 2 et limiterait grandement la mutualisation proposée par ces contrats, assurant chaque assuré au regard de son risque propre.

La conséquence est à terme que, l'assureur ne prenant plus à sa charge qu'une petite partie de la variance totale (car les informations détenues par l'assureur lui permettent de segmenter à l'extrême selon les hypothèses de départ), l'intérêt de l'assurance peut être remis en question :

- la prime est faible pour l'individu ayant un profil non risqué mais, à moins d'avoir une très forte aversion au risque, ce même individu a probablement peu de raison de s'assurer ;
- pour les individus à risque, la prime d'assurance est si élevée qu'ils pourraient considérer que l'aléa autour de la perte future est préférable.

ANNEXE 4 : TRAITEMENT DES VALEURS ABERRANTES

Lorsqu'une base de données telle que la base Open DAMIR est analysée, il est nécessaire de traiter la problématique des valeurs manquantes mais également celle des valeurs aberrantes. En effet, des erreurs dans la collecte des données peuvent parfois engendrer des valeurs quantitatives extrêmes ou des valeurs négatives pour des grandeurs qui devraient logiquement rester positives.

Du point de vue des données, une valeur quantitative est considérée comme aberrante (ou atypique) lorsqu'elle prend une valeur inattendue ou peu vraisemblable compte-tenu de ce qu'elle représente, de l'expérience passée ou de l'avis d'expert. Du point de vue des statistiques, une valeur aberrante est définie comme une valeur qui s'éloigne trop de la distribution normale des données. En pratique, seront considérées comme aberrantes les valeurs qui s'écartent de la moyenne au-delà d'un certain nombre de fois l'écart-type.

Il est possible de représenter aisément ces phénomènes en utilisant une représentation de type « boîte à moustaches » :

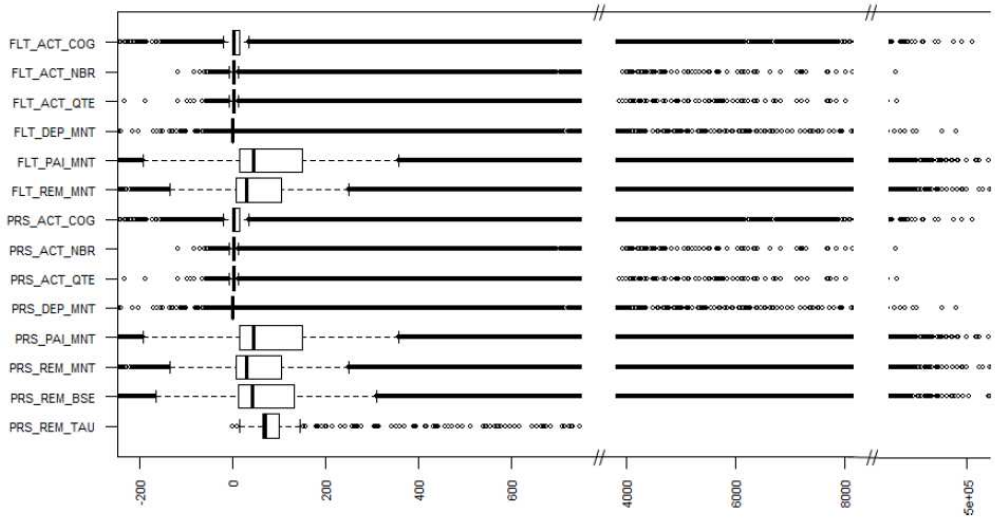


Figure 45 – Boîte à moustache des 14 variables quantitatives de la base Open DAMIR [MEKONTSO FOTSING, A. C. (2018)]

Il est clair que la base Open DAMIR présente à la fois des valeurs négatives et des valeurs très extrêmes. Pour autant, est-ce qu'il s'agit véritablement de données erronées et doivent-elles être corrigées ? En effet, les valeurs négatives peuvent être liées à des ajustements ou des corrections. Par ailleurs, étant donné que la base Open DAMIR présente des lignes agrégées, c'est-à-dire qu'une ligne représente un certain nombre de dépenses de santé distinctes, il est très difficile de juger si une valeur extrême est aberrante ou si elle est logique. Enfin, les valeurs extrêmes qui sortent des boîtes à moustaches (au-delà de 1,5 fois la distance interquartile) représentent plus de 10% des lignes de la base. Pour toutes ces raisons, nous avons considéré qu'il n'était pas pertinent de retraiter la base Open DAMIR en retirant ces valeurs aberrantes.

ANNEXE 5 : CONSTRUCTION D'UN MODELE STATISTIQUE

Qu'est-ce qu'un modèle statistique ?

Un modèle statistique correspond à une description mathématique d'un mécanisme ayant généré un certain nombre d'observations. Il s'agit par conséquent d'une représentation simplifiée de la réalité basée sur l'analyse des données observées. La représentation du mécanisme fera le plus souvent intervenir des variables aléatoires intégrées au modèle via un jeu de distributions de probabilités (explicites ou non).

Formellement, un modèle est défini par le couple $(\mathcal{D}, \mathcal{P})$ où $\mathcal{D}$ représente l'ensemble des données possibles et $\mathcal{P}$ l'ensemble des distributions de probabilités. Nous allons alors chercher à trouver une distribution $\mathbb{P} \in \mathcal{P}$ afin de représenter au mieux les données observées $\boldsymbol{X} \in \mathcal{D}$. Dans la plupart des cas, $\mathbb{P}$ est paramétrable via un vecteur de paramètres $\boldsymbol{\theta}$:

$$\mathbb{P} = \{\mathbb{P}_{\boldsymbol{\theta}} \, ; \, \boldsymbol{\theta} \in \Theta\}$$

où Θ représente l'ensemble des paramètres possibles du modèle. La recherche de $\mathbb{P}$ sera ainsi restreinte à la recherche des paramètres $\boldsymbol{\theta}$. Dans le cadre d'algorithmes de *machine learning* ce vecteur peut se décomposer en deux parties $\boldsymbol{\theta} = (\boldsymbol{\theta_I}, \boldsymbol{\theta}_H)$ où $\boldsymbol{\theta_I}$ sont les paramètres intrinsèques du modèle qui seront déterminés automatiquement par l'algorithme apprenant au cours d'une phase d'apprentissage et $\boldsymbol{\theta}_H$ représente les hyperparamètres qui déterminent la structure du modèle, indépendamment des données utilisées en entrée.

Un modèle statistique peut avoir trois objectifs distincts [KONISHI, S. et KITAGAWA, G. (2008)] :

- prédire des variables non observées ;
- construire un estimateur statistique de l'information contenue dans les données ;
- décrire le mécanisme stochastique sous-jacent.

Dans le cadre spécifique de ce mémoire, nous nous intéressons uniquement aux objectifs de prédiction et d'estimation. La sortie du modèle statistique $\widehat{\boldsymbol{Y}}$ sera ainsi une variable aléatoire qui devra être la plus proche possible de la variable réelle $\boldsymbol{Y}$.

Paramétrage du modèle

Si le modèle retenu est un modèle paramétrique, sa construction passera alors une phase d'apprentissage et une phase de validation permettant d'optimiser les paramètres par rapport aux données d'entrées et aux résultats attendus.

Concrètement, lors de la phase d'apprentissage, une base d'entraînement est utilisée afin d'optimiser la valeur des paramètres $\boldsymbol{\theta}$ du modèle afin que la sortie $\widehat{\boldsymbol{Y}}$ soit la plus proche

possible de la valeur $\boldsymbol{Y}$ observée. Dans le cadre d'algorithmes de type *machine learning*, c'est le modèle qui optimise lui-même ses paramètres intrinsèques $\boldsymbol{\theta_I}$ afin de correspondre à la base d'entraînement et le concepteur se contente de choisir[101] les hyperparamètres $\boldsymbol{\theta}_H$.

Une fois cette étape réalisée, il est nécessaire d'évaluer le modèle optimisé sur une base de validation indépendante de la base d'entraînement utilisée lors de la phase d'apprentissage. Cette étape de validation est nécessaire afin de vérifier que le paramétrage du modèle ne conduise pas à un phénomène de surapprentissage[102]. Elle va consister à vérifier la qualité statistique du modèle sur de nouvelles données, généralement en analysant l'erreur $\boldsymbol{\varepsilon}$ entre la sortie du modèle $\widehat{\boldsymbol{Y}}$ et la valeur attendue $\boldsymbol{Y}$.

Les phases d'apprentissage et de validation consistent ainsi à évaluer les performances du modèle afin de choisir les paramètres qui optimisent sa qualité statistique. Elles nécessitent la construction d'une base d'entraînement et d'une base de validation. La véritable valeur de $\boldsymbol{Y}$ n'étant évidemment pas connue dans le cas général (sinon le modèle n'aurait pas d'intérêt), afin d'évaluer les performances du modèle, il est nécessaire de construire une base de sorte à avoir accès à la donnée de sortie $\boldsymbol{Y}$: c'est la base de données complètes.

Validation croisée

Une fois cette base de données complètes constituée, il s'agit ensuite de voir comment la répartir entre le jeu d'entraînement et le jeu de validation. La méthode la plus évidente consiste à séparer simplement la base en deux : une partie servira de jeu d'entraînement et l'autre aura le rôle de jeu de validation[103]. Néanmoins, cette méthode ne permet d'évaluer le modèle que sur un unique résultat, ce qui peut remettre en cause le résultat lorsque le modèle est stochastique.

Pour pallier ce problème, il est généralement préférable d'utiliser une méthode de validation plus fiable : la validation croisée[104]. Cette méthode consiste à diviser la base des cas complets en k sous-bases (typiquement $k = 10$ [MC LACHLAN, G. J. et al. (2004)]). Chaque sous-base donne lieu à une validation au cours de laquelle la base d'entraînement est constituée des $k - 1$ sous-bases restantes. Les résultats sont ensuite agrégés via la moyenne ou le maximum.

---

[101] La plupart du temps, le choix des hyperparamètres d'un modèle de *machine learning* s'effectue via un mélange subtil de théorie, de bonnes pratiques et d'heuristique.

[102] Il y a surapprentissage (ou *overfitting*) lorsque le modèle correspond de manière trop parfaite au jeu d'entraînement à tel point qu'il intégrera non seulement l'information pertinente inclue dans la base de données d'entraînement mais également les éventuels bruits ou effets extrêmes. La conséquence est que l'algorithme sera alors très bon s'il lui est proposé exactement les mêmes entrées que celles présentent dans la base d'entraînement mais donnera de mauvais résultats dès lors que les entrées seront différentes (même très légèrement).

[103] L'usage choisit généralement environ deux tiers des données pour l'apprentissage de l'algorithme et le tiers restant pour la validation.

[104] Il existe de nombreuses méthodes de validation croisée différentes. Dans le cadre de ce mémoire, nous nous intéressons uniquement à la méthode de validation croisée à k blocs (ou *k-fold cross validation*).

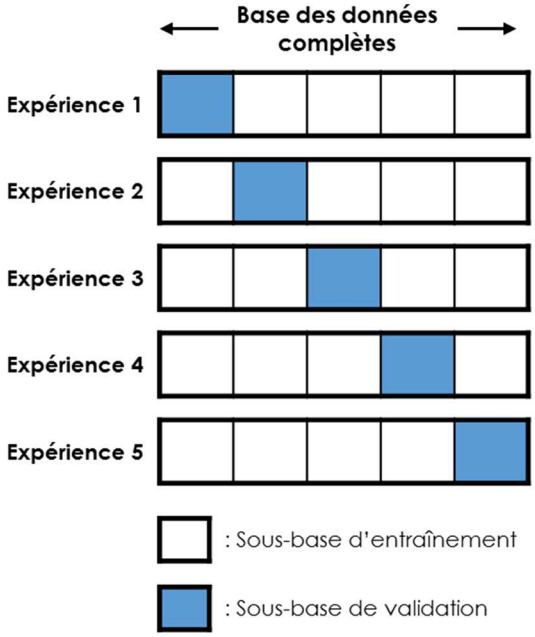


Figure 46 – Schématisation de la validation croisée

Evaluation de la qualité statistique du modèle

Il existe différentes manières de mesurer la qualité statistique d'un modèle selon l'objectif du traitement réalisé par le modèle. La mesure la plus évidente consiste à regarder l'erreur de prédiction en évaluant la distance (la définition précise de la distance utilisée sera vu au § II.1.3.2.) entre les valeurs prédites par le modèle et les valeurs réelles. Cette méthode permet de mesurer à la fois le biais et l'intervalle de confiance fournis par le modèle.

La mesure de la qualité statistique sera utilisée non seulement aux étapes d'apprentissage et de validation croisée mais également à l'étape de simulation qui permettra de déterminer la performance du modèle optimisé.

Dans le cas spécifique du modèle d'imputation des valeurs manquantes, nous obtenons ainsi le schéma suivant :

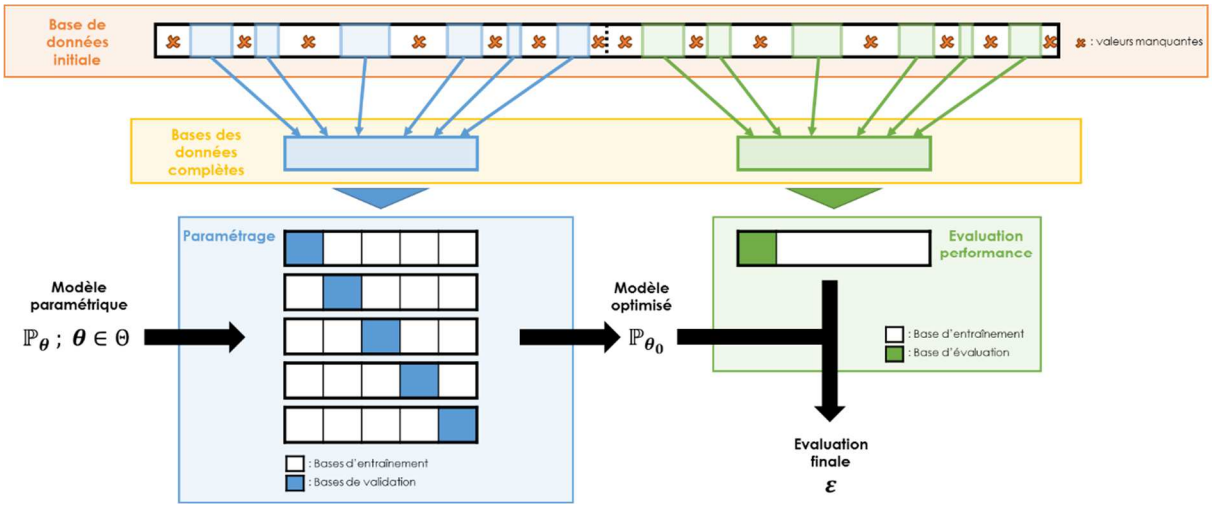


Figure 47 – Schématisation de l'évaluation de la performance du modèle

Une fois ces résultats obtenus pour l'ensemble des modèles envisagés, il est ainsi possible de les comparer sur différents critères. La qualité statistique du modèle sera bien évidemment le critère prépondérant mais selon les circonstances, il peut en exister d'autres. En particulier dans le cas du *big data*, il est nécessaire de prendre en compte la capacité du modèle de traiter de grands volumes de données dans des temps de réponses raisonnables tout en respectant les capacités de la machine en termes de mémoire vive.

ANNEXE 6 : CREATION DE VALEURS MANQUANTES ARTIFICIELLES

Afin de valider rigoureusement le modèle d'imputation des valeurs manquantes et d'évaluer correctement sa performance, il est nécessaire de pouvoir disposer d'une base de validation présentant des valeurs manquantes cohérentes par rapport à la distribution réelle d'apparition des données manquantes. Pour cela, il nous faut construire un modèle dont l'objectif est de détecter à partir d'une base de données sans aucune donnée manquante les emplacements des valeurs manquantes. Afin de répondre à cette problématique, nous avons ainsi envisagé deux modèles différents.

Modèle MCAR : création de valeurs manquantes aléatoires

Il s'agit du cas le plus simple à réaliser mais qui ne fonctionne que dans le cas de données MCAR : pour chaque variable $\boldsymbol{y_j}$ (avec $j \in [\![1, p]\!]$), il suffit de calculer le taux de valeurs manquantes π_j au niveau de la base réelle puis de réaliser un tirage aléatoire selon un processus de Bernoulli[105] selon les π_j sur la base de validation. Les données ainsi tirées seront alors considérées comme manquantes.

