UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÈLES DE PRÉDICTION DES COÛTS DES DIAGNOSTICS DE COMPLICATIONS DURANT LA GROSSESSE ET L'ACCOUCHEMENT

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN ÉCONOMIQUE

PAR

MARC-ANDRÉ GERALDO-DEMERS

DÉCEMBRE 2019

REMERCIEMENTS

J'aimerais remercier mon directeur Philip Merrigan pour sa patience, son soutien et ses conseils, qui ont permis d'alimenter ma réflexion et ont été une source de motivation majeure. En m'accordant sa confiance et une large indépendance, son encadrement m'a permis de repousser mes limites et de développer une grande autonomie.

J'aimerais remercier l'École des Sciences de la Gestion (ESG-UQÀM) et le département de sciences économiques pour leur soutien financier tout au long de ma maîtrise.

J'aimerais aussi remercier tous les professeurs et employés du département d'économique de l'ESG-UQÀM que j'ai eu la chance de côtoyer durant mon parcours à l'UQÀM. Je pense à Martine Boisselle-Lessard (UQÀM), Nicholas Lawson (UQÀM), Till Düppe (UQÀM), mais également à tous ceux et celles du 5e étage de l'ESG-UQÀM. La formation de grande qualité que vous offrez ainsi que la convivialité de l'environnement académique dans lequel j'ai pu évoluer sont remarquables.

J'aimerais également remercier ma famille, mes amis et tout mon entourage pour vos encouragements et votre soutien constant tout au long de ce projet. Vous avez été essentiels pendant mon parcours, et vous êtes la raison principale derrière cette réussite.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AUC	Aire sous la courbe ROC (*Area under the curve*)
ENN	Méthode des plus proches voisins actualisés
K-NN	Méthode des K plus proches voisins (K *Nearest Neighbors*)
ROC	*Receiver Operating Characteristics*
SMOTE	Technique de sur-échantillonnage synthétique de la classe majoritaire
SVC	*Support vector classifier*

RÉSUMÉ

La caractérisation des risques potentiels pour la santé d'une femme durant la grossesse et l'accouchement est une tâche primordiale puisque ces risques ont un impact direct sur la santé de l'enfant dès la naissance, en plus de potentiellement mettre en danger la vie de la patiente.

Les coûts engendrés par une mauvaise identification de ces risques peuvent avoir un poids important sur un système de santé en termes de ressources médicales et financières. En détectant plus rapidement si une patiente doit avoir recours à davantage de ressources le jour de son accouchement, les risques potentiels pour sa santé peuvent être minimisés en plus de permettre un gain d'efficience dans la gestion des ressources.

L'objectif de ce mémoire est de faire la prédiction des coûts des diagnostics de complications durant la grossesse et l'accouchement. Les données utilisées proviennent d'un État nord-américain et elles contiennent les dossiers médicaux de 54 000 femmes qui ont donné naissance entre les années 1998 à 2006. Les prédictions sont réalisées avec des méthodes d'apprentissage automatique et de ré-échantillonnage.

Les résultats obtenus démontrent que les méthodes d'apprentissage automatique et de ré-échantillonnage permettent la prédiction des complications. Les meilleurs modèles pour prédire les complications et les coûts sont les modèles de forêt aléatoire et AdaBoost. Les variables les plus importantes pour prédire les complications sont l'âge de la mère, les complications antérieures, les maladies transmises sexuellement, l'hypertension et le diabète.

MOTS CLÉS : santé, grossesse, accouchement, apprentissage automatique, méthodes de ré-échantillonnage, système de santé, coûts des soins de santé.

INTRODUCTION

Au Québec, le système de santé est régi en grande partie par le gouvernement provincial et est financé par les contribuables. Le gouvernement fait face à un choix entre investir aujourd'hui dans la prévention ou possiblement assumer une augmentation importante des coûts futurs. Ainsi, les prédictions de coûts sont nécessaires au bon fonctionnement du système de santé. L'intérêt économique derrière un tel exercice est d'obtenir de meilleures prévisions budgétaires pour les diagnostics entourant les grossesses et une meilleure gestion des ressources médicales.

La santé d'un individu durant les premières années de vie est primordiale puisqu'elle détermine le développement cognitif de l'enfant jusqu'à l'âge adulte. Les enfants grandissant dans un environnement favorable à leur santé dès la naissance sont susceptibles d'être plus productifs à l'âge adulte. Law *et al.* (2015) ont montré qu'un nouveau-né en bonne santé a de fortes chances d'être moins coûteux en termes de soins de santé, en plus de gagner de meilleurs salaires à l'âge adulte. La santé des femmes avant qu'elles deviennent enceintes et durant la grossesse est tout aussi importante, car leur état de santé a un impact direct sur la santé de l'enfant à la naissance et peut donc avoir des conséquences de santé publique sur le long terme. Les femmes ayant de lourds antécédents médicaux sont plus susceptibles d'avoir des complications pendant la grossesse. Les coûts supplémentaires causés par ces patientes peuvent être importants pendant et après la grossesse, en nécessitant davantage de visites, de prescriptions et de soins. Par ailleurs, Law *et al.* (2015) ont montré que les complications durant la grossesse et l'accouchement

sont communes et certaines d'entre elles peuvent amener une hausse des coûts de soins de santé importante pour les enfants immédiatement après l'accouchement.

Les quantités massives de données administratives sur les soins de santé récoltées par les gouvernements représentent une source d'information riche et importante pouvant servir à quantifier les coûts de différentes catégories de soins durant la grossesse. Lorsqu'elle est correctement exploitée, cette masse de données peut informer les intervenants et améliorer l'état de santé des patients tout en réduisant les coûts (Adam *et al.*, 2017). Ces données peuvent permettre d'identifier les différentes complications en utilisant les antécédents médicaux de plusieurs patientes pour ensuite faire une prédiction des diagnostics et déterminer les ressources financières et opérationnelles nécessaires pour les traitements associés à ces derniers. Dû à la quantité importante de données collectées, l'utilisation de ces sources d'information amène des problèmes particuliers nécessitant des outils différents des méthodes économétriques généralement utilisées pour faire ce type de prédiction (Varian, 2014). De plus, le nombre important d'observations et de variables explicatives nécessitent des outils puissants permettant de faire une sélection des variables servant à faire la prédiction (Varian, 2014).

Une meilleure gestion des ressources a plusieurs avantages pour les médecins et pour les administrateurs des établissements de santé. Par exemple, en détectant plus rapidement si une patiente doit avoir recours à un spécialiste ou si elle doit se rendre à un endroit où davantage de soins sont disponibles, ils auront une meilleure évaluation de la situation et des soins nécessaires avant l'accouchement. Il en résultera donc une meilleure prévention des complications possibles en lien avec la grossesse et l'accouchement, ce qui causera possiblement un gain d'efficience en termes de gestion des ressources, tout en ayant potentiellement un effet bénéfique à long terme sur la santé des femmes et des enfants. Des enfants et des mères avec une meilleure santé seront moins coûteux pour les établissements de santé, ce qui

permettra de débloquer des ressources financières additionnelles dans le futur tout en contribuant à l'objectif de pérennisation de la qualité du système de santé. De plus, les algorithmes servant à la prévision des diagnostics peuvent être un outil d'aide à la décision pour les médecins au moment de poser des diagnostics. En analysant des dizaines de milliers de cas simultanément et en traitant une quantité d'information importante n'étant pas disponible au médecin lors des rendez-vous prénataux, ces algorithmes sont susceptibles d'aider les médecins à poser de meilleurs diagnostics.