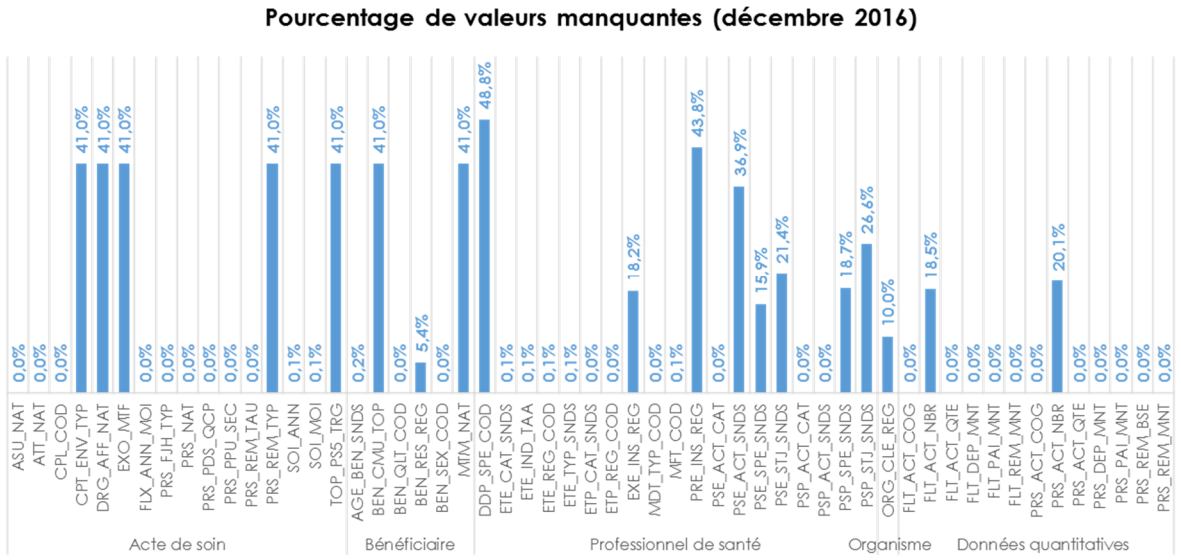


Figure 48 – Proportion de valeurs manquantes par variable dans la base Open DAMIR (décembre 2016)

Bien qu'il s'agisse d'un modèle évident et simple à construire, nous sentons bien que ce modèle ne permettra pas de représenter fidèlement l'apparition des valeurs manquantes au sein de la base Open DAMIR dès lors que celle-ci n'est pas MCAR. Il est donc nécessaire d'envisager la construction d'un second modèle plus évolué.

[105] Pour chaque donnée y_{ij} du vecteur $\boldsymbol{y_j} = (y_{1j}, y_{2j}, \dots, y_{nj})$, il y a une probabilité π_j de tirer y_{ij} (et donc logiquement une probabilité $1 - \pi_j$ de ne pas tirer y_{ij}).

Modèle MAR : base de validation avec corrélations

Patterns de valeurs manquantes

La construction du jeu de validation est plus complexe lorsque les données manquantes sont MAR car il faut s'assurer de respecter les corrélations entre les variables observées et les données manquantes. Or celles-ci ne sont pas nécessairement connues[106]. Il est néanmoins possible de générer plus simplement un jeu de validation avec des données MAR en travaillant autour des « *patterns* » des valeurs manquantes [BRAND, J. P. L. (1999)].

Nous appelons ici « *pattern* » le schéma de répartition des données manquantes au sein des différentes variables de la base. Par exemple, en considérant uniquement deux variables $\boldsymbol{x}$ et $\boldsymbol{y}$, il existe quatre *patterns* différents :

- $\boldsymbol{x}$ et $\boldsymbol{y}$ sont tous les deux observées (*pattern* 00) ;
- $\boldsymbol{x}$ est observé et $\boldsymbol{y}$ est manquante (*pattern* 01) ;
- $\boldsymbol{x}$ est manquante et $\boldsymbol{y}$ est observée (*pattern* 10) ;
- $\boldsymbol{x}$ et $\boldsymbol{y}$ sont tous les deux manquantes (*pattern* 11).

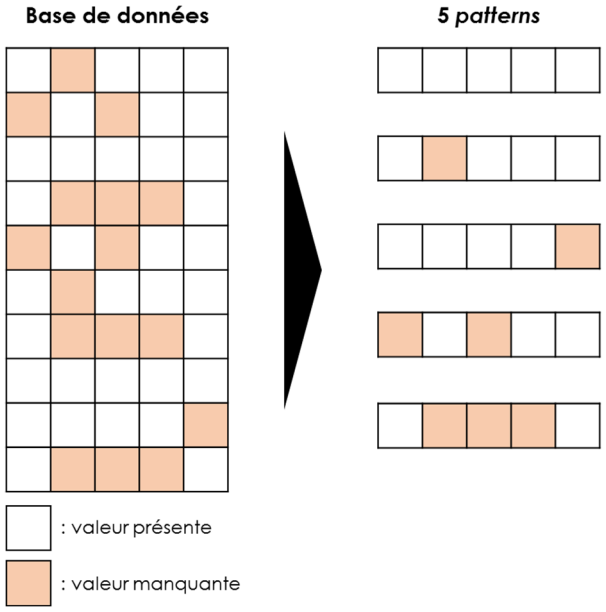


Figure 49 – Schématisation des patterns de valeurs manquantes

[106] En toute rigueur, il faudrait estimer la loi de distribution des valeurs manquantes à l'aide d'une analyse multivariée afin de pouvoir de pouvoir la simuler pour créer artificiellement les valeurs manquantes. Cependant, une telle analyse pourrait faire l'objet d'un mémoire entier.

Principe général

L'ensemble des *patterns* possibles est ainsi enregistré au sein d'une matrice binaire $\boldsymbol{R_{pat}} = \left(r_{pat_{ij}}\right)$, de dimension $n_{pat} \times p$ où n_{pat} est le nombre de *patterns* existant dans la base[107] et p est le nombre de variables avec pour valeur $r_{pat_{ij}} = 1$ si la donnée est manquante et $r_{pat_{ij}} = 0$ sinon. Chaque *pattern* $\tau \in [\![1, n_{pat}]\!]$ possède parmi l'ensemble des lignes incomplètes une fréquence relative d'apparition f_τ au sein de la base de données et un modèle de probabilité de réponse spécifique $\mathbb{P}\left[\boldsymbol{R_{pat}^{(\tau)}}|\boldsymbol{Y}\right]$[108].

L'algorithme de création de valeurs manquantes au sein de la base de données $\boldsymbol{Y}$ (qui est pour rappel de dimensions $n \times p$) est alors le suivant :

- Pour chaque ligne $i \in [\![1, n]\!]$, un tirage de la loi multinomiale de fréquences $f_1, f_2, \dots, f_{n_{pat}}$ va sélectionner le *pattern* τ des valeurs manquantes envisagé.
- En notant $\alpha = \frac{\sum_{i=1}^{n} \mathbf{1}_{\left\{\sum_{j=1}^{p} r_{ij} \geq 1\right\}}}{n}$ la proportion de lignes incomplètes au sein de la base, nous allons appliquer le modèle du pattern $\mathbb{P}\left[\boldsymbol{R_{pat}^{(\tau)}}|\boldsymbol{Y}\right]$ avec la probabilité α.

Nous obtenons ainsi $f_\tau \times n$ candidats pour le *pattern* τ et $f_\tau \times n \times \alpha$ lignes incomplètes avec le *pattern* τ. Comme $\sum_{\tau=1}^{n_{pat}} f_\tau = 1$ par construction, le nombre total de lignes incomplètes sera $n \times \alpha$ et donc la proportion de lignes incomplètes au sein de la base sera bien respectée.

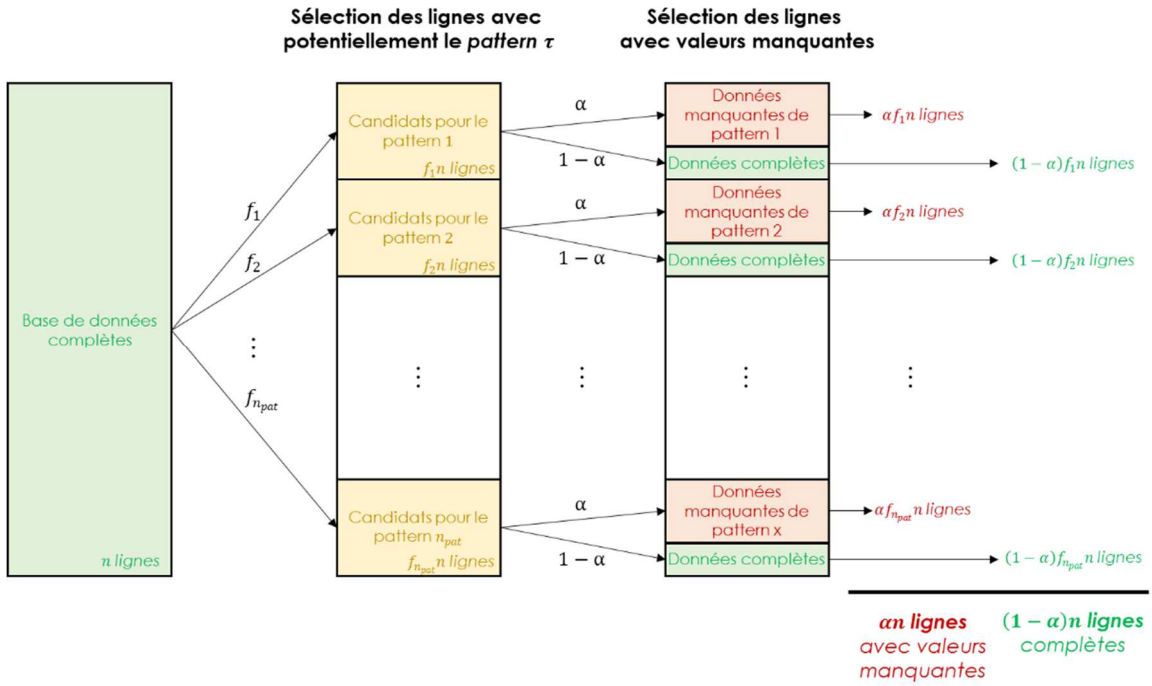


Figure 50 – Modèle de création des valeurs manquantes au sein de la base complète

[107] La valeur maximum de n_{pat} est donc 2^p.

[108] Pour des raisons de simplicité, nous modéliserons les probabilités de réponse spécifiques en utilisant une simple combinaison linéaire d'un petit nombre de variables qualitatives observées bien choisies.

Sélection des variables explicatives

Afin de déterminer les variables explicatives les plus intéressantes (pour des raisons de simplicité, nous ne considérons ici que les variables qualitatives[109]) pour construire nos modèles de probabilité de réponse, en plus de la proportion de valeur manquantes, nous allons définir deux indicateurs nous permettant de quantifier l'utilité d'une variable dans l'explication de l'apparition d'une valeur manquante. Le premier a été introduit par Stef van Buuren [VAN BUUREN, S. (2018)] ; nous proposons le second pour les besoins spécifiques de notre mémoire.

Nous considérons un $j \in [\![1, p]\!]$ fixé de sorte que la variable $\boldsymbol{y_j}$ soit une variable qualitative. L'*outflux* O_j de la variable $\boldsymbol{y_j}$ représente le nombre de paires de variables $(\boldsymbol{y_j}, \boldsymbol{y_k})$ avec $k \in [\![1, p]\!]$ pour lesquels la donnée sur $\boldsymbol{y_j}$ est observée alors que celle sur $\boldsymbol{y_k}$ est manquante par rapport au nombre total de valeurs manquantes :

$$O_j = \frac{\sum_{k=1}^{p} \sum_{i=1}^{n} r_{ij}(1 - r_{ik})}{\sum_{k=1}^{p} \sum_{i=1}^{n} (1 - r_{ik})}$$

L'*outflux* dépend de la proportion de valeurs manquantes : si la variable est toujours observée, celui-ci vaudra 1 et inversement, si la donnée est systématiquement manquante, l'*outflux* sera nul. En considérant deux variables avec la même proportion de valeurs manquantes, alors la variable avec l'*outflux* le plus important sera mieux connecté aux données manquantes et sera donc potentiellement plus utile pour expliquer leur apparition. En appliquant cette définition à la base Open DAMIR, nous trouvons les résultats suivants :

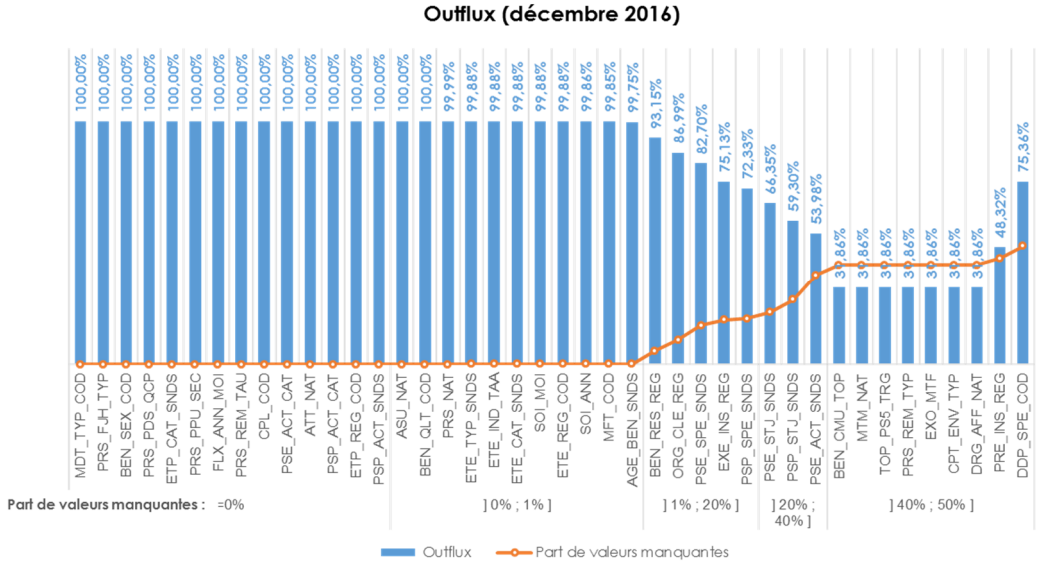


Figure 51 – Valeur de l'outflux et de la proportion des valeurs manquantes par variable qualitative

[109] Cette hypothèse reste raisonnable dans la mesure où les données manquantes concernent principalement ces variables et que la base Open DAMIR est administrative et non déclarative (les montants n'ont donc pas de raison d'entraîner l'apparition de données manquantes).

Sans surprise, l'*outflux* est dans la grande majorité expliqué par la proportion de valeurs manquantes : plus celle-ci est faible et plus l'*outflux* est important. Deux variables se démarquent néanmoins grâce à cette analyse avec une valeur d'*outflux* supérieure à celle de variables présentant moins de données manquantes : PRE_INS_REG (région du professionnel de santé prescripteur) et DPP_SPE_COD (discipline de prestation de l'établissement de santé exécutant).

Nous définissons par ailleurs l'indicateur puissance de dispersion Π_j de la variable $\boldsymbol{y_j}$ qui sert à mesurer la dispersion du nombre de valeurs manquantes au sein des autres variables $\boldsymbol{Y_{-j}}$ selon les valeurs prises par $\boldsymbol{y_j}$. Ainsi, si considère la variable $\boldsymbol{y_j}$ peut prendre L_j valeurs parmi $\left\{y_j^{(1)}, y_j^{(2)}, \dots, y_j^{(L_j)}\right\}$ (l'une de ces valeurs pouvant correspondre éventuellement au cas d'une donnée manquante), alors pour $k \in [\![1, p]\!]$, il est possible de déterminer l'écart-type σ_{jk} de la proportion de valeurs manquantes de la variable $\boldsymbol{y_k}$ selon les valeurs prises par la variable $\boldsymbol{y_j}$:

$$\sigma_{jk} = \sqrt{\frac{1}{L_j}\sum_{l=1}^{L_j}\left(\frac{n_{kj}^{miss(l)}}{n_j^{(l)}} - \frac{n_k^{miss}}{n}\right)^2}$$

avec :

- $n_k^{miss} = \sum_{i=1}^{n}(1 - r_{ik})$ le nombre de valeurs manquantes de $\boldsymbol{y_k}$;
- $n_j^{(l)} = \sum_{i=1}^{n} \mathbf{1}_{\left\{y_{ij} = y_j^{(l)}\right\}}$ le nombre de lignes $i \in [\![1, n]\!]$ telles que $y_{ij} = y_j^{(l)}$;
- $n_{kj}^{miss(l)} = \sum_{i=1}^{n}\left[\mathbf{1}_{\left\{y_{ij} = y_j^{(l)}\right\}} \times (1 - r_{ik})\right]$ le nombre de valeurs manquantes de $\boldsymbol{y_k}$ parmi les lignes $i \in [\![1, n]\!]$ telles que $y_{ij} = y_j^{(l)}$.

Ainsi, σ_{jk} sera nul si toutes les valeurs de la variable $\boldsymbol{y_j}$ induisent la même proportion de données manquantes sur la variable $\boldsymbol{y_k}$ et inversement sera maximale si les apparitions des données manquantes de $\boldsymbol{y_k}$ sont entièrement déterminées par les valeurs de $\boldsymbol{y_j}$.

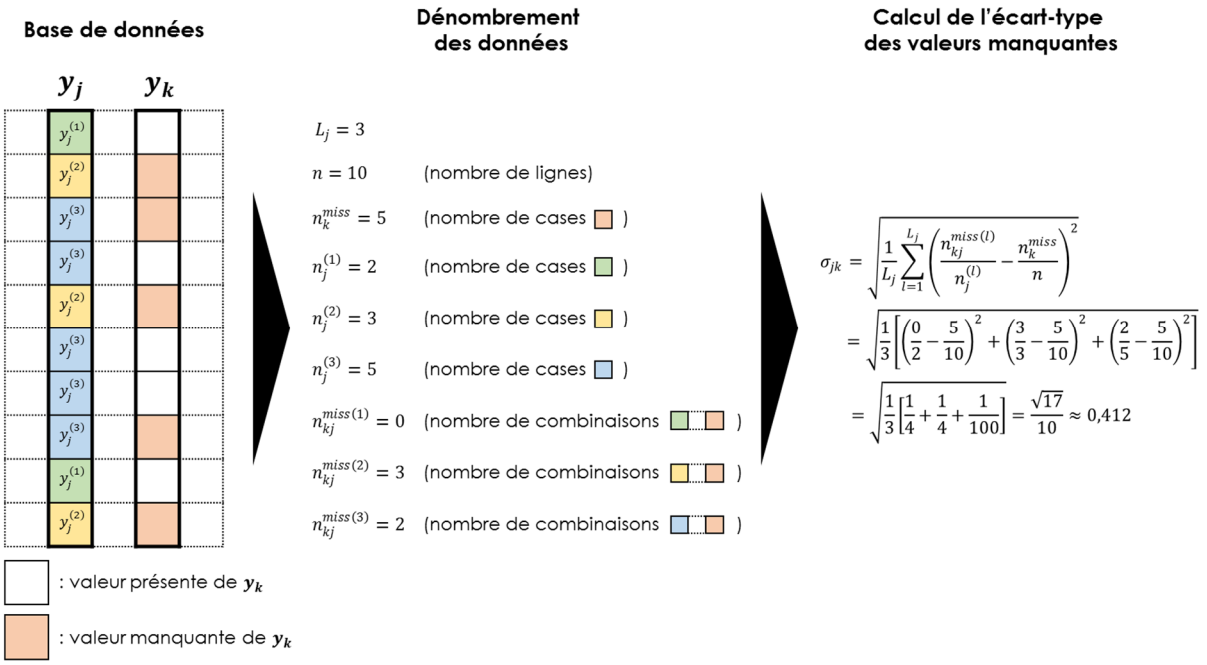


Figure 52 – Schématisation du calcul de l'écart-type lié des valeurs manquantes

La puissance de dispersion Π_j est alors la moyenne des σ_{jk} pour toutes les variables $\boldsymbol{y_k}$ différentes de $\boldsymbol{y_j}$:

$$\Pi_j = \frac{1}{p-1}\sum_{\substack{k=1 \\ k \neq j}}^{p} \sigma_{jk}$$

Cet indicateur permet ainsi d'identifier les variables les plus différenciantes pour l'apparition des données manquantes.

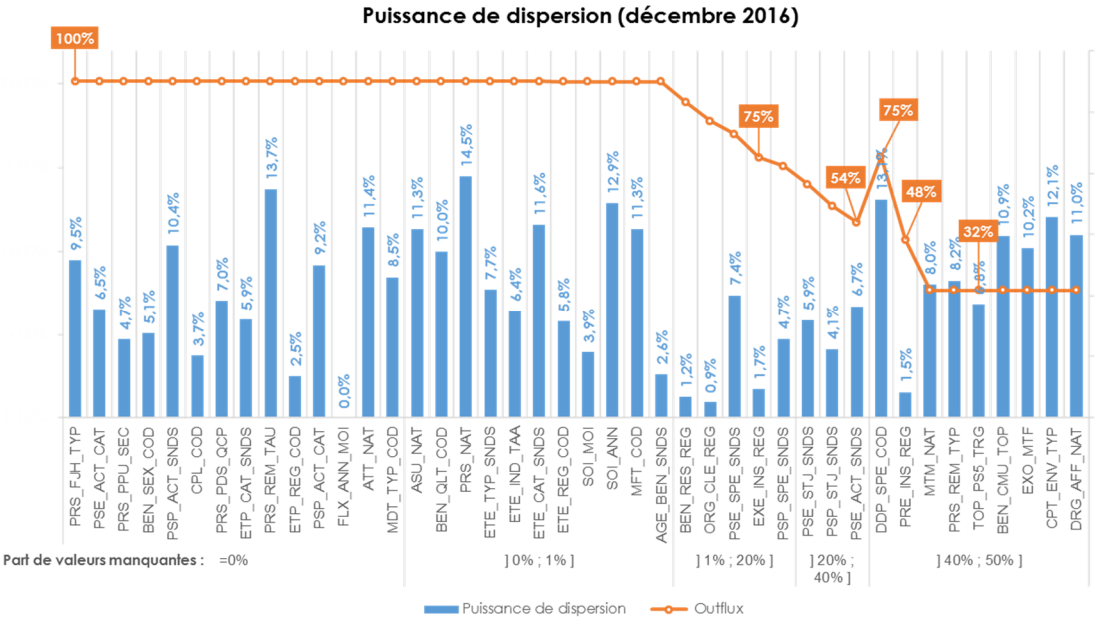


Figure 53 – Valeur de la puissance de dispersion et de l'outflux par variable qualitative

En combinant ces deux indicateurs, il est ainsi possible de sélectionner les variables explicatives les plus pertinentes et de construire un modèle simple de distribution des valeurs manquantes $\mathbb{P}\left[\boldsymbol{R}_{\boldsymbol{pat}}^{(\tau)}|\boldsymbol{Y}^{(\tau)}\right]$ pour le *pattern* τ. Nous voyons en effet, comme nous pouvions nous y attendre, que la principale variable explicative est PRS_NAT[110] (nature de la prestation). L'analyse permet également de mettre en lumière quatre autres variables explicatives :

- PRS_REM_TAU[111] (taux de remboursement du régime obligatoire) ;
- DPP_SPE_COD[112] (discipline de l'établissement de soin exécutant) ;
- SOI_ANN (année de l'acte de soin) ;
- ETE_CAT_SNDS[113] (catégorie de l'établissement de soin exécutant).

Afin de simplifier l'élaboration des modèles $\mathbb{P}\left[\boldsymbol{R}_{\boldsymbol{pat}}^{(\tau)}|\boldsymbol{Y}^{(\tau)}\right]$, nous proposons en conclusion de retenir les deux variables explicatives PRS_NAT et SOI_ANN

Elaboration du modèle

La première étape d'élaboration du modèle de création des valeurs manquante consiste à identifier les différents *pattern* τ existant au sein de la base Open DAMIR avec les fréquences empiriques f_τ associées (en retirant les lignes complètes de l'assiette). Nous trouvons ainsi 1 170 patterns différents.

Nous avons identifié les variables explicatives $\boldsymbol{y}_{\boldsymbol{exp}}^{(\boldsymbol{1})}$ et $\boldsymbol{y}_{\boldsymbol{exp}}^{(\boldsymbol{2})}$. Nous calculons ainsi les fréquences empiriques $f_l^{(\tau)}$ du *pattern* τ selon la combinaison l des valeurs de $\boldsymbol{y}_{\boldsymbol{exp}}^{(\boldsymbol{1})}$ et $\boldsymbol{y}_{\boldsymbol{exp}}^{(\boldsymbol{2})}$.

Sur la base des $n = 134\,476$ lignes complètes, nous sélectionnons donc les $n \times f_\tau$ lignes candidates au *pattern* τ selon un processus de Bernoulli. Parmi les lignes obtenues, nous effectuons un second processus de Bernoulli avec les probabilités $f_l^{(\tau)} \times \alpha$ (où pour rappel $\alpha = 99{,}57\%$ représente la proportion de lignes incomplètes au sein de la base). Dit plus simplement, le facteur α détermine le nombre de lignes complètes et les probabilités $f_l^{(\tau)}$ vont permettre de sélectionner ces lignes parmi toutes les lignes candidates au *pattern* τ.

[110] Cela justifie *a posteriori* le choix de cette variable pour l'analyse effectuée au § II.1.2.2.
[111] Cependant, cette donnée présente en réalité une information redondante avec la donnée PRS_NAT car le taux de remboursement dépend logiquement de la nature de l'acte de soin.
[112] Malheureusement, en creusant la piste de la donnée DPP_SPE_COD, nous remarquons que 49% des données sont manquantes et 48% sont sans objet ; ce qui ne laisse que 3% des lignes contenant une réelle information.
[113] Etant donné que cette variable a subi des modifications dans le cadre des traitements préalables, il est délicat de la conserver en tant que variable explicative.

Résultats des modèles

Les deux modèles MCAR et MAR de construction de la base de validation ont été implémentés sous Python. Comme dans les deux cas il ne s'agit pas d'un algorithme de type *machine learning*, il n'y a pas besoin d'étape d'apprentissage ou de validation.

Par ailleurs, il est extrêmement délicat de quantifier la performance de ces modèles créant des valeurs manquantes artificielles. Il est donc proposé d'évaluer et de comparer la performance de nos modèles via un contrôle visuel sur la répartition des valeurs manquantes obtenue.

En regardant la fréquence d'apparition des valeurs manquantes sur l'ensemble des lignes, nous obtenons les résultats suivants :

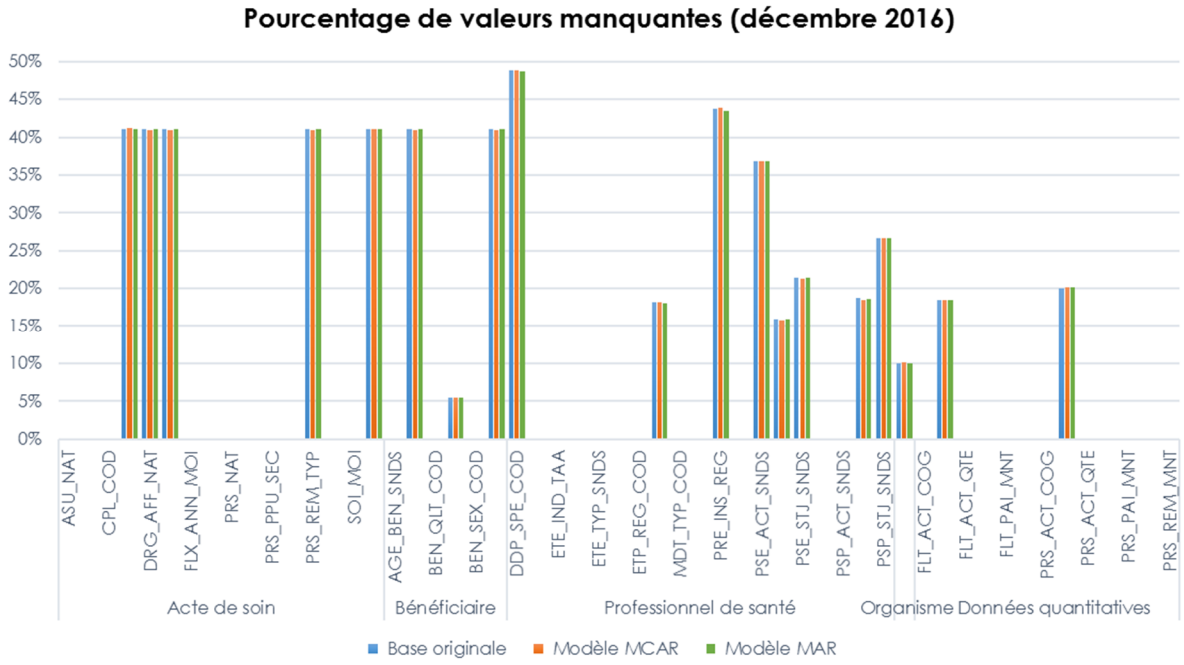


Figure 54 – Comparaison de la proportion de valeurs manquantes des modèles MCAR et MAR

Nous voyons que la fréquence d'apparition des données manquantes par variable est conforme à la base Open DAMIR originale pour les deux modèles, ce qui est logique car les deux ont été construits afin de respecter cette contrainte[114].

Afin de conclure, il est donc nécessaire de regarder la répartition des valeurs manquantes lignes par lignes. Ainsi, en regardant les 500 premières lignes, nous obtenons les résultats suivants :

[114] Les légers écarts constatés sont imputables à l'aspect stochastique des modèles.

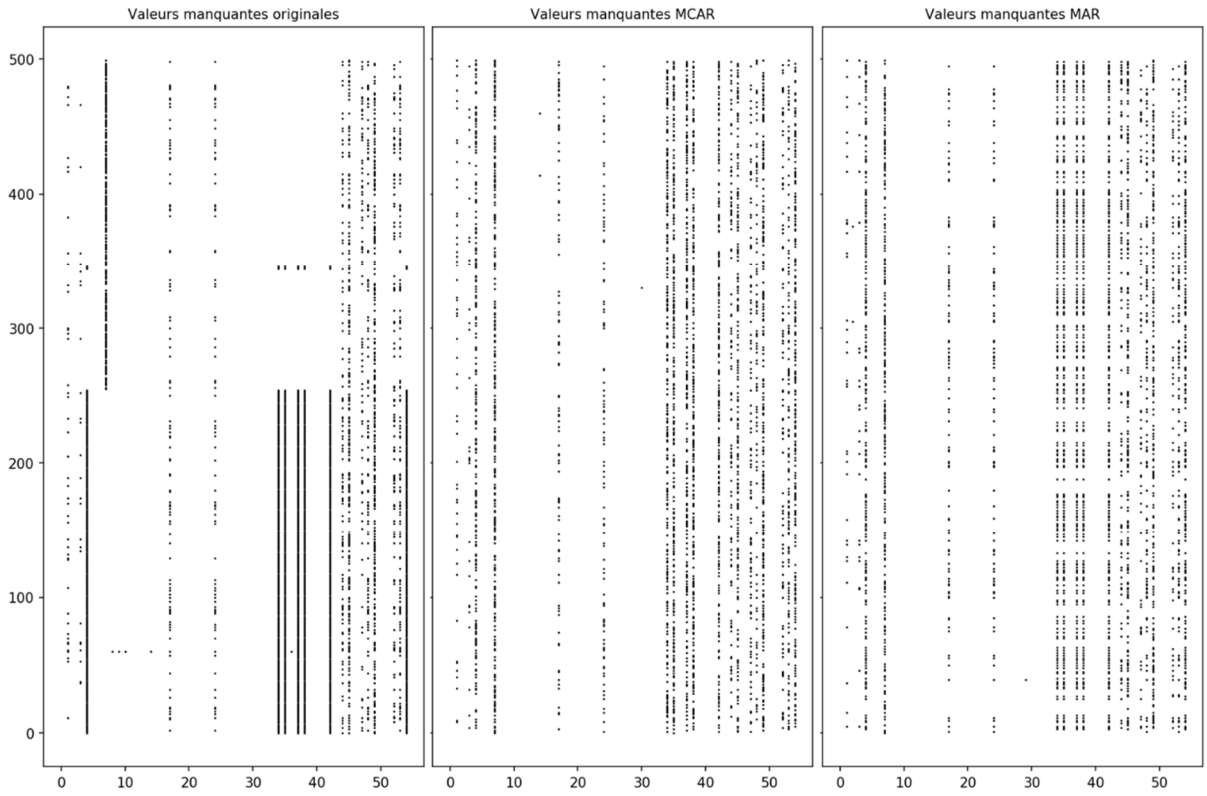


Figure 55 – Comparaison de la répartition des valeurs manquantes des modèles MCAR et MAR

Sans surprise, nous constatons que les corrélations existantes entre les variables dans l'apparition des données manquantes sont perdues dans le modèle MCAR alors qu'elles apparaissent bien au sein du modèle MAR.