La présente recherche a pour but la prédiction des coûts des diagnostics de complications durant la grossesse à l'aide de données administratives provenant d'un État nord-américain. Ces données contiennent de l'information relative aux dossiers médicaux, aux programmes d'assurance médicament et aux coûts des prescriptions et des actes médicaux pour 54 000 femmes ayant donné naissance durant les années 1998 à 2006. Pour chaque patiente, l'historique médical est disponible deux ans avant l'accouchement et cinq ans après. Les prévisions sont effectuées un mois avant l'accouchement à partir de l'historique médical de chacune des patientes en considérant les coûts associés aux soins de santé de ces dernières [1].

En premier lieu, il sera question de déterminer les diagnostics de complications les plus fréquents et les coûts associés à ces derniers afin de créer la variable dépendante pour les modèles. En second lieu, une prédiction des diagnostics de complications un mois avant l'accouchement sera effectuée en utilisant les antécédents médicaux des patientes. Ensuite, cette prévision sera utilisée pour prévoir le coût moyen des soins de santé associés à ces diagnostics. L'objectif est donc d'utiliser les prédictions de diagnostics afin de prédire les risques associés à la santé de la patiente ainsi que les coûts des soins de santé le jour de l'accouchement et

1. Cette période a été choisie puisqu'elle permet d'utiliser l'information disponible durant la période prénatale sans affecter la qualité de la prédiction.

les mois suivants. En prévoyant quelles patientes sont à risque d'avoir des complications durant l'accouchement, il est possible de personnaliser l'offre de soins de santé avant même que la patiente se présente dans un établissement de santé pour son accouchement.

Les méthodes utilisées pour la prédiction sont des méthodes d'apprentissage automatique. L'avantage de ces méthodes est qu'elles parviennent à découvrir des relations généralisables, avec des formes fonctionnelles complexes n'étant pas spécifiées à l'avance et qui travaillent bien hors de l'échantillon (Mullainathan et Spiess, 2017). La combinaison de ces méthodes avec des données médicales permettra de concilier différentes sources d'information importantes et d'avoir une idée approfondie des conséquences budgétaires et pratiques liées aux complications durant la grossesse. En identifiant les patientes à risque d'augmenter la fréquence de leurs visites, de leurs prescriptions et l'importance des soins en lien avec ces diagnostics, les prévisions effectuées à l'aide de ces méthodes peuvent aider à rendre plus efficiente l'utilisation des ressources et permettre aux établissements œuvrant dans le réseau de la santé de sauver des sommes annuelles importantes (Adam *et al.*, 2017).

L'ensemble d'information entre le premier jour du dernier mois avant l'accouchement et le jour de l'accouchement ne sera pas utilisé pour la prédiction afin de prédire les diagnostics de complications. La variable dépendante est une variable binaire indiquant si la patiente est à risque d'avoir au moins un diagnostic de complications[2]. La prédiction de la variable dépendante binaire sera utilisée afin de prédire les coûts des soins de santé. Les variables explicatives servant à faire l'exercice de prévision sont des variables binaires indiquant le groupe d'âge de la patiente, le régime d'assurance médicament, différentes variables pour représenter

2. Plusieurs diagnostics de complications différents ont été identifiés dans les données afin de créer la variable dépendante.

les coûts des soins de santé, les visites à l'hôpital, les prescriptions, les durées de traitement pour les différents médicaments, les actes médicaux et les diagnostics.

À partir des données disponibles sur les visites et les prescriptions, un panel a été créé afin de représenter l'évolution des coûts des soins de santé reçus à travers le temps pour les patientes dans l'échantillon. Ces coûts seront estimés en utilisant les montants facturés par les établissements qui offrent les soins de santé. Ces données permettront de représenter la trajectoire de coûts de la patiente pour les vingt-trois premiers mois de la période d'observation. L'intuition économique derrière cette méthode est que la trajectoire de coûts représente une bonne approximation des conditions de santé puisque les patientes les plus coûteuses sont celles qui seront généralement en moins bonne santé. Ces trajectoires de coûts serviront à créer un ensemble de variables explicatives pour l'entrainement des modèles.

Le présent mémoire débutera avec une revue de littérature sur l'hypothèse des origines fœtales, les déterminants des complications durant la grossesse et la prédiction des coûts et des diagnostics. Le deuxième chapitre présentera la méthodologie. Le troisième chapitre présentera les données. Le quatrième chapitre présentera les résultats. Le cinquième chapitre présentera une discussion sur les différents résultats obtenus. Finalement, le dernier chapitre présentera la conclusion.

CHAPITRE I

REVUE DE LITTÉRATURE

1.1 Hypothèse des origines fœtales

Une des hypothèses principales derrière la littérature sur le lien entre les conditions *in utero* et la santé de l'enfant suite à la naissance a été établie par le médecin et épidémiologiste David J. Barker (Almond et Currie, 2011). Ce dernier argumentait que les neuf mois durant la période prénatale représentent une des périodes les plus importantes dans la vie d'un individu puisqu'elle affecte les habiletés futures de l'enfant, les conditions de santé jusqu'à l'âge adulte et le statut socioéconomique. Barker (1990) argumente que les conditions intra-utérines programment le métabolisme du fœtus et peuvent causer le développement de certaines maladies dans le futur. Par exemple, la sous-nutrition fœtale en milieu de gestation tardive conduit à une croissance fœtale disproportionnée qui programme le fœtus de sorte qu'il soit favorable pour le développement de maladie coronarienne (Barker, 1995).

Cette hypothèse, appelée l'hypothèse des origines fœtales, suggère donc que le fœtus passe à travers certaines périodes qui sont décisives pour son développement. Barker (1997) argumente que les fœtus doivent s'adapter à une quantité limitée de nutriments, et ce faisant ils changent de manière permanente leur physiologie et leur métabolisme. Ces changements peuvent être l'origine d'un bon nombre

de maladies survenant plus tard dans la vie de l'individu, incluant les maladies coronaires, mais également les troubles connexes aux arrêts cardiovasculaires, le diabète et l'hypertension (Barker, 1997). Dans leurs recherches empiriques pour tenter de vérifier l'hypothèse, Barker *et al.* (1989) ont utilisé des données nationales provenant de la Grande-Bretagne et ont montré que le lien entre l'environnement *in utero* et la pression artérielle à l'âge adulte explique possiblement une part de la variation dans les taux de mortalité due aux maladies cardiovasculaires.

Almond et Currie (2011) ont résumé l'hypothèse des origines fœtales en trois idées principales. Premièrement, ils argumentent que les effets des conditions fœtales sont persistants. Deuxièmement, ils argumentent que les effets de ces conditions sur la santé de l'individu peuvent demeurer latents pendant plusieurs années[3]. Troisièmement, ils argumentent que les effets hypothétiques des conditions fœtales reflètent un mécanisme biologique spécifique, soit la programmation du fœtus. Ce mécanisme se manifesterait à travers les effets de l'environnement sur l'épigénome, qui commencent tout juste à être compris selon les auteurs. Une acceptation complète de l'hypothèse des origines fœtales aurait des implications radicales pour les décisions individuelles et les politiques publiques, suggérant par exemple que le meilleur moment pour intervenir dans le but d'améliorer la vie des enfants est avant même qu'ils soient nés, et peut-être avant même que les mères se rendent compte qu'elles sont enceintes (Almond et Currie, 2011).