Nous retiendrons donc sans surprise le modèle MAR pour créer nos bases d'entrainement et de validation pour la construction de nos modèles d'imputation des données manquantes.

ANNEXE 7 : DESCRIPTION DES ALGORITHMES UTILISES

Listwise deletion

L'analyse des cas complets (ou *listwise deletion*) consiste à considérer uniquement les observations pour lesquelles toutes les données sont disponibles en supprimant les lignes présentant des valeurs manquantes. Il s'agit du traitement par défaut effectué dans de nombreux cas (parfois sans être précisé).

L'avantage évident de cette méthode réside dans sa simplicité d'implémentation. Par ailleurs, si les données sont MCAR, alors l'analyse des cas complets produira des estimateurs non biaisés pour la moyenne et la variance [LITTLE, R. J. A. et RUBIN, D. B. (2002)].

Cependant, la méthode présente de nombreux inconvénients. Le premier est qu'il gâche une grande partie des informations contenues dans la base (parfois au-delà de la moitié des données initiales sont supprimées et cela peut atteindre 90% dans certaines études [KING, G. et al. (2001)]). De plus, dès que les données ne sont pas MCAR, l'analyse des cas complets biaisera les estimateurs produits [SCHAFER, J. L. et GRAHAM, J.W. (2002)].

Imputation par combinaison linéaire des observations

Cette méthode d'imputation consiste à remplacer les valeurs manquantes présentent sur une variable quantitative par une combinaison linéaire des valeurs observées sur cette variable. L'imputation par la moyenne et l'imputation par la médiane sont les méthodes les plus fréquemment utilisées.

Il s'agit d'une méthode extrêmement simple à mettre en œuvre. L'imputation par la moyenne donnera un estimateur sans biais de la moyenne [MEKONTSO FOTSING, A. C. (2018)] (et respectivement pour l'imputation par la médiane). En revanche, elle aura tendance à sous-estimer la variance, à biaiser tous les autres estimateurs et à perturber les relations entre les variables dès lors que les données ne sont pas MCAR. En outre, cette méthode n'a de sens que pour les variables quantitatives.

CMCF

La méthode CMCF (*Concept Most Common Attribute Value Fitting*) ou imputation par le mode correspond à l'équivalent de l'imputation par la moyenne pour les variables qualitatives. Il s'agit simplement de remplacer les données manquantes sur une variable par la valeur la plus fréquemment représentée au sein des observations. Elle présente également les mêmes avantages et inconvénients que l'imputation par la moyenne.

kNN (k Nearest Neighbors)

L'algorithme kNN [TROYANSKAYA, O. et al. (2001)] est une méthode de type *hot deck*[115] : la prédiction du modèle se base sur les valeurs trouvées sur des lignes similaires. Nous entendons par ligne similaire une ligne qui partage des valeurs identiques ou proches sur un certain nombre de variables explicatives. Concrètement, l'algorithme va regarder les k plus proches voisins de la ligne considérée en entrée. Afin de déterminer ces derniers, il est nécessaire de définir une mesure de distance[116] appropriée à la base de données. Usuellement, sont utilisées la distance euclidienne[117] lorsque les données sont quantitatives et la distance de Hamming[118] lorsque les données sont qualitatives. Lorsque la base de données est mixte, il est possible d'utiliser comme distance la combinaison d'une distance quantitative et d'une distance qualitative [MEKONTSO FOTSING, A. C. (2018)]. Une fois que les k lignes les plus proches de la ligne présentant la donnée à imputer ont été identifiées, il suffit d'agréger les résultats qu'elles contiennent, soit via la moyenne[119] (pour une variable quantitative), soit via le mode (pour une variable qualitative) pour obtenir la sortie du modèle.

L'algorithme kNN ne nécessite pas de modéliser la distribution des données et son emploi est adapté aux bases mixtes [SCHWENDER, H. (2012)]. Par ailleurs, il est très simple à mettre en œuvre puisque les seuls paramètres à déterminer sont le nombre de voisins k à considérer et la mesure de distance à utiliser. Enfin, cette méthode donne généralement de très bons résultats. Toutes ces raisons font que la méthode kNN est particulièrement populaire, notamment pour l'imputation de valeurs manquantes.

CART

Parmi les différents algorithmes de *machine learning*, la technique des forêts aléatoires (ou *random forest*) est particulièrement populaire. Afin de pouvoir expliquer le concept de forêt aléatoire, il est avant tout nécessaire de définir le concept d'arbre de décision. Popularisé par Breiman [BREIMAN, L. et al. (1984)] dans son algorithme CART (*Classification and Regression Random Tree*), un arbre de classification (pour les données qualitatives) ou de régression (pour les données quantitatives) est un modèle de description d'une variable d'une base de données construit de manière récursive basé sur une division en deux d'une variable explicative afin de produire deux sous-ensembles le plus homogène possible. Cet algorithme est qualifié d'arbre car sa représentation

[115] L'expression « *hot deck* » fait référence aux cartes perforées utilisées pour le stockage de l'information et signifie que l'information utilisée pour l'imputation provient directement des données, par opposition aux méthodes d'imputation « *cold deck* » (non étudiées dans le cadre de ce mémoire) qui consistent à se servir des informations contenues dans d'autres bases de données.

[116] Formellement, une distance d sur un ensemble E entre deux points x et y est une application de $E \times E$ dans $\mathbb{R}_+$ vérifiant les propriétés de symétrie ($d(x,y) = d(y,x)$), de séparation ($d(x,y) = 0 \Leftrightarrow x = y$) et d'inégalité triangulaire ($d(x,z) \leq d(x,y) + d(y,z)$).

[117] La distance euclidienne $d_E(\boldsymbol{x},\boldsymbol{y})$ entre deux vecteurs $\boldsymbol{x}(x_1, x_2, \dots, x_n)$ et $\boldsymbol{y}(y_1, y_2, \dots, y_n)$ est la racine de la somme des écarts au carré : $d_E(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$.

[118] La distance de Hamming $d_H(\boldsymbol{x},\boldsymbol{y})$ entre deux vecteurs $\boldsymbol{x}(x_1, x_2, \dots, x_n)$ et $\boldsymbol{y}(y_1, y_2, \dots, y_n)$ est le nombre de cas où $x_i \neq y_i$ pour $i \in [\![1, n]\!]$: $d_H(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n} \mathbf{1}_{\{x_i \neq y_i\}}$.

[119] Éventuellement pondéré par l'inverse des distances.

graphique évoque la forme d'un arbre inversé : la base complète est la racine ; chaque nœud va donner naissance à deux branches[120] représentant les valeurs de la variable explicative considérée ; les nœuds terminaux[121] représentent les feuilles et contiennent les différentes valeurs possibles.

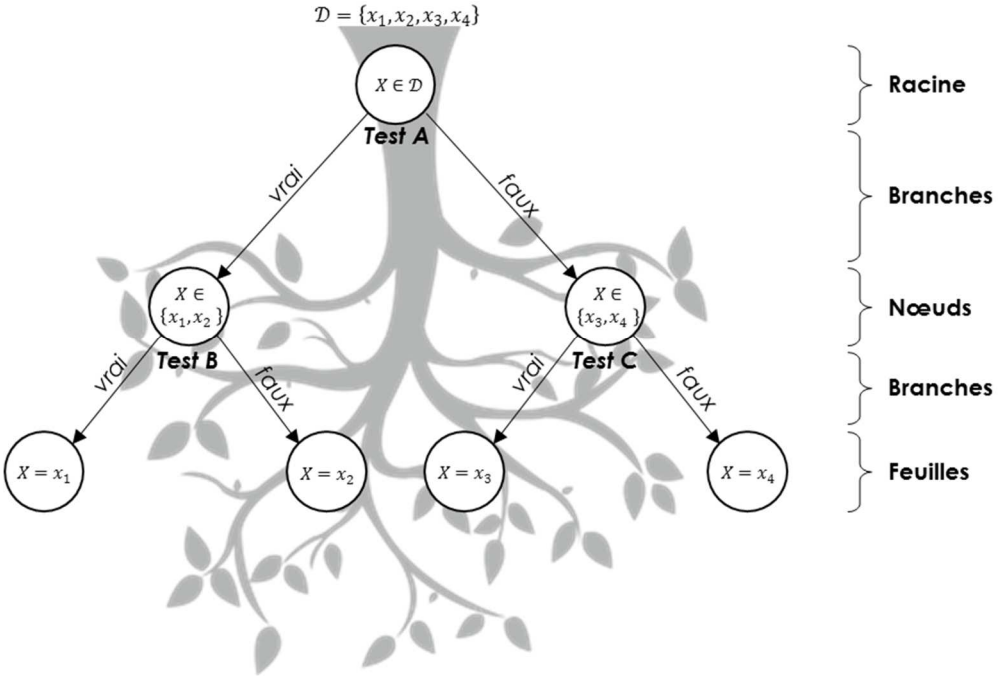


Figure 56 – Schématisation d'un arbre de décision de type CART

Formellement, nous considérons une base de données $(\boldsymbol{X}, \boldsymbol{Y})$ contenant $p+1$ colonnes et n lignes avec une variable à expliquer $\boldsymbol{Y}$ et p variables explicatives $\boldsymbol{X_k}$ ($k \in [\![1, p]\!]$) observées sur un échantillon de n observations. La construction d'un CART consiste à déterminer la séquence des nœuds qui vont chacun partitionner en deux l'espace des valeurs possibles prises par $\boldsymbol{Y}$. Pour cela, il est nécessaire de définir trois critères :

- <u>Critère de division :</u> Il s'agit de définir comment sélectionner la meilleure division[122] à appliquer au nœud. L'idée est rendre les deux sous-arbres obtenus le plus homogène possible au sens de $\boldsymbol{Y}$. Nous allons donc définir une mesure de l'hétérogénéité[123] $\mathfrak{D}_\kappa$ d'un nœud κ puis nous allons choisir la division qui minimise $\mathfrak{D}_{\kappa_G} + \mathfrak{D}_{\kappa_D}$ au niveau des deux nœuds κ_G et κ_D nés de la division (autrement dit, le gain d'homogénéité apporté par la division sera maximisé).

[120] Il existe d'autres algorithmes basés sur le concept d'arbre de décision (avec par exemple davantage de branches) mais nous nous concentrerons dans le cadre de ce mémoire uniquement sur l'algorithme CART qui est le plus simple et le plus utilisé dans la pratique.

[121] Un nœud ne donnant pas naissance à des branches et donc à des sous-arbres s'appelle un nœud terminal.

[122] Une division correspond au choix d'une variable explicative $\boldsymbol{X_i}$ parmi $\boldsymbol{X}$ et d'un test par rapport à la valeur prise par $\boldsymbol{X_i}$. La base de donnée $(\boldsymbol{X}, \boldsymbol{Y})$ est ainsi divisée selon que le test ainsi défini est vrai ou faux. Pour qu'une division soit admissible, il est nécessaire que les deux résultats possibles du test contiennent au moins une observation.

[123] Une mesure $\mathfrak{D}_\kappa$ de l'hétérogénéité du nœud κ est une fonction non négative qui est nulle si et seulement si toutes les valeurs de $\boldsymbol{Y}$ sont identiques et maximale lorsqu'elles sont dispersées de manière équiprobable. Lorsque la variable $\boldsymbol{Y}$ est quantitative, la mesure d'hétérogénéité du nœud κ pourra être la variance empirique $\mathfrak{D}_\kappa = \frac{1}{n_\kappa} \times \sum_{i=1}^{n_\kappa} (y_i - \overline{\boldsymbol{y_\kappa}})^2$ en notant n_κ le nombre de lignes au niveau du nœud κ. Lorsque que la variable $\boldsymbol{Y}$ est qualitative selon M modalités $\forall\, i \in [\![1, n_\kappa]\!], y_i \in \{y^{(1)}, y^{(2)}, \dots, y^{(M)}\}$, la mesure d'hétérogénéité du nœud κ pourra être la concentration de Gini $\mathfrak{D}_\kappa = \sum_{m=1}^{M} p_\kappa^m (1 - p_\kappa^m)$ où p_κ^m est la fréquence empirique d'avoir $\boldsymbol{Y} = y^{(m)}$ (ie. $p_\kappa^m = \frac{1}{n_\kappa} \times \sum_{i=1}^{n_\kappa} \mathbf{1}_{\{y_i = y^{(m)}\}}$).

- Critère d'arrêt : Il s'agit de déterminer quand un nœud sera considéré comme terminal (qu'il deviendra donc une feuille et ne donnera plus naissance à de nouvelles branches). La règle d'arrêt peut être lorsque toutes les valeurs de $\boldsymbol{Y}$ au sein du nœud κ sont identiques, lorsqu'il n'existe plus de division admissible ou bien lorsque le nombre d'observations n_κ devient inférieur à un certain seuil.

- Critère d'affectation : Il s'agit de décider quelle valeur sera affectée à la variable $\boldsymbol{Y}$ pour chaque feuille de l'arbre. De manière instinctive, il est possible par exemple de considérer la moyenne empirique lorsque $\boldsymbol{Y}$ est quantitative et la valeur la plus représentée lorsque $\boldsymbol{Y}$ est qualitative.

Grâce à ces trois critères et à un jeu d'entraînement (pour lequel les valeurs de $\boldsymbol{Y}$ sont connues), il est ainsi possible de construire un arbre maximal $\mathcal{A}_{max}$ qui servira à prédire des valeurs de $\boldsymbol{Y}$ inconnues à partir des valeurs de $\boldsymbol{X}$ observées. Cependant, ce modèle de prévision sera très instable car fortement dépendant du jeu d'entraînement utilisé. Il va par conséquent être nécessaire d'ajouter une étape d'élagage (ou *pruning*) afin de réduire la complexité de l'arbre et ainsi éviter le surapprentissage.

Cette étape d'élagage consiste à choisir l'arbre optimal qui minimise l'erreur de généralisation[124] entre l'arbre maximal $\mathcal{A}_{max}$ et l'arbre trivial $\mathcal{A}_1$ ne comportant qu'un unique nœud. Pour des raisons d'efficacité [BREIMAN, L. et al. (1984)], la recherche de cet arbre optimal sera restreinte à la suite emboîtée des sous-arbres d'hétérogénéité croissante[125] ayant permis de passer de $\mathcal{A}_1$ à $\mathcal{A}_{max}$. Par construction, le gain d'homogénéité entre deux sous-arbres $\mathcal{A}_\kappa$ et $\mathcal{A}_{\kappa+1}$ est une fonction décroissante de κ. Nous allons donc partir de l'arbre maximal $\mathcal{A}_{max}$ déterminé à partir d'un jeu d'entraînement et remonter la suite de sous-arbres afin de sélectionner celui qui possède l'erreur de généralisation minimale[126], calculée à partir d'un jeu de validation.

Une fois l'arbre optimal construit et sélectionné, il peut être utilisé pour imputer des données manquantes. Il suffit en effet de le parcourir en regardant les variables explicatives afin de trouver la feuille correspondante pour y trouver la valeur à imputer. Le même vecteur de variables explicatives donnera ainsi nécessairement la même valeur à imputer.

L'immense avantage d'un modèle CART est qu'il ne nécessite aucune hypothèse sur la distribution des variables et fonctionne d'autant mieux que les variables explicatives sont

---

[124] L'erreur de généralisation $e_{\mathcal{A}}$ d'un arbre $\mathcal{A}$ possédant $\mathrm{K}_{\mathcal{A}}$ nœuds correspond à la mesure de la distance entre la prédiction $\widehat{\boldsymbol{y}}_{\boldsymbol{\kappa}}$ de $\boldsymbol{Y}$ donnée par le modèle au nœud κ et sa vraie valeur $\boldsymbol{y}$ rapporté au nombre d'observations n_κ. Comme vu au § II.1.3.2., il existe différentes mesures de distance selon que $\boldsymbol{Y}$ est une variable quantitative (par exemple $e_{\mathcal{A}} = \frac{1}{\sum_{\kappa=1}^{\mathrm{K}_{\mathcal{A}}} n_\kappa} \times \sqrt{\sum_{\kappa=1}^{\mathrm{K}_{\mathcal{A}}} (\widehat{\boldsymbol{y}}_{\boldsymbol{\kappa}} - y)^2}$) ou une variable qualitative (par exemple $e_{\mathcal{A}} = \sum_{\kappa=1}^{\mathrm{K}_{\mathcal{A}}} \frac{1}{n_\kappa} \times \mathbf{1}_{\{\widehat{\boldsymbol{y}}_{\boldsymbol{\kappa}} \neq y\}}$). Il y a un lien très fort entre l'erreur de généralisation et la mesure d'hétérogénéité puisque que ces deux mesures sont proportionnelles et servent à quantifier le risque d'erreur du modèle.

[125] L'hétérogénéité $\mathfrak{D}_{\mathcal{A}}$d'un arbre $\mathcal{A}$ à $\mathrm{K}_{\mathcal{A}}$ feuilles correspond à la somme de l'hétérogénéité de toutes ses feuilles $\mathfrak{D}_{\mathcal{A}} = \sum_{\kappa=1}^{\mathrm{K}_{\mathcal{A}}} \mathfrak{D}_\kappa$.

[126] Plus précisément, nous allons calculer l'erreur de généralisation $e_{\mathcal{A}_\kappa}$ sur l'ensemble des sous-arbres de la suite des $\mathcal{A}_\kappa$ puis nous choisirons comme arbre optimal celui avec l'erreur de généralisation minimale e_{min} (il ne s'agit donc que d'un minimum local puisque tous les arbres possibles n'ont pas été parcourus). Il est également possible de sélectionner le plus petit arbre dont l'erreur de généralisation est comprise dans l'intervalle $[e_{min}\ ;\ e_{min} + \varepsilon]$ (avec $\varepsilon > 0$ suffisamment petit) afin de gagner sur la complexité en sacrifiant un petit peu sur la précision du modèle.

nombreuses. Il permet en outre de traiter facilement des bases mixtes qui possèdent à la fois des données qualitatives et quantitatives. Cependant, l'algorithme est par construction très instable (c'est-à-dire qu'il peut donner des résultats très différents selon l'échantillon qui lui servira de jeu d'entraînement) puisqu'une erreur de division à la racine de l'arbre va se propager sur l'ensemble de ses branches. C'est pour pallier ce problème qu'ont émergé les techniques d'agrégation de modèle, parmi lesquelles se trouve la méthode des forêts aléatoires.

Random forest

Le concept d'agrégation de modèles consiste à améliorer un modèle particulièrement instable via une combinaison d'un grand nombre de modèles. Parmi les différentes techniques d'agrégation, nous nous intéressons plus particulièrement ici aux forêts aléatoires introduit par Breiman [BREIMAN, L. (2001)] afin d'agréger spécifiquement des modèles de CART.

La technique des forêts aléatoires repose sur le principe d'agrégation du *bagging*. Il s'agit de décomposer les n observations de la base de données $(\boldsymbol{X}, \boldsymbol{Y})$ en B échantillons indépendants $\{(\boldsymbol{X_i}, \boldsymbol{y_i})\}_{i \in [\![1,B]\!]}$ obtenus par une méthode de *bootstrap*[127]. Chaque échantillon $(\boldsymbol{X_i}, \boldsymbol{y_i})$ va alors donner lieu à une prédiction $\widehat{\boldsymbol{y}}_{\boldsymbol{i}} = G_i(\boldsymbol{X})$ par un modèle CART. L'ensemble sera ensuite agrégé pour constituer le résultat final, soit pour les variables quantitatives par une moyenne $\widehat{\boldsymbol{Y}} = G(\boldsymbol{X}) = \frac{1}{B}\sum_{i=1}^{B} G_i(\boldsymbol{X})$; soit pour les variables qualitatives par un vote majoritaire $\widehat{\boldsymbol{Y}} = G(\boldsymbol{X}) = \arg\max_{\boldsymbol{Y}} |i \mid G_i(\boldsymbol{X}) = \boldsymbol{Y}|$. Cette agrégation a ainsi pour objectif de réduire la variance des estimateurs en les moyennant.

Néanmoins, il n'est pas évident que les B échantillons *bootstrap* soient effectivement indépendants. Bien au contraire, le fait que les $\boldsymbol{X_i}$ soient obtenus par un tirage avec remise de la même base de données montre que ces variables sont certes aléatoires et identiquement distribuées mais corrélées. Plus précisément, Breiman [BREIMAN, L. (2001)] a montré qu'en notant μ la moyenne des $\boldsymbol{X_i}$, σ^2 leur variance et ρ leur coefficient de corrélation deux à deux, alors la variance de la moyenne de ces B variables aléatoires est :

$$Var\left[\frac{1}{B} \times \sum_{i=1}^{B} \boldsymbol{X_i}\right] = \rho\sigma^2 + \frac{1-\rho}{B\sigma^2} \xrightarrow[B \to +\infty]{} \rho\sigma^2$$

Par conséquent, si les B échantillons ne sont pas indépendants, alors les $G_i(\boldsymbol{X})$ ne le seront pas non plus et même en faisant grandir la valeur de B, nous serons bloqués par le premier terme de la formule précédente pour véritablement réduire la variance.

[127] Le *bootstrap* est une technique de simulation d'un échantillon particulièrement utile lorsque la loi de la variable est inconnue et qu'il n'est pas envisageable de prendre l'hypothèse qu'elle soit normale. Le principe du *bootstrap* est de ré-échantillonner une base de données constituées de n observations par un tirage aléatoire de n_B lignes de la base avec remise afin d'obtenir un nouvel échantillon de *bootstrap* de n_B lignes. Usuellement, l'ordre de grandeur est $n_B = \frac{2}{3} \times n$.

Pour contourner ce problème, la méthode des forêts aléatoires va chercher à faire baisser le degré de corrélation entre les $G_i(\boldsymbol{X})$ en ajoutant une étape de *randomisation* consistant pour chaque arbre $\mathcal{A}_i$ à choisir aléatoirement q variables explicatives parmi les p variables disponibles[128]. Cette sélection aléatoire à chaque étape de construction d'un CART va ainsi accroître la variabilité des arbres et ainsi assurer l'indépendance entre les $G_i(\boldsymbol{X})$. Dans ces conditions et grâce au grand nombre d'arbres considérés, l'étape d'élagage présenté dans la partie précédente n'est plus nécessaire[129].

Le choix du nombre d'arbre B constituant la forêt aléatoire s'effectue par optimisation de l'erreur OOB (*Out Of Bag*)[130] : l'algorithme continue à générer des échantillons bootstrap (donc à augmenter B) et à générer des CART tant que l'erreur OOB du modèle diminue.

L'algorithme des forêts aléatoires conserve tous les avantages de l'algorithme CART mais en étant beaucoup plus robuste tout en restant efficace pour traiter les très grandes bases de données ainsi que les problèmes non linéaires, si bien qu'il est considéré comme la méthode de référence pour résoudre un grand nombre de problèmes concernant les bases de données, en particulier l'imputation des valeurs manquantes.

Réseau de neurones

Afin de bien comprendre ce qu'est un algorithme de réseau de neurones, il est nécessaire de bien définir ce qu'est un neurone informatique. Un neurone (ou perceptron) correspond tout simplement à une fonction avec un certain nombre n de variables en entrée $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ et une variable en sortie y. Très basiquement, la sortie sera une combinaison linéaire des entrées selon différents poids $\boldsymbol{w} = (w_1, w_2, \dots, w_n)$.

Cependant, si le neurone se cantonnait à ce type d'agrégation, l'algorithme résultant, quel que soit la complexité du réseau élaboré, ne pourrait traiter que des problèmes linéaires sans aucun avantage par rapport à un simple modèle de régression linéaire. Il est donc indispensable d'introduire un aspect non linéaire au sein du neurone en appliquant à la somme pondérée des entrées une fonction de seuil f_s. Nous parlons d'effet de seuil car la fonction f_s est construite afin d'introduire trois intervalles :

- en-dessous du seuil, le neurone n'est pas activé (sa sortie vaut 0 ou 1) ;
- au-dessus du seuil, le neurone est actif (sa sortie vaut 1) ;
- au voisinage du seuil, le neurone est dans une phase de transition.

[128] Le choix du nombre de variables explicatives pour la construction des différents CART est généralement optimisée par une procédure de validation croisée (cf. § annexe 5). La valeur par défaut utilisée est $q = \sqrt{p}$ si $\boldsymbol{Y}$ est une variable qualitative et $q = \frac{p}{3}$ si $\boldsymbol{Y}$ est une variable quantitative.

[129] Usuellement, la taille de l'arbre est simplement restreinte en indiquant un nombre minimum d'observations par nœud.

[130] L'erreur *Out Of Bag* correspond à la moyenne des erreurs de généralisation des G_i pour lesquelles le jeu de validation correspond à la partie de la base de données ne faisant pas partie de l'échantillon *bootstrap* $(\boldsymbol{X_i}, \boldsymbol{y_i})$.

Nous obtenons donc *in fine* :

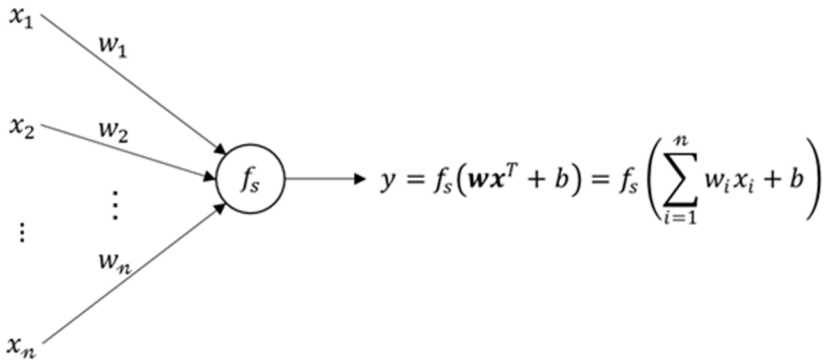


Figure 57 – Schématisation d'un neurone artificiel

Il existe plusieurs fonctions[131] permettant d'obtenir cet effet de seuil avec une phase de transition plus ou moins abrupte. Selon la nature de cette fonction, une très légère variation au sein des entrées pourra ainsi entraîner une forte variation sur la sortie.

Le principe d'un réseau de neurones[132] est d'emboîter différentes couches successives de perceptrons de sorte que les sorties de la première couche (la couche d'entrées) viennent alimenter les entrées des suivantes (les couches cachées) jusqu'à aboutir à la dernière couche (la couche de sortie).

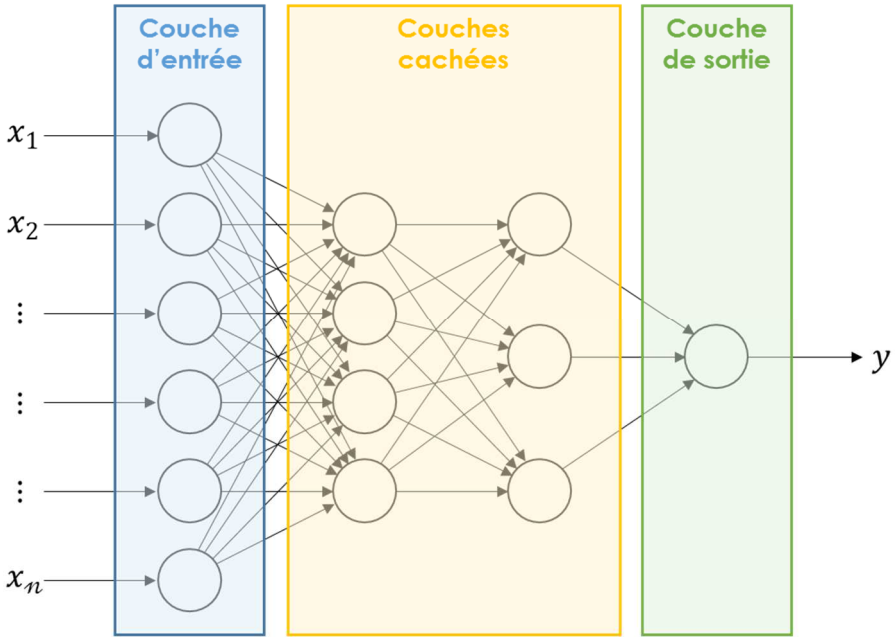


Figure 58 – Schématisation d'un réseau de neurones

Le nombre de neurones au sein d'une couche n'est imposé que pour la couche d'entrée (en fonction du nombre de données en entrée de l'algorithme) et pour la couche de sortie (en fonction du nombre de données en sortie de l'algorithme, généralement une seule). Le

[131] Par exemple, la fonction sigmoïde (aussi appelée fonction logistique) $z \longmapsto \phi(z) = \frac{1}{1+e^{-z}}$ présente un aspect relativement lisse tandis que la fonction *Rectified Linear Unit* $z \longmapsto ReLU(z) = \max(0, z)$ présente un effet de seuil plus violent (mais qui permet une descente du gradient plus efficace).

[132] Dans le cadre de ce mémoire, nous ne nous intéresserons qu'au type de réseaux de neurones le plus simple, à savoir les réseaux de neurones à propagation avant. Cependant il existe un grand nombre de structures diverses et variées de réseaux de neurones.

nombre de couches cachées et le nombre de neurones au sein de chaque couche cachée sont quant à eux des paramètres à optimiser en fonction de l'efficacité et de la performance attendues pour l'algorithme.

Une fois ces hyperparamètres fixés, le schéma du réseau de neurones est arrêté et les paramètres restant à optimiser grâce à une base d'entraînement (qui représentent donc l'apprentissage de l'algorithme) sont les différents poids en entrée de chaque neurone. Afin de mettre à jour ces poids, il est essentiel de mesurer la performance du réseau de neurones en comparant la valeur prédite[133] en sortie de modèle à la valeur réelle au moyen d'une fonction de perte[134]. A l'aide de la perte mesurée, nous allons ainsi remonter les différentes couches afin de modifier la valeur des poids selon leur contribution relative[135]. Les changements à appliquer aux différents poids sont déterminés avec la méthode de descente du gradient[136].

Comme pour tous les algorithmes de *machine learning*, le réseau de neurones va ainsi être optimisé à l'aide d'une base d'entraînement et pourra ainsi être utilisé pour imputer les valeurs manquantes. Cependant, contrairement aux forêts aléatoires, il existe encore très peu de littérature sur cette utilisation spécifique des réseaux de neurones même s'il commence à apparaître depuis deux ans des modèles très complexes d'imputation de données manquantes via des algorithmes de *deep learning* [CHE, Z. et al. (2018)]. Dans le cadre de ce mémoire, nous nous contenterons de tester les performances d'imputation d'un réseau de neurones basique.

Régression logistique

La régression logistique est un modèle de classification binaire qui permet de restituer une prédiction sous la forme d'une probabilité (la classification est en réalité déduite de cela en appliquant un seuil, ou « *cutoff* », permettant de se ramener à une prédiction d'appartenance à la classe 0 ou 1). Elle cherche une relation linéaire entre les variables explicatives et le *log-odds*[137] de la cible binaire.

---

[133] Il s'agit de propagation *forward* car le réseau de neurones est parcouru dans le sens des entrées vers les sorties.

[134] Il existe de nombreuses fonctions de perte selon la tâche que l'algorithme doit effectuer, en particulier s'il s'agit d'un problème de classification (données qualitatives) ou de régression (données quantitatives). Comme pour l'erreur de généralisation évoquée précédemment, la fonction de perte consiste à mesurer la distance entre la prédiction du modèle et la valeur réelle. Ce point sera abordé en détail plus loin.