Plusieurs auteurs ont tenté de tester empiriquement cette hypothèse, le plus souvent à partir d'expériences naturelles. Par exemple, en utilisant l'épidémie d'influenza qui est survenue aux États-Unis en 1918 afin de déterminer si l'exposition prénatale a eu des effets sur les conditions économiques à l'âge adulte, Almond

3. En prenant comme exemple le développement de maladies cardiaques, les auteurs expliquent qu'un individu peut avoir des conditions favorables pour le développement de ces maladies, mais qu'elles ne surgissent pas avant un certain âge. La probabilité d'avoir une maladie cardiaque est donc une variable latente inobservable.

(2006) trouve un effet négatif important sur l'ensemble des mesures économiques pour les années de recensement de 1960, 1970 et 1980. Almond (2006) démontre notamment que les enfants *in utero* pendant l'épidémie ont un niveau d'éducation plus faible et que les enfants qui sont nés de mères qui ont été infectées sont 15 % moins susceptibles d'obtenir leur diplôme d'études secondaires. Les salaires des hommes étaient entre 5 % et 9 % plus faibles à cause de l'infection (Almond, 2006). Les dépenses publiques ont également augmenté : ceux qui se trouvaient dans leur premier trimestre de gestation pendant le pic de la pandémie ont connu les paiements d'aide sociale les plus élevés, soit d'avril 1911 à avril 1925 (Almond, 2006). De plus, le statut socioéconomique est beaucoup plus faible et la probabilité d'être pauvre est supérieure de 15 % lorsque comparée avec d'autres cohortes de naissances dont les mères n'ont pas été exposées à l'épidémie (Almond, 2006).

Almond et Mazumder (2005) ont également trouvé des résultats similaires en s'intéressant à la même expérience naturelle. Ils ont démontré que la cohorte *in utero* pendant l'exposition à la pandémie a des moins bonnes conditions de santé par rapport aux cohortes nées quelques mois plus tôt ou plus tard. Almond et Mazumder (2005) constatent également que le cancer, les problèmes cardiaques, les problèmes rénaux, l'hypertension et les problèmes d'estomac sont tous statistiquement significatifs pour les individus qui étaient *in utero* pendant la pandémie et dont le mois de naissance est au plus fort de la pandémie.

Almond et Mazumder (2011) ont utilisé le ramadan comme expérience naturelle afin de tester l'hypothèse que les conditions de santé prénatale, dans ce cas-ci un jeûne, ont un effet sur la santé de l'enfant à la naissance. Ils ont montré que les enfants qui sont nés d'une grossesse qui coïncide avec le ramadan ont de moins bonnes conditions de santé et sont économiquement désavantagés par rapport à leurs pairs où la grossesse ne coïncide pas avec le ramadan. Les taux d'invalidité chez les adultes qui sont nés de mères pour lesquelles la grossesse a croisé le

ramadan sont environ 20 % plus élevés, avec des handicaps mentaux spécifiques qui ont des effets beaucoup plus importants sur leur santé (Almond et Mazumder, 2011). Ainsi, leurs résultats suggèrent que le jeûne et la réduction de l'apport calorique durant la grossesse ont des conséquences à long terme sur la santé de l'enfant, et ce jusqu'à l'âge adulte.

Certains auteurs se sont intéressés aux autres facteurs susceptibles d'avoir un effet sur les conditions socioéconomiques de l'enfant une fois à l'âge adulte. Parmi ces facteurs, on retrouve notamment l'exposition au stress *in utero* et le gain de poids durant la grossesse. Aizer *et al.* (2009) ont montré que l'exposition *in utero* à des niveaux élevés de cortisol, soit l'hormone qui cause le stress, affecte négativement la cognition, la santé et le niveau d'éducation de la progéniture une fois à l'âge adulte. En ce qui concerne le gain de poids durant la grossesse, Ludwig et Currie (2010) ont montré qu'il existe une relation positive entre le gain de poids durant la grossesse et le poids de l'enfant à la naissance, et ce indépendamment des facteurs génétiques. Puisque le poids élevé à la naissance prédit l'indice de masse corporelle à l'âge adulte, un gain de poids excessif pendant la grossesse pourrait augmenter le risque d'obésité chez les enfants une fois à l'âge adulte et augmenter le risque de développer d'autres maladies (Ludwig et Currie, 2010).

De meilleures conditions prénatales causeraient non seulement des enfants en bonne santé, mais également des individus qui seraient en meilleure santé à l'âge adulte. Dans ce contexte, Costa et Lahey (2005) se sont intéressés à la relation entre l'amélioration des conditions prénatales et postnatales par rapport au vieillissement de la population. En comparant une cohorte d'individus âgés de 60 ans en 1900 et une cohorte environ du même âge entre 1960 et 1980 aux États-Unis, ils ont trouvé que l'amélioration de ces conditions a eu une contribution se chiffrant entre 16 % et 17 % au déclin des taux de mortalité de dix ans. Combinés aux résultats précédents, ces résultats suggèrent que le lien qui existe entre

le vieillissement de la population et l'amélioration des conditions *in utero* passe par une amélioration des conditions de santé dès la naissance. L'amélioration des conditions postnatales est susceptible de diminuer le risque que l'enfant contracte des maladies pendant une période qui est critique pour son développement. Case et Paxson (2009) ont montré que la présence d'un environnement infectieux dans les premières années de vie est associée avec de plus faibles résultats aux tests cognitifs tard à l'âge adulte. Les résultats de Case et Paxson (2009) suggèrent donc que les facteurs environnementaux, notamment la contagion des lieux, sont associés avec le développement cognitif à l'âge adulte.

En démontrant empiriquement l'hypothèse de Barker, ces études suggèrent que les conditions *in utero* sont importantes pour le développement de l'enfant jusqu'à l'âge adulte. La vérification de cette hypothèse motive la prédiction des complications durant la grossesse dans le sens où, s'il est possible de prévoir une détérioration des conditions foetales avant le jour de l'accouchement, ceci permettrait d'améliorer la prévention et d'investir plus rapidement dans la santé de la femme en cas de prévision positive de diagnostic de complications.

1.2 Déterminants des complications durant la grossesse

Plusieurs études se sont intéressées aux déterminants et aux différentes conditions de santé associés aux complications durant la grossesse. Une grande partie de ces études s'intéresse principalement à des effets causaux dans le but d'améliorer la compréhension de ces déterminants et leurs effets sur la santé de la mère et l'enfant pendant la grossesse et suite à l'accouchement. Parmi ces déterminants, on retrouve les troubles d'obésité, l'âge maternel, l'hypertension, le diabète, la consommation de tabac et les déterminants socioéconomiques.