[135] Il s'agit de propagation *backward* car le réseau de neurones est parcouru dans le sens des sorties vers les entrées.

[136] La méthode de descente du gradient est une technique itérative classique d'optimisation, typiquement afin de trouver les paramètres $\boldsymbol{w}$ minimisant une fonction de perte $\boldsymbol{w} \longmapsto f_p(\boldsymbol{w})$ supposée dérivable. Elle consiste à adapter le pas de changement de valeur des paramètres en fonction de leur variabilité au voisinage du point courant (le gradient peut être vue comme une généralisation multidimensionnelle de la dérivée) : plus cette variabilité est faible et plus le pas sera petit. Concrètement, à chaque étape k, tant que la valeur absolue du gradient $\|\nabla f_p(\boldsymbol{w_k})\|$ est supérieure à un certain seuil de tolérance $\varepsilon > 0$, la nouvelle valeur des paramètres $\boldsymbol{w_{k+1}}$ est recherchée sur la demie-droite définie par l'opposée du gradient $-\nabla f_p(\boldsymbol{w_k})$, à une distance proportionnelle à celui-ci : $\boldsymbol{w_{k+1}} = \boldsymbol{w_k} - \gamma_k \nabla f_p(\boldsymbol{w_k})$. Cet algorithme fait ainsi décroître à chaque étape la fonction de perte : $f_p(\boldsymbol{w_{k+1}}) < f_p(\boldsymbol{w_k})$.

[137] En considérant un événement A ayant une probabilité p, le rapport de chances ou « *odds* » correspondent à la probabilité de réalisation de A par rapport à sa probabilité de non-réalisation, soit $odds = \frac{p}{1-p}$. Le *log-odds* correspond au logarithme de l'*odds*, soit $log\text{-}odds = \ln\left(\frac{p}{1-p}\right)$.

En notant p la probabilité *a posteriori* d'obtenir la modalité 1 de la cible Y sachant la valeur prise par les variables explicatives X (c'est-à-dire $p = \mathbb{P}[Y = 1|X]$), cela signifie la recherche des coefficients linéaires a_i pour obtenir la relation suivante :

$$log\text{-}odds = \ln\left(\frac{p}{1-p}\right) = a_0 + a_1x_1 + a_2x_2 + \cdots$$ [138]

Elle est ensuite optimisée par la maximisation de la vraisemblance via la méthode itérative du gradient. Elle a l'avantage d'avoir une solution interprétable puisque les cotes (ou « *odds* ») sont obtenues en passant à l'exponentielle les coefficients. Ainsi :

- si $e^{a_1} > 1$: quand x_1 augmente, la probabilité d'être dans la classe 1 augmente
- si $a_1 > a_2 > 0$: x_1 joue plus sur l'évolution de la probabilité que x_2.

Machines à vecteurs de supports

Les *Support Vector Machines,* souvent traduit par l'appellation de Séparateur à Vaste Marge (SVM), sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination c'est-à-dire la prévision d'une variable qualitative binaire via la recherche d'un hyperplan optimal qui sépare les observations afin de les classer.

L'algorithme consiste à l'identification de vecteurs supports, qui sont les observations permettant de définir la frontière de décision pour la classification.

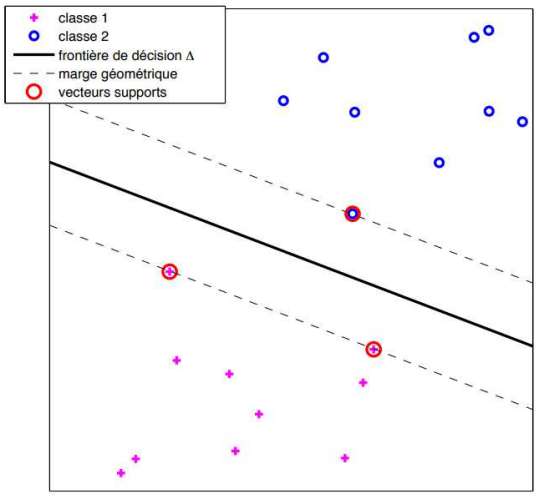


Figure 59 – Illustration de la notion de vecteur support dans le cas d'un problème linéairement séparable [FRANCOEUR, D. (2010)]

C'est une méthode robuste, qui s'avère efficace mais qui est très gourmande en capacité de calcul et qui fournit des résultats difficiles à interpréter[139].

[138] La régression logistique se nomme ainsi car la solution de cette équation donne $p = \frac{e^{a_0+a_1x_1+a_2x_2+\cdots}}{1+e^{a_0+a_1x_1+a_2x_2+\cdots}}$, ce qui correspond à la densité d'une loi logistique.

[139] Pour ces raisons, les machines à vecteurs de supports sont de moins en moins utilisées depuis l'avènement des réseaux de neurones.

Analyse discriminante

L'analyse discriminante tente de répondre à la question suivante : quelles sont les combinaisons linéaires de variables qui permettent de séparer le mieux possible les k catégories ?

La méthode consiste à estimer la probabilité conditionnelle pour chaque classe « k » $\mathbb{P}[Y = y_k|\boldsymbol{X}]$. Pour un individu ω à classer dans une des classes k, cela revient donc à chercher à minimiser l'erreur d'affectation :

$$y_k^* = \arg\max_k \mathbb{P}[Y(\omega) = y_k|\boldsymbol{X}(\omega)]$$

Nous affectons ainsi l'individu à la classe qui maximise sa probabilité d'appartenance.

Le théorème de Bayes permet de reformuler la problématique de la façon suivante :

$$\mathbb{P}[Y = y_k|\boldsymbol{X}] = \frac{\mathbb{P}[\boldsymbol{X}|Y = y_k] \times \mathbb{P}[Y = y_k]}{\mathbb{P}[\boldsymbol{X}]} = \frac{\mathbb{P}[\boldsymbol{X}|Y = y_k] \times \mathbb{P}[Y = y_k]}{\sum_l \mathbb{P}[\boldsymbol{X}|Y = y_l] \times \mathbb{P}[Y = y_l]}$$

La règle d'affectation aux classes peut donc s'écrire ainsi :

$$y_k^* = \arg\max_k \mathbb{P}[\boldsymbol{X}|Y = y_k] \times \mathbb{P}[Y = y_k]$$

Par ailleurs, $\mathbb{P}[Y = y_k]$ peut être estimé facilement par la fréquence empirique $\frac{n_k}{n}$. L'estimation du terme $\mathbb{P}[\boldsymbol{X}|Y = y_k]$ est plus compliquée et nécessite de supposer les hypothèses suivantes :

- normalité : la probabilité conditionnelle suit une loi normale multivariée, soit $\mathbb{P}[\boldsymbol{X}|Y = y_k] \rightsquigarrow \mathcal{N}(\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$;
- homoscédasticité : les matrices de covariances conditionnelles sont identiques, soit $\boldsymbol{\Sigma_k} = \boldsymbol{\Sigma}$.

Ces deux hypothèses, permettent, en passant au logarithme et en retirant les termes qui ne dépendent pas de k, de réécrire la fonction de classement sous la forme suivante :

$$\ln(\mathbb{P}[\boldsymbol{X}|Y = y_k]) = -\frac{1}{2}(x - \boldsymbol{\mu_k})\boldsymbol{\Sigma}^{-1}(x - \boldsymbol{\mu_k})^t + C_k$$

L'espérance $\boldsymbol{\mu_k}$ et la matrice de variance-covariance $\boldsymbol{\Sigma}$ peuvent être estimées empiriquement sur l'échantillon :

- $\boldsymbol{\mu_k}$ par $\overline{\boldsymbol{x_k}}$, ce dernier étant le vecteur moyenne pour les individus correspondant à la classe y_k
- $\boldsymbol{\Sigma}$ par $\frac{1}{n-K}\sum_{k=1}^{K}(n_k - 1) \times \boldsymbol{V_k}$, K étant le nombre de classes k et V_k la matrice de variance-covariance empirique pour y_k

Le terme principal correspond à une distance aux moyennes conditionnelles pondérée par l'inverse de la matrice de variance-covariance. Ce terme est positif mais précédé d'un signe négatif dans la fonction d'affectation aux classes, ce qui correspond au fait que l'individu a tendance à être associé à la classe dont le barycentre est le plus proche.

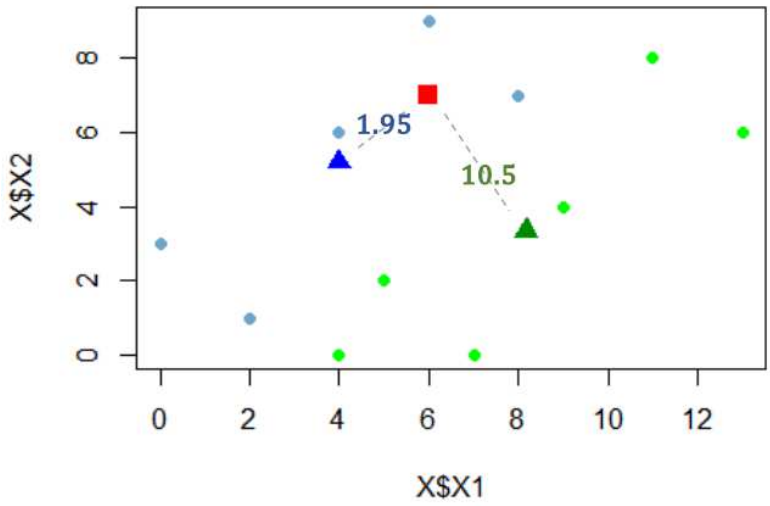


*Figure 60 – Exemple dans un cas avec **X** à deux dimensions*

Dans l'exemple ci-dessus, le point à classer (en rouge) est plus proche du barycentre (les triangles) de la classe bleue et y sera donc affecté [RAKOTOMALALA, R. (2020a)].

Enfin, en développant et en retirant à nouveau les termes qui ne dépendent pas de k, nous obtenons finalement une fonction de classement qui est linéaire et la règle d'affectation aux classes qui peut s'écrire :

$$y_k^* = \arg\max_k \left(a_{k0} + a_{k1}x_1 + a_{k2}x_2 + \cdots \right)$$

Via la fonction de classement, l'analyse discriminante a l'avantage de fournir un modèle lisible et facile à interpréter. Comme la régression logistique, elle permet d'estimer une probabilité d'appartenance aux classes, ce qui nous intéresse en particulier.

Boosting

Le *boosting* est une méthode qui consiste à agréger des classifieurs (en général simples, comme les arbres de décisions ou les régressions logistiques) élaborés sur des échantillons d'apprentissage en ajustant au fil des agrégations la pondération des observations[140].

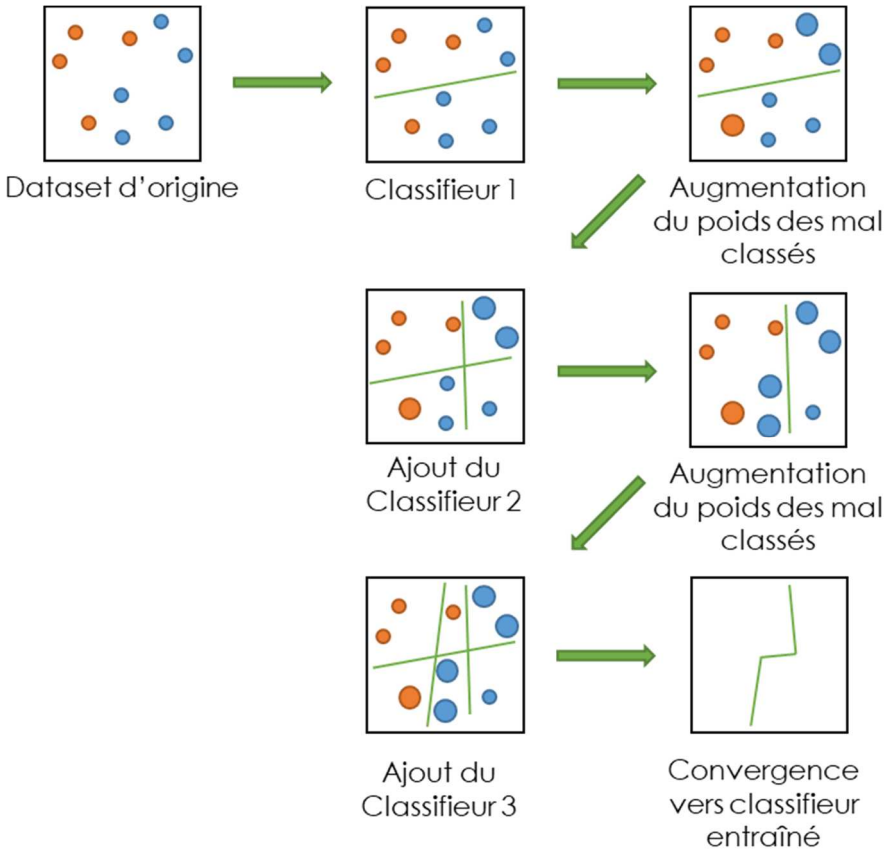


Figure 61 – Description du fonctionnement du boosting

Une des méthodes de *boosting* les plus connues est celle du *gradient boosting*. Son objectif est de minimiser une fonction de perte globale J, somme d'une fonction de coût j calculée sur l'ensemble des observations x_i, confrontant la valeur observée de la cible y_i et la prédiction d'un modèle f pour l'observation donnée :

$$J(f) = \sum_{i=1}^{n} j\left(y_i, f(x_i)\right)$$

[140] A noter que cette pondération est l'essentielle différence avec le *bagging* qui consiste plutôt à moyenner un ensemble de modèles.

C'est sur le modèle f que nous allons itérer, en partant d'un modèle f_0 et en appliquant la méthode de descente de gradient avec la relation suivante :

$$f_b(x_i) = f_{b-1}(x_i) - \eta \times \nabla j\big(y_i, f(x_i)\big)$$

- f_b étant le classifieur de l'étape b,
- η est la constante d'apprentissage, qui définit l'information conservée aux étapes précédentes et joue sur la vitesse de convergence de l'algorithme,
- ∇ est le gradient, c'est-à-dire la dérivée partielle de la fonction de coût j par rapport au modèle :

$$\nabla j\big(y_i, f(x_i)\big) = \frac{\partial j\big(y_i, f(x_i)\big)}{\partial f(x_i)}$$

Si le classifieur f est classiquement un arbre de régression, la fonction de coût est, en revanche, adaptable à la problématique étudiée.

La méthode du *gradient boosting* est reconnue empiriquement comme très efficace [RAKOTOMALALA, R. (2020b)]. Un des inconvénients est qu'elle ne fournit pas de résultats interprétables. Par ailleurs, son bon paramétrage peut s'avérer très complexe (nombreux paramètres à ajuster).

ANNEXE 8 : DESCRIPTION DE LA BASE DE DONNEES D'ETUDE

L'ensemble des données récupérées a été consolidé dans une seule table de manière à la préparer pour appliquer les modèles de *machine learning*. La table obtenue contient les données suivantes, classées par provenance.

Données issues de l'assureur

Le traitement de récupération des données, décrit en II.2.1.2., a permis de construire la table suivante :

Maille	Champ	Temporalité
Par assuré / ouvrant droit	Code sexe	
	Age de l'assuré	au 01/01/N
	Situation de famille de l'assuré	au 01/01/N
	Code grand régime	au 01/01/N
	Code régime local	au 01/01/N
	Département de résidence	au 01/01/N
	Région de résidence	au 01/01/N
	Pays de résidence	au 01/01/N
	Taille de la famille[141] de l'assuré	au 01/01/N
	Nombre d'enfants	au 01/01/N
	Nombre de femmes dans la famille	au 01/01/N
	Nombre d'hommes dans la famille	au 01/01/N
	Nombre de jours d'affiliation Santé	sur exercice N-1
	Nombre de jours d'affiliation Prev	sur exercice N-1
	Collège du contrat en Santé	au 01/01/N
	Collège du contrat en Prev	au 01/01/N
	Secteur d'activité de l'entreprise rattaché	au 01/01/N
	Convention collective de l'entreprise	au 01/01/N

[141] En fait, nombre de bénéficiaire (y compris l'assuré lui-même)

Maille	Champ	Temporalité
	Montant Dépense réelle totale	sur exercice N-1 + historique N-2 et N-3, vu au 01/01/N
	Montant Reste à charge total	sur exercice N-1 + historique N-2 et N-3, vu au 01/01/N
	Montant Dépassement honoraire total	sur exercice N-1 + historique N-2 et N-3, vu au 01/01/N
	Nombre d'actes pondérés	sur exercice N-1 + historique N-2 et N-3, vu au 01/01/N
	Nombre d'actes en dépassement honoraire	sur exercice N-1 + historique N-2 et N-3, vu au 01/01/N
Par assuré et par type d'acte	Montant Dépense réelle	sur exercice N-1 + historique N-2 et N-3, vu au 01/01/N
	Montant Reste à charge	sur exercice N-1 + historique N-2 et N-3, vu au 01/01/N
	Montant Dépassement honoraire	sur exercice N-1 + historique N-2 et N-3, vu au 01/01/N
	Nombre d'actes pondérés	sur exercice N-1 + historique N-2 et N-3, vu au 01/01/N
	Nombre d'actes en dépassement honoraire	sur exercice N-1 + historique N-2 et N-3, vu au 01/01/N
Par assuré et par risque prévoyance[142]	Nb survenance de sinistre	sur exercice N-1, vu au 01/01/N
	Dossier en cours	sur exercice N-1, vu au 01/01/N
	Montant versé en N	sur exercice N-1, vu au 01/01/N

Les types d'actes sont les suivants :

- Actes de spécialité
- Frais médicaux
- Auxiliaires médicaux
- Pharmacie
- Optique
- Prothèse Orthopédie
- Prothèse Dentaire
- Analyses Radiographie
- Traitement orthodontique
- Hospitalisation
- Soins dentaires
- Cures

Comme vu en II.2.1.2.3., les données historiques ont été agrégées puis appariées aux assurés.

[142] Incapacité, invalidité, rente éducation et rente de conjoint

Données issues d'Open DAMIR

Les données issues d'Open DAMIR ont été rattachées à l'assuré grâce à la méthode décrite en II.2.2.2. .

Maille	Champ	Temporalité
Par assuré et par type d'acte	Nombre actes	Données appariées N-1 et N-2
	Paiement total	Données appariées N-1 et N-2
	Remboursement total	Données appariées N-1 et N-2
	Dépassement total	Données appariées N-1 et N-2
	Nombre dépassement	Données appariées N-1 et N-2
	Fréquence des actes par rapport à la population INSEE concernée	Données appariées N-1 et N-2
	Cout moyen	Données appariées N-1 et N-2
	Fréquence des dépassements parmi les actes de même type	Données appariées N-1 et N-2
	Cout moyen dépassement	Données appariées N-1 et N-2

Nous nous sommes ramenés aux mêmes types d'actes que pour les données de l'assureur grâce à une table de correspondance, construite par nos soins.

Données calculées

Certaines données ont été calculées à partir des informations déjà à disposition afin de mettre en évidence certains aspects susceptibles d'améliorer nos résultats (évoqué en II.2.2.3.)

Maille	Champ	Temporalité
Par assuré et par type d'acte	Ecart cout moyen avec données DAMIR sur l'individu	
	Ecart cout moyen des dépassements avec données DAMIR sur l'individu	
	Ecart à la fréquence des dépassements avec les données DAMIR sur l'individu	

ANNEXE 9 : STATISTIQUES DESCRIPTIVES DE LA BASE D'ETUDE

Pour vérifier la nature et la qualité des données sur lesquelles nous allons appliquer nos modèles, nous observons les éléments suivants.

Pyramide des âges

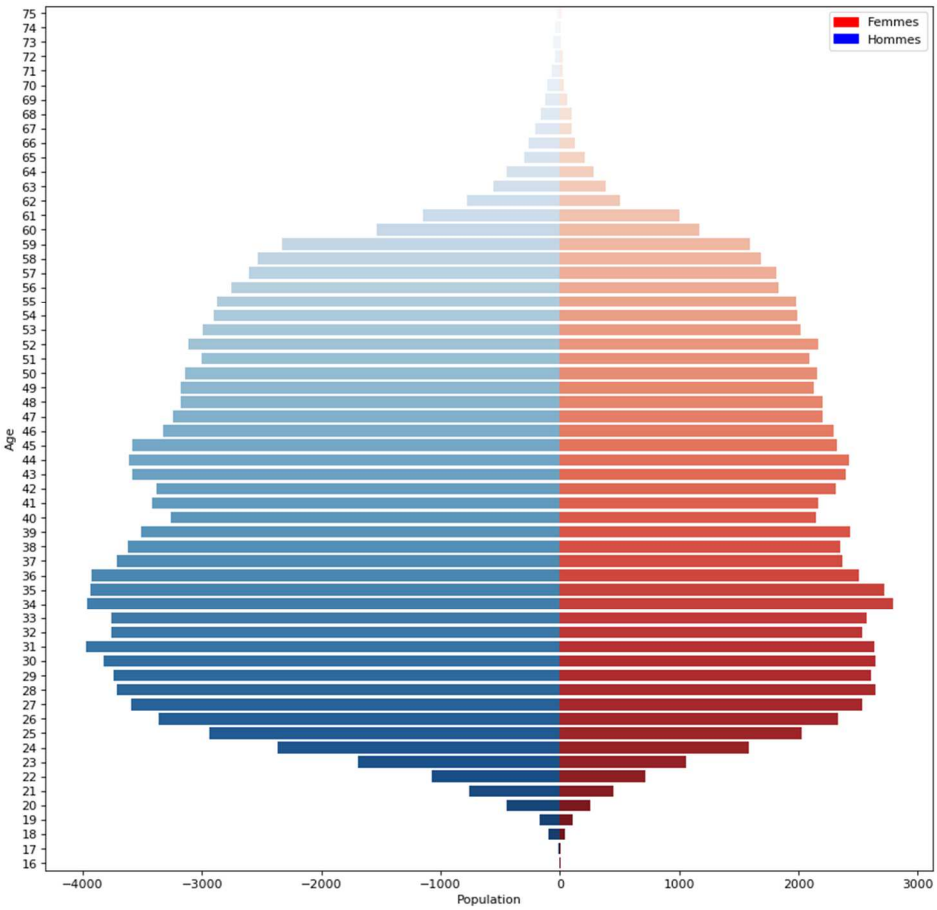


Figure 62 – Pyramide des âges de notre base d'étude

Nous remarquons que les hommes sont sur-représentés. Mis à part ce point, la pyramide ne semble pas déformée outre mesure.

Répartition géographique des individus

Région de Résidence	Individus dans le Dataset	Population totale région	Pourcentage de représentativité du dataset
Régions et Départements d'outre-mer	629	2 136 391	0,03%
Ile-de-France	55 308	12 138 930	0,46%
Centre-Val de Loire	7 666	2 580 581	0,30%
Bourgogne-Franche-Comté	8 315	2 819 635	0,29%
Normandie	8 997	3 341 440	0,27%
Hauts-de-France - Nord-Pas-de-Calais-Picardie	21 640	6 016 992	0,36%
Grand Est	15 879	5 557 095	0,29%
Pays de la Loire	14 265	3 742 638	0,38%
Bretagne	8 317	3 309 220	0,25%
Aquitaine-Limousin-Poitou-Charentes	12 037	5 940 517	0,20%
Languedoc-Roussillon-Midi-Pyrénées	15 133	5 819 131	0,26%
Auvergne-Rhône-Alpes	33 509	7 933 200	0,42%
Provence-Alpes-Côte d'Azur et Corse	16 022	5 359 093	0,30%

Hormis l'outre-mer, la distribution est relativement uniforme suivant les régions.

Montants de dépenses réelles par famille d'acte

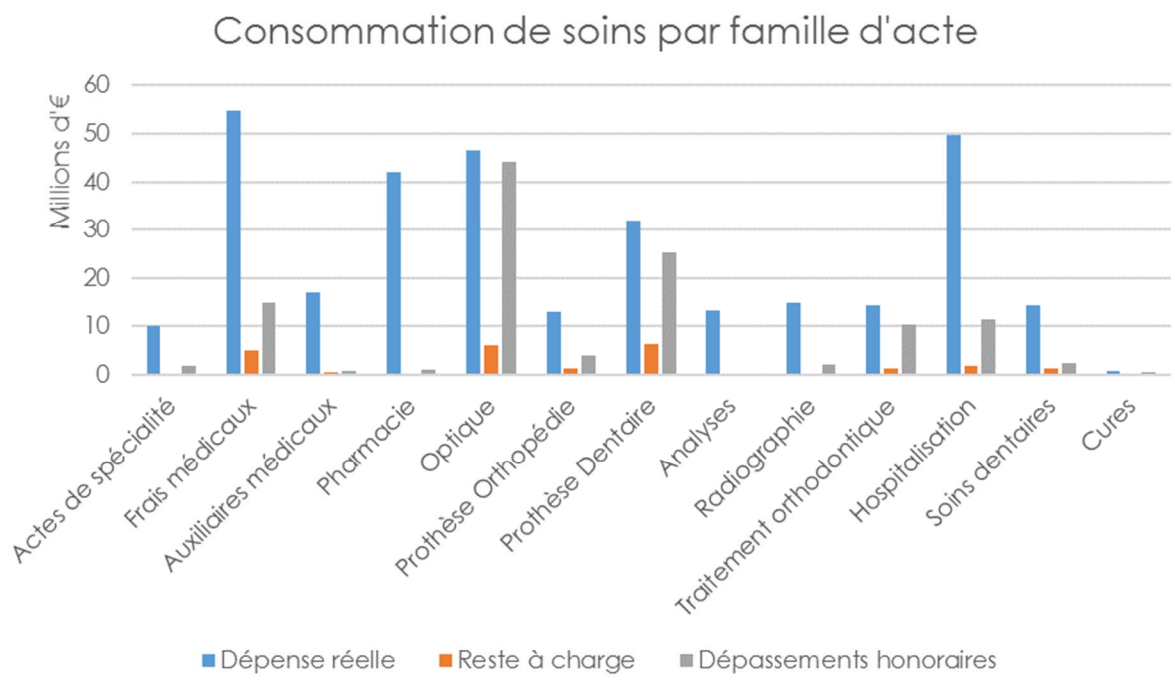


Figure 63 – Totaux de consommation de soin par famille d'acte dans notre base d'étude

A titre de comparaison, nous pouvons observer également la distribution, suivant les familles d'actes transcodées par notre table de correspondance, dans les données DAMIR :

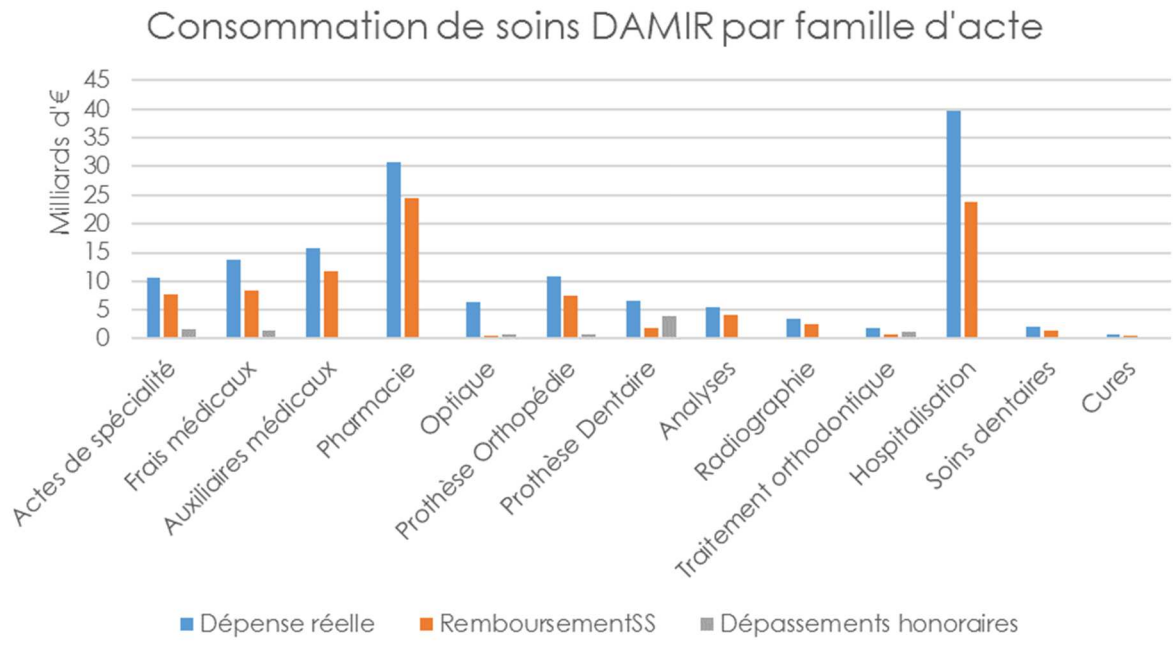


Figure 64 – Totaux de consommation de soin par famille d'acte dans la base DAMIR

Distribution des montants de dépense réelle

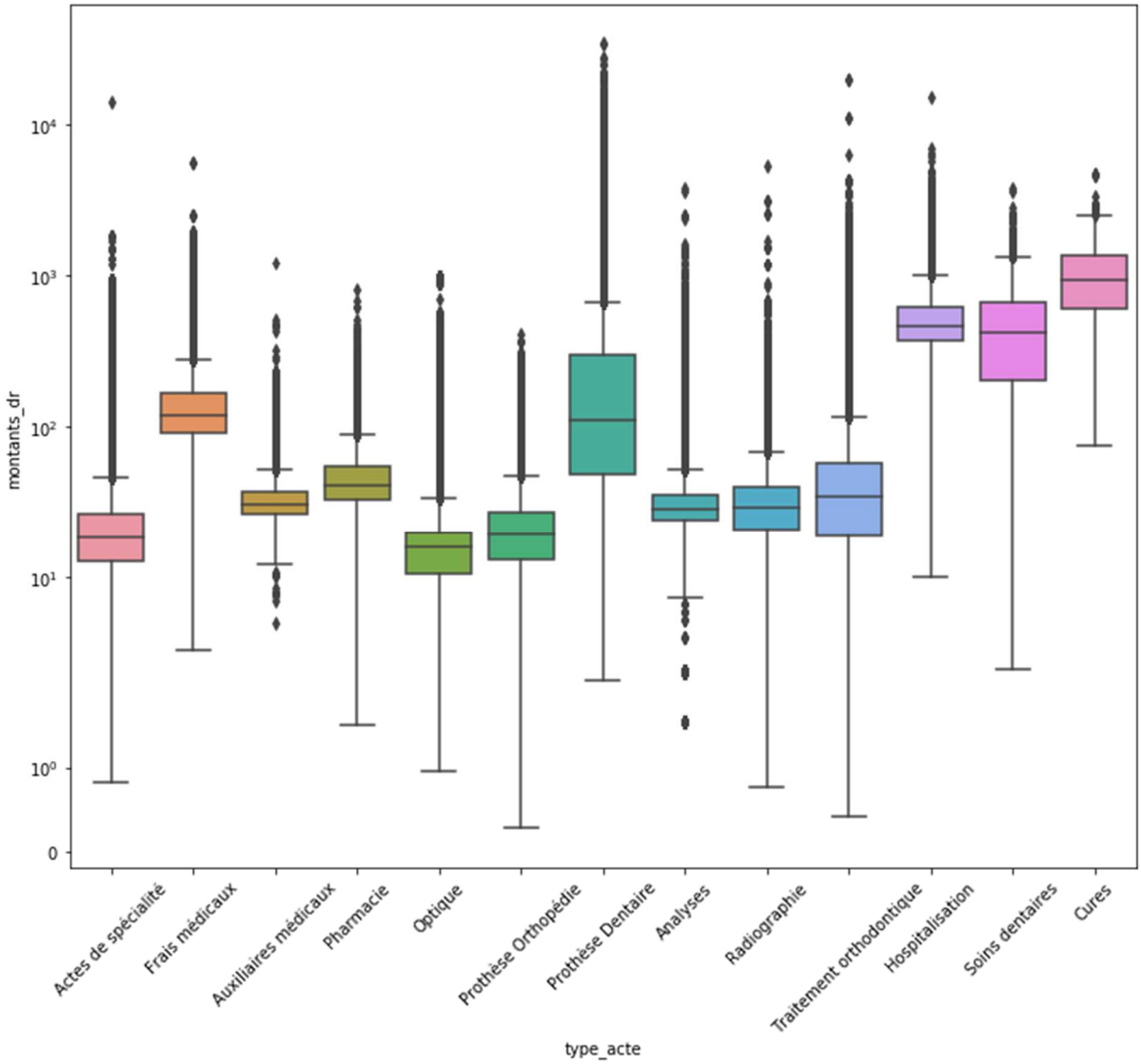


Figure 65 – Boites à moustaches de la dépense réelle ramenée à des actes unitaires et par famille d'acte dans notre base d'étude (échelle logarithmique)

On remarque en particulier les longues traines et certains montants bien plus élevés que la médiane.