Ananth et Vintzileos (2006) ont comparé les accouchements prématurés aux accouchements à terme afin de déterminer les conditions de santé maternelle et fœtale qui résultent en accouchements prématurés[4]. La prééclampsie, les détresses fœtales, la petite taille pour l'âge gestationnel et la rupture du placenta étaient les indicateurs médicaux les plus fréquents pour une intervention médicale résultant en accouchement prématuré (Ananth et Vintzileos, 2006). Le taux d'accouchements prématurés était de 4,6 % avec 23,5 % de ces accouchements médicalement indiqués. Considérant la taille de l'échantillon, ceci représente un nombre important de cas[5]. Les auteurs ont également observé que les proportions de femmes âgées de moins de 20 ans, avec moins de douze années d'éducation, de race noire, célibataires et qui ont une consommation positive de tabac étaient plus grandes parmi le groupe de femmes qui ont un accouchement prématuré comparativement à celles qui ont un accouchement à terme. Ces résultats suggèrent que le statut socioéconomique de la mère peut aider à déterminer quelles patientes sont susceptibles d'avoir un accouchement prématuré.

Kramer *et al.* (2000) se sont intéressés aux déterminants socioéconomiques des accouchements prématurés et des restrictions de croissance intra-utérine. Ils ont démontré que vivre dans la pauvreté conduit à une accumulation de multiples facteurs de stress chroniques qui sont liés à la condition socioéconomique de la mère. Ceci peut augmenter de manière coordonnée le risque d'issues défavorables durant la grossesse dans une mesure beaucoup plus grande que ce qui peut être expliqué par les contributions étiologiques individuelles des mères (Kramer *et al.*, 2000). Les principaux facteurs quantitativement importants en ce qui concerne les disparités socioéconomiques dans la restriction de la croissance intra-utérine sont

4. Les accouchements prématurés sont déterminés par les auteurs comme étant des accouchements qui ont lieu à moins de 35 semaines de gestation.

5. L'échantillon utilisé provient du *Missouri's live birth data files* durant la période entre 1989 et 1997. Leur échantillon contient 684 711 naissances après les exclusions.

le tabagisme, le faible gain de poids gestationnel et la petite taille à la naissance (Kramer *et al.*, 2000). Les auteurs argumentent également que les femmes avec des statuts socioéconomiques faibles sont plus enclines à consommer du tabac, de l'alcool et des drogues. Les diagnostics d'abus de tabac, d'usage de drogue et d'abus d'alcool peuvent donc aider à approximer le statut socioéconomique de la mère en captant une partie de l'effet du niveau d'éducation sur les habitudes de santé.

Cavazos-Rehg *et al.* (2015) se sont intéressés au lien entre l'âge maternel et le risque de complications durant la grossesse et l'accouchement. Ils ont démontré que le risque de complication est plus grand pour les femmes âgées entre 11-14 ans, 15-19 ans et les femmes qui ont 35 ans et plus. Ils ont également démontré que l'effet de l'âge maternel sur les complications est hétérogène, soit que le risque n'est pas associé aux mêmes complications selon l'âge de la mère (Cavazos-Rehg *et al.*, 2015). Contrairement à ce qui est avancé par Kramer *et al.* en ce qui concerne les variables socioéconomiques, Cavazos-Rehg *et al.* n'ont pas trouvé de lien significatif après avoir contrôlé pour les facteurs démographiques, le revenu et la nationalité.

Quelques études se sont également intéressées à la relation causale entre les complications et les coûts. Galtier-Dereure *et al.* (2000) ont démontré que le coût moyen des soins prénataux et postnataux est plus élevé pour les enfants nés de mères en surpoids que pour les mères de poids normal, et les enfants nés de mères en surpoids sont plus souvent admis dans les unités de soins intensifs néonatals que les enfants nés de mères de poids normal. De plus, ils ont trouvé que les troubles d'hypertension sont significativement plus fréquents chez les femmes enceintes en surpoids que chez leurs homologues en santé. L'obésité entraîne également des séjours hospitaliers postpartum significativement plus longs en raison d'accouchements avec césarienne plus fréquents, en plus de causer un risque plus

élevé d'infections pour la mère et de surpoids pour ces enfants à l'âge de 12 mois (Galtier-Dereure *et al.*, 2000). Cette étude démontre également que lorsque le diabète complique le cours de la grossesse, les nouveau-nés sont prédisposés à développer un surpoids et de l'obésité pendant l'enfance, en particulier dans le cas d'un poids élevé à la naissance.

Law *et al.* (2015) se sont intéressés au lien causal entre les complications durant la grossesse et les coûts des soins de santé dans les trois premiers mois de vie de l'enfant. Parmi un échantillon de 137 040 enfants, 75,4 % sont des enfants qui sont nés de mères ayant au moins une complication durant la grossesse ou l'accouchement. Les complications les plus fréquentes sont les anomalies du fœtus (26,2 %), le travail hâtif ou menacé (16,6 %), les hémorragies (10,8 %), le diabète (8,0 %) et l'hypertension (7,7 %). Similairement aux résultats obtenus par Galtier-Dereure *et al.* et Cavazos-Rehg *et al.*, Law *et al.* ont trouvé qu'il existe une relation entre l'âge maternel et le nombre de femmes avec des problèmes de santé chroniques ou des problèmes de santé en lien avec la grossesse. Les accouchements prématurés sont plus communs chez les femmes avec de l'hypertension, du diabète, des problèmes de dépendance à la drogue et à l'alcool, des problèmes de santé mentale, des problèmes alimentaires, des maladies cardiovasculaires, des maladies rénales et des maladies thyroïdes (Law *et al.*, 2015).

En comparant le coût des soins de santé pour les nouveau-nés durant les trois premiers mois avec le coût des soins pour les enfants nés de mères avec et sans complications, Law *et al.* (2015) ont révélé que dans la plupart des cas, le coût des soins est plus élevé pour les enfants qui sont nés de mères ayant eu des complications durant la grossesse et l'accouchement, allant de 132 $ jusqu'à 44 720 $ pour les gestations multiples et les troubles cardiovasculaires congénitaux (Law *et al.*, 2015). De plus, ils ont démontré que l'utilisation de ressources médicales était significativement plus élevée pour les enfants nés de mères ayant eu des

complications durant la grossesse et l'accouchement, pour toutes les catégories de complications confondues (Law *et al.*, 2015).

1.3 Prédiction des coûts et des diagnostics

Plusieurs chercheurs se sont attardés aux prévisions des coûts et des diagnostics en santé à l'aide de méthodes en apprentissage automatique. Milovic (2012) s'est intéressé aux méthodes possibles pour faire des prévisions en santé et aux avantages d'utiliser des méthodes en apprentissage automatique. Un avantage important mentionné par ce dernier est qu'une fois que l'exploitation de données est intégrée dans les systèmes d'information, les établissements de santé peuvent réduire la subjectivité dans les prises de décisions médicales et fournir de nouvelles connaissances. Les problèmes de prédiction dans un contexte de prise de décision face à un diagnostic peuvent bénéficier considérablement de méthodes prédictives. De plus, les algorithmes prédictifs peuvent servir de diagnostic comportemental, nous aidant à comprendre la nature de l'erreur humaine (Kleinberg *et al.*, 2015).