Distribution et codistribution des montants

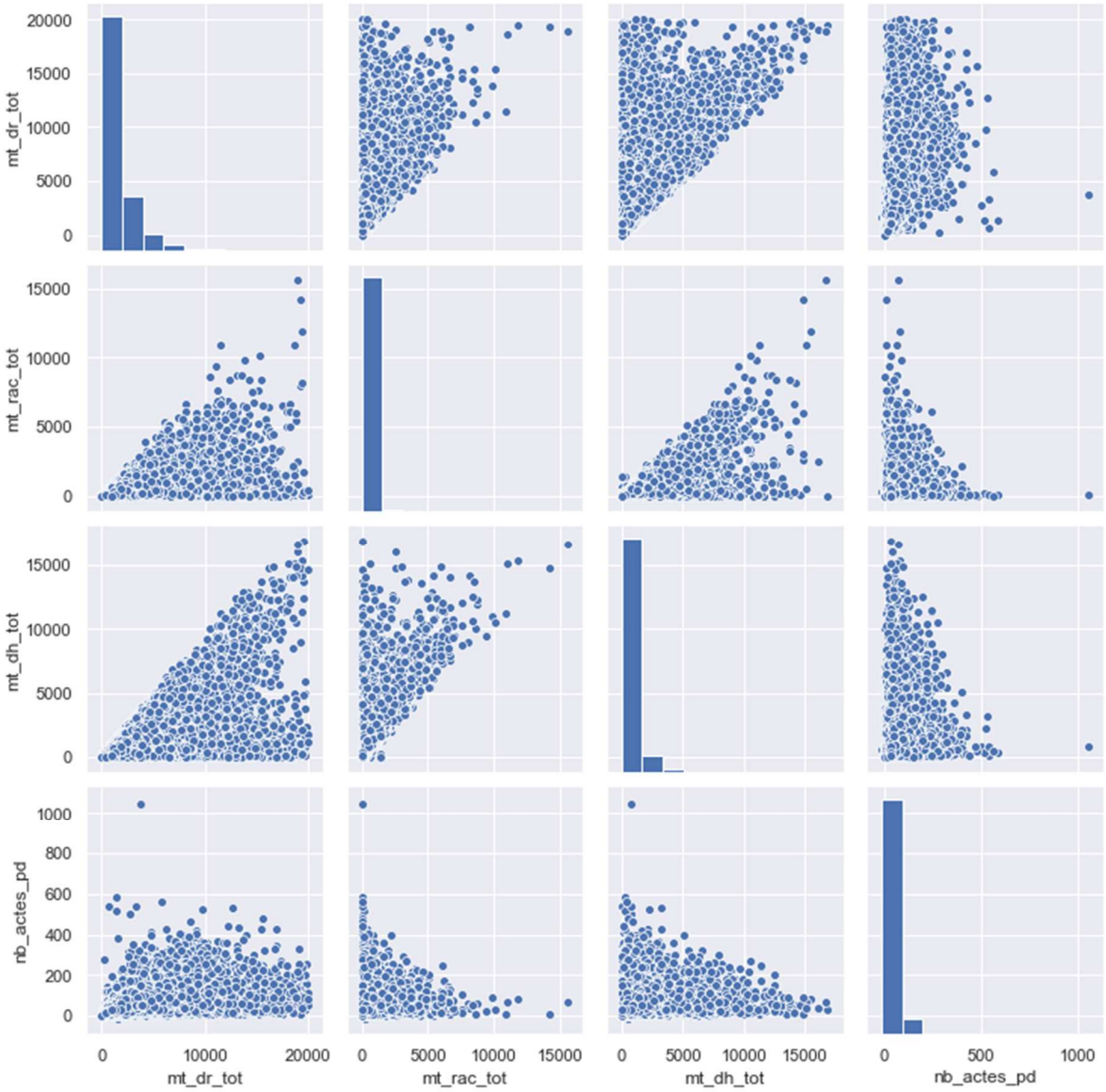


Figure 66 – Distribution et codistribution des montants de dépense réelle, reste à charge, dépassements d'honoraires et nombre d'actes de notre base d'étude

On remarque sans surprise les corrélations entre les différents montants par acte.

ANNEXE 10 : METHODES DE REECHANTILLONNAGE

Différentes méthodes existent pour effectuer un rééchantillonnage et elles répondent à des configurations de déséquilibre différentes. Pour les illustrer, nous proposons de faire tourner une petite sélection d'algorithmes sur des données fictives générées aléatoirement, déséquilibrées (1 pour 100), séparées en deux classes distinctes (0 et 1) et projetés sur deux axes :

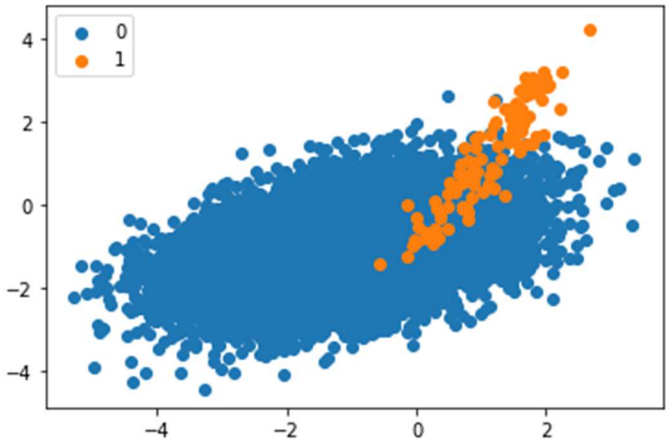


Figure 67 – Représentation d'une ACP sur individus de 2 classes déséquilibrées (0 et 1)

Méthodes de sous-échantillonnage

Les techniques de sous-échantillonnage permettent de retirer des données d'entraînement les exemples qui font partie de la classe majoritaire. Il peut s'agir d'une sélection aléatoire (*random undersampling*) mais une des limites est que par ce choix arbitraire, certains exemples importants pour déterminer la frontière entre les deux classes peuvent être retirés. Des informations utiles seraient alors supprimées.

Il existe donc des algorithmes de sous-échantillonnages qui permettent d'éviter cette suppression d'information, en tentant de supprimer en priorité les échantillons qui ne présentent que des informations redondantes avec le reste du *dataset* ou au contraire, en tentant d'identifier les échantillons importants pour éviter de les supprimer.

Les méthodes de sous-échantillonnage se répartissent en ces deux catégories :

- celles qui cherchent à identifier les échantillons à conserver. Il est en général possible de définir une cible sur un certain nombre d'échantillons à conserver pour équilibrer les données ;
- celles qui identifient les échantillons supprimables car ne contenant que de l'information redondante. Ici, les algorithmes balayent tout le *dataset* à la recherche d'échantillons à supprimer et cela ne suffit pas toujours à l'équilibrer.

Méthodes par conservation des échantillons les plus pertinents

Nous prenons ici en exemple l'algorithme *CondensedNearestNeighbour*. Cette méthode simule une classification des données par l'algorithme *kNN* et ne conserve que les échantillons qui apportent de l'information à cette classification pour éviter toute perte lors du sous-échantillonnage. Dit autrement, cela signifie qu'un échantillon ne permettant pas de classifier mieux un autre point que les autres échantillons déjà conservés pourra être supprimé. Elle est cependant sensible au bruit et très coûteuse en temps de traitement [SUTTON, O. (2012)].

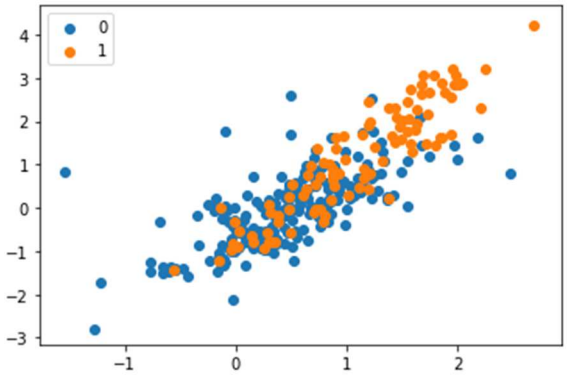


Figure 71 – Echantillons restants après application de CondensedNearestNeighbour

Méthodes par sélection d'échantillons à supprimer

Nous prenons ici en exemple l'algorithme *TomekLinks*. Cette méthode cherche à former des paires entre des échantillons très proches et de classes différentes. Une paire (ou *Tomek Link*) est constituée par A et B si :

- A est le plus proche voisin de B
- B est le plus proche voisin de A
- A et B sont de classes différentes

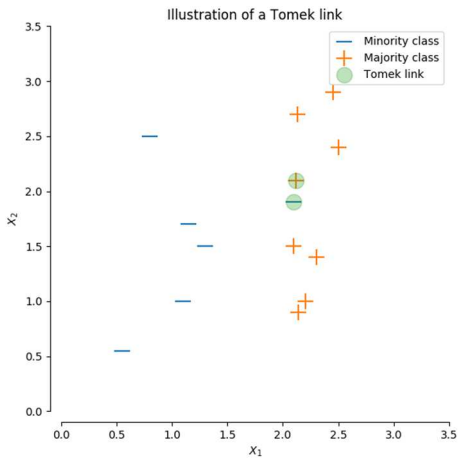


Figure 72 – Exemple d'identification de Tomek Link [imbalanced-learn.readthedocs.io]

A partir de là, il est possible de choisir la stratégie de sous-échantillonnage de manière à supprimer l'échantillon de la classe majoritaire ou bien de supprimer la paire complète. Ces paires sont intéressantes car elles sont particulièrement significatives[143] lorsqu'il s'agit de définir la frontière entre les classes.

Cette méthode ne retire cependant qu'un petit nombre d'échantillons (le bruit ainsi que les exemples les plus ambigus) et doit être combinée avec une autre méthode.

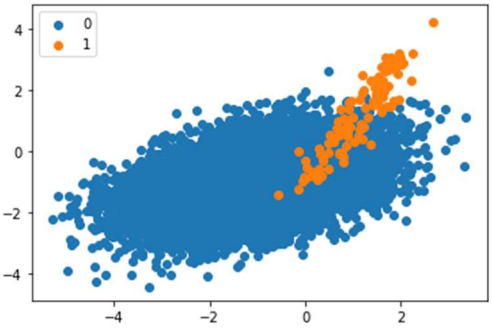


Figure 73 – Echantillons restants après application de TomekLinks

Méthodes de sur-échantillonnage

Les techniques de sur-échantillonnage sont l'exact opposé du sous-échantillonnage : il s'agit d'ajouter des échantillons d'entraînement extrapolés d'exemples de la classe minoritaire. De nouveaux échantillons sont donc générés sur la base de ceux déjà présents dans les données. Comme pour le sous-échantillonnage, il faut toutefois veiller à utiliser une méthode pertinente pour cela, le risque étant le surapprentissage ou l'ajout de bruit.

SMOTE (Synthetic Minority Oversampling TEchnique)

Cet algorithme génère de nouveaux échantillons de la classe minoritaire en appliquant la méthode suivante :

1. Identification des k plus proches voisins de même classe de chacun des échantillons x_i de la classe minoritaire

2. Choix aléatoire de x_{z_i} parmi ces k plus proches voisins

3. Création d'un nouvel échantillon x_{new} positionné aléatoirement entre ces deux points :

$$x_{new} = x_i + \lambda \times (x_{z_i} - x_i) \text{ avec } \lambda \longmapsto \mathcal{U}(0,1)$$

[143] Significatives en bien ou en mal : ces paires peuvent constituer un « support » de la frontière ou au contraire ajouter du bruit difficile à traiter par le classifieur. Il convient donc d'être vigilant par rapport à la configuration des deux classes des données étudiées afin que la stratégie de sous-échantillonnage choisie ne viennent pas dégrader l'information contenue dans les données.

La procédure peut être répétée de manière à générer autant de nouveaux échantillons que nécessaire.

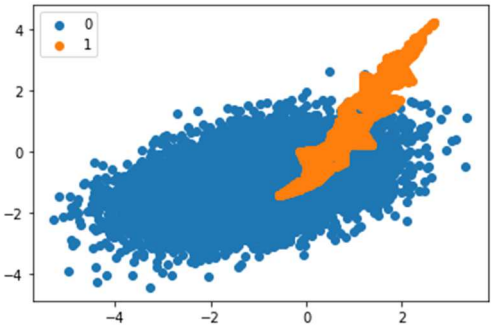


Figure 76 – ACP des données rééquilibrées par l'ajout des échantillons créés par SMOTE

Cette méthode est très intéressante, notamment combinée avec un sous-échantillonnage en vue de la définition de la zone de classification de la classe minoritaire. En effet, elle permet d'agrandir cette zone de classification grâce aux échantillons générés, ce qui a un effet positif sur les performances des modèles [CHAW, N. V. et al. (2002)].

Des variantes existent qui diffèrent principalement sur l'ajout de critères de sélection des échantillons qui vont servir de points de départ au sur-échantillonnage, notamment *BorderlineSMOTE*, qui cible comme échantillons de départ ceux dont la classification semble ambiguë (situés à la « frontière » des deux classes).

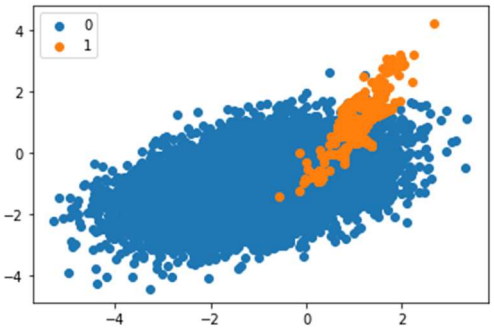


Figure 77 – ACP des données rééquilibrées par l'ajout des échantillons créés par BorderlineSMOTE

Sur-échantillonnage suivi par un sous-échantillonnage

Les deux approches ayant leurs avantages, il paraît intéressant de les combiner, en sur-échantillonnant la classe minoritaire et en sous-échantillonnant la majoritaire avec des algorithmes bien choisis. Par exemple, nous avons appliqué un sur-échantillonnage via SMOTE de manière à ce que la classe minoritaire soit représentée à 10% puis nous avons sous-échantillonné la classe majoritaire avec *One-Sided Selection*.

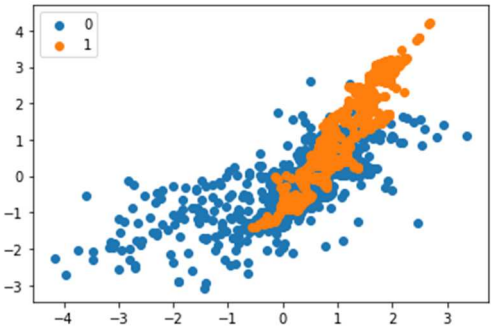


Figure 80 – Données rééchantillonnées par SMOTE puis One-Sided Selection

En conclusion, ces méthodes de rééchantillonnage présentent différentes caractéristiques et le choix d'appliquer l'une ou l'autre doit se faire en fonction des données. Est-ce que la frontière entre les classes est nette ? Y a-t-il beaucoup de bruit, d'échantillons qui semblent non significatifs ? Est-ce que le sur-échantillonnage ne va pas générer un surajustement du modèle ? Il est difficile de répondre à toutes ses questions sans tester les résultats du modèle lui-même [BATISTA, G. E. A. P. A. et al. (2004)], c'est pourquoi le choix et le paramétrage de la méthode de rééchantillonnage fera partie de la stratégie d'arbitrage à mettre en place.