Milovic (2012) s'est intéressé aux méthodes possibles pour faire des prédictions en santé et aux avantages d'utiliser des méthodes d'exploration de données (*data mining*) et des méthodes d'apprentissage automatique. Il argumente que l'utilisation des technologies de l'information permet une amélioration du système de santé et de la santé publique en offrant des meilleurs soins de santé aux utilisateurs du système tout en réduisant les coûts et en économisant du temps (Milovic, 2012). De plus, les prévisions effectuées à partir de ces méthodes peuvent réduire l'erreur subjective induite par le comportement humain en analysant une grande masse d'information dans le but d'aider un spécialiste à prendre une décision affectant la santé d'un individu (Milovic, 2012).

Un exemple d'application de ces méthodes afin de réduire l'erreur induite par

de mauvais diagnostics est une étude réalisée par Kleinberg *et al.* (2015). Dans leur étude, les auteurs proposent de prévoir la probabilité de mortalité afin de déterminer si un individu devrait recevoir une chirurgie ou non. Puisqu'il y a une désutilité à subir la chirurgie pour le patient s'il n'y a pas de réel risque de mortalité, en plus des coûts monétaires imposés sur le système de santé, la prédiction pourrait aider à déterminer si celle-ci est rentable ou non avant même qu'elle soit donnée au patient. Pour répondre à cette question, Kleinberg *et al.* (2015) ont prédit la probabilité de mortalité dans les douze mois suivants la chirurgie avec des variables sur le sexe, l'âge, la région géographique et l'utilisation des soins de santé. Ils ont aussi utilisé des variables sur les comorbidités, les symptômes, les blessures, les conditions aiguës et leur évolution à travers le temps.

En simulant la réception des chirurgies futiles par des personnes qui ont une probabilité de mortalité suffisamment élevée plutôt que de les donner à ceux qui ont une probabilité faible, soit le 10 % plus risqué avec les patients moins risqués éligibles, ils démontrent que 10 512 chirurgies futiles auraient pu être évitées et que 158 millions de dollars par année auraient pu être réalloués vers des patients qui bénéficieraient de la chirurgie. Suite à ces résultats, Kleinberg *et al.* (2015) suggèrent d'aborder les problèmes d'analyse de politiques publiques avec des méthodes de prédiction plutôt qu'avec des méthodes d'inférence causale.

Bertsimas *et al.* (2008) se sont intéressés à faire une prédiction des coûts de soins de santé en utilisant la trajectoire des coûts antérieurs et l'historique médical. Afin de faire la prédiction des coûts, les auteurs ont utilisé des méthodes modernes d'exploration de données, soit les arbres de classification et les algorithmes de regroupement. Les algorithmes de regroupement sont utilisés dans le but d'identifier les patients qui ont des caractéristiques de coûts similaires afin de construire différents groupes pour ensuite utiliser les caractéristiques médicales pour créer des sous-groupes à l'intérieur de ceux-ci, et finalement faire la prédiction de la

catégorie de coûts. Ils ont effectué leurs prévisions en utilisant un échantillon test, soit un sous-ensemble de l'échantillon initial contenant des exemples qui n'ont pas été utilisés pour la calibration des modèles. Cet échantillon est conservé jusqu'à la fin de l'estimation afin d'évaluer la performance des modèles lorsqu'ils rencontrent de nouvelles observations.

En utilisant des données provenant de réclamations d'assurances aux États-Unis, Bertsimas *et al.* démontrent que les méthodes d'exploration de données fournissent des prévisions précises pour les coûts médicaux et que ces méthodes représentent un outil puissant pour réaliser cet objectif. Pour chaque patient, ils disposent de trois ans d'historique médical. Ils utilisent l'information relative aux coûts des deux premières années afin de prévoir les coûts pour la troisième année[6]. Les auteurs procèdent par validation croisée pour calibrer les modèles, c'est-à-dire qu'ils séparent leur échantillon en plusieurs échantillons aléatoires tout en conservant un sous-échantillon afin de tester les modèles. Le modèle final est ensuite estimé avec l'échantillon test afin d'en évaluer la performance.

Les variables indépendantes sont les différents groupes de diagnostics, les groupes de procédures médicales, les groupes de médicaments, et des facteurs de risque qu'ils ont développés[7]. Vingt-deux variables de coûts ont été utilisées : les coûts mensuels pour les douze derniers mois individuellement, le coût total pour les médicaments, le coût total pour les services médicaux, le coût global pour les trois et pour les six derniers mois, une variable qui capture si une patiente a un bond soudain dans ses coûts, soit un choc temporaire de dépenses en soins de santé (point de vue de la patiente), le coût mensuel maximum, le nombre de mois

6. La troisième année représente l'échantillon auquel les auteurs font référence comme étant l'échantillon test.

7. Leur base de donnée contient environ 1 500 variables explicatives. Les codes de diagnostics et les codes de procédures médicales ont été utilisés pour réduire le risque encouru par des erreurs d'entrée de données par les différents intervenants.

où les coûts sont supérieurs à la moyenne, et une variable qui indique si les coûts ont une tendance positive ou négative.

Les résultats de Bertsimas *et al.* montrent que les arbres de classification fonctionnent mieux pour les membres les moins coûteux, tandis que les algorithmes de classification fonctionnent mieux pour les membres qui sont les plus coûteux. La méthode de base à laquelle ils ont comparé leurs modèles est la somme des coûts des soins de santé dans la deuxième année, soit l'année précédant la prédiction. Leurs modèles améliorent la prédiction pour les membres dans toutes les catégories de coûts, mais plus particulièrement pour ceux qui sont dans la catégorie supérieure.

Banjari *et al.* (2015) ont également utilisé des algorithmes de classification, mais leur objectif était de déterminer le résultat de l'accouchement au début de la grossesse. Ils ont recours à ces algorithmes dans le but de classifier les patientes en différents groupes et déterminer les patientes qui sont à risque de complications durant l'accouchement. Les auteurs démontrent que les femmes enceintes âgées avec un indice de masse corporelle plus élevé avant la grossesse ont une incidence significativement plus élevée d'accouchement d'un enfant d'un poids important pour l'âge gestationnel. Cependant, ces patientes seraient celles qui prennent le moins de poids pendant la grossesse (Banjari *et al.*, 2015). Leurs bébés sont également plus longs, et ces femmes ont une probabilité significativement plus élevée de complications pendant la grossesse et une plus grande probabilité d'accouchement provoqué ou césarien. Banjari *et al.* démontrent aussi que les méthodes d'analyses par regroupement peuvent classer de manière appropriée les femmes au début de la grossesse afin de prédire le résultat de l'accouchement.

CHAPITRE II

MÉTHODOLOGIE

2.1 Méthodes d'apprentissage automatique

Plusieurs méthodes empiriques ont été utilisées pour faire la prédiction d'une variable d'intérêt avant que les méthodes d'apprentissage automatique soient utilisées par les économistes. L'utilisation des méthodes d'apprentissage automatique est motivée essentiellement par trois raisons principales. Premièrement, la disponibilité d'ordinateurs ayant une puissance de calcul suffisamment grande pour ce type de modélisation a permis une certaine démocratisation de l'accès à ces méthodes. Deuxièmement, l'importance de la numérisation des transactions et le développement des technologies de l'information ont causé une augmentation considérable de la taille des bases de données, nécessitant l'utilisation de méthodes permettant une certaine sélection des variables explicatives[8]. Troisièmement, le développement de communautés entourant ces méthodes ont permis une avancée considérable dans le partage d'information et dans la disponibilité de logiciels libres permettant de réaliser ce type de modélisation.

Les méthodes économétriques régulières ne sont pas optimisées pour faire de la prévision puisqu'elles se concentrent essentiellement sur l'absence de biais dans l'obtention d'estimateurs servant à quantifier la relation entre des variables d'in-

8. Varian (2014) discute de cet aspect plus en détail.

térêt (Kleinberg *et al.*, 2015). Dans le cas d'un estimateur par moindres carrés ordinaire, il peut être démontré qu'il s'agit du meilleur estimateur linéaire non biaisé à variance minimale lorsque l'hypothèse d'homoscédasticité est respectée[9]. Cependant, la relation entre la variable dépendante et les variables explicatives étant inconnue au moment de l'estimation, des méthodes permettant des formes fonctionnelles flexibles, souvent non linéaires, sont nécessaires afin d'obtenir de bonnes prédictions.

L'objectif du présent mémoire étant un exercice de prédiction dans un contexte de données massives, des méthodes d'apprentissage automatique et des méthodes de ré-échantillonnage seront utilisées. Les méthodes d'apprentissage utilisées pour la prédiction des complications sont le modèle logit avec pénalité de type *elastic net*, la méthode à vecteurs de support linéaire, la méthode de forêt aléatoire (*random forest*), la méthode *boosting* par gradient et la méthode *boosting* avec l'algorithme AdaBoost[10]. Les meilleurs modèles estimés à partir de chaque méthode seront ensuite combinés avec la méthode d'ensemble d'apprentissage (*ensemble learning*) afin de construire un méta estimateur[11].

Plusieurs variables explicatives seront créées à partir des déterminants des complications identifiés dans la littérature. De plus, similairement à Bertsimas *et al.* (2008), les coûts mensuels des soins de santé de chaque patiente seront estimés afin de créer un ensemble de variables explicatives en lien avec l'évolution des coûts des soins de santé durant la période d'observation. La même méthodologie sera appliquée pour le nombre d'hospitalisations et le nombre de visites dans un département de gynécologie ou d'obstétrique. Les variables de coûts seront com-

9. Le théorème de Gauss-Markov démontre que l'estimateur des moindres carrés ordinaire est l'estimateur sans biais à variance minimale lorsque l'hypothèse d'homoscédasticité est effectuée.

10. La méthode à vecteurs de support est également appelée machine à vecteurs de support (*support vector machine*).

11. Cette méthode consiste à construire un seul modèle à partir de plusieurs modèles individuels afin de faire une prédiction.

binées à un ensemble de variables médicales dans le but de faire la prédiction des diagnostics de complications.

Initialement, une prédiction de la variable dépendante binaire sera effectuée. Ensuite, la moyenne des coûts des complications conditionnelle à la prédiction sera estimée afin de prédire les coûts. Donc, pour les patientes qui ont des complications, la moyenne conditionnelle des coûts sachant qu'un diagnostic de complications a été prédit sera utilisée. La même méthode sera appliquée pour les patientes qui n'ont pas de complications. Étant donné que l'accouchement représente une hausse de coûts spontanée et que les coûts bruts sont disponibles pour les actes médicaux en lien avec les complications, il est plus facile de prédire les diagnostics à l'aide de l'historique médical des patientes et ensuite utiliser cette information pour prédire les coûts que de tenter de prédire les coûts des soins médicaux durant l'accouchement directement.

2.1.1 Définition des méthodes et de l'apprentissage

Les méthodes d'apprentissage automatique serviront à faire une prédiction dans un contexte de classification avec une variable dépendante binaire[12]. Les modèles seront entrainés à partir d'un échantillon d'apprentissage avec 80 % des observations totales. Puisque des méthodes de ré-échantillonnage synthétique seront utilisées, aucune procédure de validation croisée ne sera utilisée afin de calibrer les modèles durant l'estimation[13]. Les 20 % restants constitueront l'échantillon test, qui sera mis de côté pendant l'estimation et la calibration des modèles. L'échantillon test servira seulement à la fin de la procédure d'estimation dans le

12. L'ensemble des méthodes d'apprentissage automatique utilisées proviennent des livres *An introduction to statistical learning : with applications in R* de James *et al.* (2013) et *The elements of statistical learning : data mining, inference, and prediction* de Trevor *et al.* (2009).

13. Les problèmes causés par la combinaison entre ces méthodes et la validation croisée sont illustrés dans la section 2.2.

but de mesurer la performance des modèles lorsqu'ils rencontrent de nouvelles observations. Cet échantillon sera également utilisé pour faire la prédiction des coûts des complications.

La variable dépendante est une variable binaire qui est égale à 1 si une patiente a eu une complication durant la grossesse et 0 sinon :

$$y_i = \begin{cases} 1 & \text{avec probabilité } p \\ 0 & \text{avec probabilité } 1-p. \end{cases} \tag{2.1}$$

Dans un problème de classification, la performance des modèles en apprentissage automatique est typiquement évaluée en utilisant la matrice de confusion. Le tableau 2.1 présente un exemple de matrice de confusion lorsque la variable dépendante est binaire. Les éléments sur la diagonale de la matrice sont le nombre d'observations pour lesquelles la classe de la variable dépendante a été correctement prédite, tandis que les observations hors diagonale représentent le nombre d'observations pour lesquelles une erreur de prédiction a été commise[14].

Dans la matrice de confusion, le nombre de vrais négatifs (VN) représente les patientes n'ayant pas de diagnostic de complications correctement identifiées, le nombre de faux positifs (FP) représente les patientes n'ayant pas de diagnostic de complications incorrectement identifiées, le nombre de faux négatifs (FN) représente les patientes ayant un diagnostic de complications incorrectement identifiées et le nombre de vrais positifs (VP) représente les patientes ayant un diagnostic de complications correctement identifiées. Il y a donc deux types d'erreurs commises. Les faux positifs, qui sont équivalents à commettre une erreur de type I, et les faux négatifs, qui sont équivalent à commettre une erreur de type II. Le modèle

14. Dans le but d'assurer une uniformité dans la nomenclature et d'être conforme avec la littérature, le terme classe sera utilisé afin de désigner les catégories de la variable dépendante.

Tableau 2.1: Exemple de matrice de confusion

Valeur observée		Valeur prédite $\hat{y}_i = 0$	Valeur prédite $\hat{y}_i = 1$
	$y_i = 0$	Vrai négatif (VN)	Faux positif (FP)
	$y_i = 1$	Faux négatif (FN)	Vrai positif (VP)

peut prévoir un diagnostic de complication alors qu'il n'y en a pas et il peut ne pas prévoir un diagnostic de complication alors qu'il y en a un.

Ces mesures permettent d'estimer la sensibilité et la spécificité du modèle, qui sont respectivement le pourcentage de vrais diagnostics de complications qui ont été correctement identifiés ($VP/(VP+FN)$) et le pourcentage de diagnostics de complications négatifs qui sont correctement identifiés ($VN/(VN+FP)$)[15]. En utilisant l'information contenue dans la matrice de confusion, d'autres mesures de performance sont généralement utilisées pour comparer la performance des modèles. Parmi celles-ci, on retrouve la précision, qui est définie telle que :

$$\text{précision} = \frac{VP+VN}{VP+FP+VN+FN}. \tag{2.2}$$

La précision représente le pourcentage des observations pour lesquelles la variable dépendante a été correctement prédite. Elle permet également de définir l'erreur moyenne de classification, qui est égale à 1 − précision.

15. Plusieurs termes sont utilisés dans la littérature afin de définir la sensibilité, soit le taux de rappel (*recall*), le taux de vrais positifs, et la probabilité de détection.

Une manière de représenter graphiquement les deux types d'erreurs de classification est à l'aide de la courbe ROC (*Receiver Operating Characteristics*). La figure 2.1 présente un exemple de représentation de la courbe ROC. Sur cette courbe, l'axe vertical représente la sensibilité tandis que l'axe horizontal représente 1 – spécificité. La performance globale du modèle pour tous les niveaux de seuil de sensibilité et de spécificité est donnée par l'aire sous la courbe ROC, soit l'AUC (*Area Under the Curve*). Plus le modèle sera performant, et plus la courbe ROC aura une forme coudée, ce qui implique des valeurs de sensibilité et de spécificité élevées pour tous les niveaux de seuil, avec une valeur pour l'AUC qui se rapprochera de 1. La ligne pointillée correspond à une valeur pour l'AUC de 0,5, qui est la plus faible valeur possible.

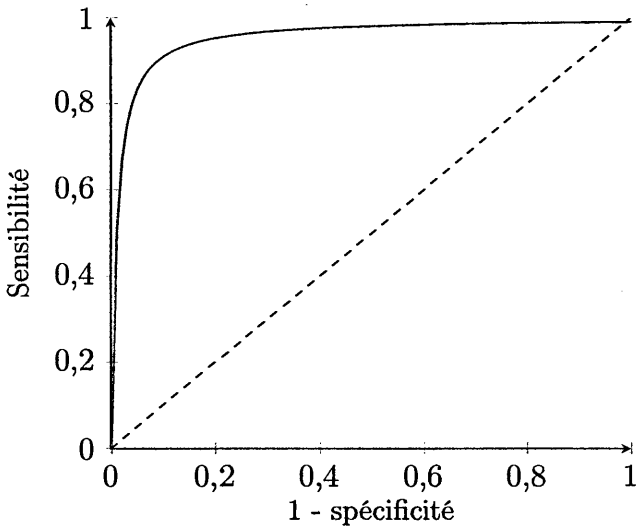


Figure 2.1: Exemple de représentation graphique de la courbe ROC

Un modèle avec une valeur de 0,5 pour l'AUC est un modèle qui prédit aléatoirement les classes de la variable dépendante. Lorsque les classes de la variable dépendante ne sont pas équilibrées et que les coûts des erreurs de prédictions sont asymétriques, utiliser comme mesure la précision ou l'erreur moyenne de classi-

fication pour déterminer le meilleur modèle peut faire en sorte que les meilleurs modèles estimés prédisent la classe majoritaire à l'ensemble de l'échantillon. Dans un tel contexte, il est préférable d'utiliser d'autres mesures afin de déterminer les meilleurs modèles (Chawla *et al.*, 2002).

Puisque la variable dépendante est une variable binaire indiquant la présence de complications durant l'accouchement, les erreurs de prédictions sont asymétriques. L'hypothèse derrière cette asymétrie est qu'il est plus coûteux de ne pas prévoir de diagnostics de complications alors qu'il y en a un plutôt que de prévoir un diagnostic de complications alors qu'il n'y en a pas. Il est donc nécessaire de s'intéresser aux différents types d'erreurs commises par les modèles estimés et d'utiliser plusieurs mesures pour évaluer la performance de ces derniers. L'AUC sera la mesure utilisée pour déterminer les meilleurs modèles durant l'estimation. La sensibilité, la spécificité, la précision, l'erreur de classification et l'AUC seront les mesures utilisées afin de comparer les meilleurs modèles estimés.

2.1.2 Méthode de régression avec régularisation

La méthode de régression avec régularisation utilisée sera le modèle logit avec une pénalité de type *elastic net* (Tibshirani, 1996 ; Friedman *et al.*, 2010 ; Zou et Hastie, 2005) [16]. Cette méthode estime un modèle logit en ajoutant une pénalité à la fonction de vraisemblance de sorte à pénaliser l'ajout de variables explicatives dans le modèle. En ajoutant la pénalité, ceci permet de faire un choix quant aux variables explicatives qui devraient être incluses dans le modèle afin d'améliorer la prédiction. L'inclusion de l'ensemble des variables explicatives risque de provoquer un problème de suridentification (*overfitting*) au moment de calibrer les modèles

16. Les détails des méthodes de régularisation lasso et Ridge proviennent des chapitres 3 et 4 du livre de Trevor *et al.* (2009) et du chapitre 6 du livre de James *et al.* (2013). Les détails du modèle logit proviennent du livre *Microeconometrics : methods and applications* de Cameron et Trivedi (2005).

en plus de réduire la performance hors échantillon, soit lorsque la performance des modèles est évaluée avec l'échantillon test. L'ajout de la pénalité permet donc de réduire la variance des prédictions tout en ayant un modèle interprétable à partir d'un sous-ensemble des variables disponibles pour construire le modèle.

Pour le modèle logit, la probabilité d'observer un diagnostic positif de complications durant la grossesse peut être définie comme étant :

$$Pr(y = 1|X) = \Lambda(X\beta) = \frac{\exp(X\beta)}{1 + \exp(X\beta)}. \tag{2.3}$$

Ensuite, le modèle logit pénalisé est estimé en maximisant la fonction de vraisemblance suivante :

$$\max_{\beta} \sum_{i=1}^{N} y_i log(\Lambda(X_i\beta)) + (1 - y_i) log(\Lambda(X_i\beta)) - \lambda \left(\alpha ||\beta||_1 + \frac{1-\alpha}{2} ||\beta||_2^2 \right). \tag{2.4}$$

La pénalité est représentée par le deuxième terme de l'équation. Elle sélectionne les variables explicatives en écrasant la valeur des coefficients des variables qui n'améliorent pas le modèle à 0 ou a des valeurs très près de 0. Le premier terme dans la parenthèse $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ sert à pénaliser la somme des valeurs absolues des coefficients, tandis que le deuxième terme $||\beta||_2^2 = \sum_{j=1}^{p} \beta_j^2$ sert à pénaliser la norme des coefficients, ce qui permet de conserver un certain effet de groupe des variables explicatives dans le modèle[17]. Le terme $||\beta||_1$ écrase complètement les coefficients des variables explicatives à 0, alors que le terme $||\beta||_2^2$ réduit les valeurs des coefficients simultanément sans les égaliser complètement à 0. Ce type de pénalité est appelé *elastic net* puisqu'il combine les pénalités du modèle lasso et du modèle Ridge.

17. L'effet de groupe survient lorsqu'un groupe de variables explicatives étant corrélées entre-elles tendent à être ajoutées ou retirées du modèle simultanément.

Les termes λ et α sont des paramètres dont les valeurs devront être déterminées avant l'estimation du modèle. Le paramètre λ détermine le poids de la contrainte lors de la maximisation de la vraisemblance, tandis que le paramètre α affecte la distribution de la pénalité entre $||\beta||_1$ et $||\beta||_2^2$, avec $\alpha \in [0, 1]$. Une valeur de λ nulle équivaudrait à estimer un modèle logit avec l'ensemble des variables explicatives. Dans les cas particuliers où $\alpha = 1$ et $\alpha = 0$, on retrouve respectivement le modèle lasso et le modèle Ridge. Le terme $||\beta||_1$ fait en sorte que le modèle tend vers une solution clairsemée lorsque λ est suffisamment grand, tandis que le terme $||\beta||_2^2$ permet une meilleure stabilité numérique et converge plus rapidement que le modèle lasso.

Les valeurs des paramètres λ et α seront déterminées en effectuant une procédure de recherche par grille (*grid search*). Cette procédure consiste à définir des vecteurs à partir des valeurs possibles pour chacun des paramètres et estimer plusieurs modèles, où chaque modèle utilise des valeurs différentes. Les modèles estimés sont ensuite ordonnés à l'aide d'une mesure de performance afin de déterminer le meilleur modèle, soit celui qui maximise ou qui minimise la mesure utilisée, et les valeurs de λ et α pour ce modèle sont les valeurs qui sont considérées optimales. Cette procédure sera utilisée pour déterminer les valeurs optimales des paramètres de tous les différents modèles en apprentissage automatique estimés.

2.1.3 Méthode à vecteurs de support linéaire

La méthode à vecteurs de support linéaire (SVC) prédit la classe d'une observation en définissant un hyperplan et en déterminant de quel côté cette dernière se situe par rapport à celui-ci [18]. L'hyperplan est choisi afin de discriminer le mieux possible les classes des observations dans l'échantillon d'entrainement, tout en permettant

18. Les détails de la méthode à vecteurs de support proviennent du chapitre 9 du livre de James *et al.* (2013) et du chapitre 12 du livre de Trevor *et al.* (2009).

au modèle de commettre des erreurs lorsqu'il effectue une prédiction. L'hyperplan est déterminé en solutionnant le problème d'optimisation suivant :

$$\max_{\beta_0,\beta_1,...,\beta_p,\varepsilon_1,...,\varepsilon_n,M} \quad M \tag{2.5}$$

$$\text{sujet à} \quad \sum_{j=1}^{p} \beta_j^2 = 1 \tag{2.6}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}) \geq M(1 - \varepsilon_i) \tag{2.7}$$

$$\varepsilon_i \geq 0 \tag{2.8}$$

$$\sum_{i=1}^{N} \varepsilon_i \leq C. \tag{2.9}$$

La variable C est un paramètre qui sert à régler le modèle et M représente la largeur de la marge. Puisque M est la fonction objectif du problème de maximisation, l'objectif est tel que M prenne la valeur la plus grande possible. Les variables $\varepsilon_i, ..., \varepsilon_n$ sont des variables de jeu qui permettent aux observations d'être du mauvais côté de la marge ou de l'hyperplan. La figure 2.2 présente un exemple graphique de vecteur de support linéaire avec quinze observations, où les points noirs et gris représentent les classes de la variable dépendante.

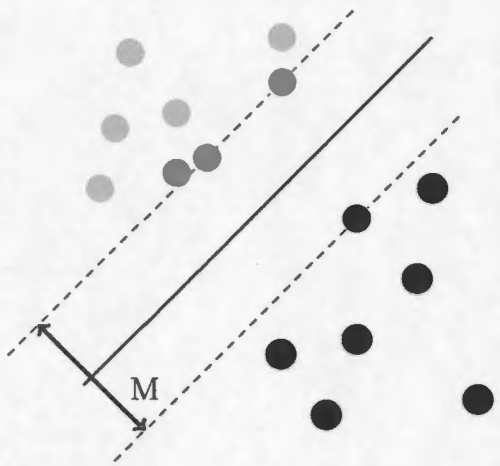


Figure 2.2: Exemple de vecteur de support linéaire avec $N = 15$

Une fois que le problème est solutionné, la classe d'une observation x^* est prédite en déterminant le côté de l'hyperplan où cette variable se situe, soit en prenant le signe de $f(x^*) = \beta_0 + \beta_1 x_1^* + ... + \beta_p x_p^*$. La variable de jeu ε_i indique l'endroit où se situe la $i^{ème}$ observation par rapport à la position de l'hyperplan et de la marge. L'observation se situe du bon côté de la marge lorsque $\varepsilon_i = 0$, du mauvais côté de la marge lorsque $\varepsilon_i > 0$ et du mauvais côté de l'hyperplan lorsque $\varepsilon_i > 1$. En bornant la somme des variables ε_i, le paramètre C détermine le nombre et la sévérité des violations par rapport à la marge et à l'hyperplan qui est toléré durant l'estimation. Lorsque $C = 0$, aucune violation n'est permise et $\varepsilon_1 = ... = \varepsilon_n = 0$, ce qui équivaut à la marge maximale possible[19]. Pour des valeurs de $C > 0$, seulement C observations peuvent être du mauvais côté de l'hyperplan puisque si une observation est du mauvais côté de l'hyperplan, la variable $\varepsilon_i > 1$ et la dernière contrainte du problème de maximisation requiert que $\sum_{i=1}^{N} \varepsilon_i \leq C$. Ainsi, lorsque C augmente, la tolérance pour les erreurs commises augmente et la marge M permise s'élargit. Le paramètre C contrôle donc l'arbitrage entre le biais et la

19. Une telle marge existe seulement si les classes de la variable dépendante sont complètement séparables (James *et al.*, 2013).

variance de cette méthode. Plus les valeurs de C seront petites, plus la marge sera petite et le modèle aura un faible biais et une grande variance. À l'opposé, une grande valeur de C augmente le biais du modèle et diminue la variance.

La solution du problème de maximisation ci-dessus implique de faire un produit scalaire des observations dans plusieurs dimensions (James *et al.*, 2013). En pratique, ce produit scalaire est remplacé par une fonction de noyau (*kernel function*) linéaire ou non linéaire afin de formaliser le problème de maximisation. L'estimation avec cette méthode est ensuite réalisée en solutionnant le problème de maximisation avec la fonction de noyau qui a été spécifiée par l'utilisateur. Ici, une fonction de noyau linéaire sera utilisée. Le paramètre C doit être spécifié avant l'estimation de cette méthode. Un vecteur de valeurs pour le paramètre C sera défini et une procédure de recherche par grille sera utilisée dans le but de déterminer la valeur qui permet d'obtenir le meilleur modèle.

2.1.4 Méthodes en arbre

Les méthodes en arbre consistent a estimer un ou plusieurs arbres de décision pour représenter les données, pour ensuite faire une prédiction basée sur l'endroit où se trouve chaque observation dans le modèle final [20]. En plus de permettre l'obtention de résultats facilement interprétables, ces méthodes sont utiles puisqu'elles permettent de considérer des effets d'interactions entre les variables indépendantes, ce qui n'est pas le cas avec les méthodes de régression à moins que les interactions soient incluses dans la spécification du modèle.

Un arbre de décision estime un modèle en appliquant une série de divisions binaires et de règles de décisions, à chaque fois en utilisant une variable pour former

20. Les détails des méthodes en arbre utilisées proviennent du chapitre 8 du livre de James *et al.* (2013) et des chapitres 8, 10 et 15 du livre de Trevor *et al.* (2009).

un nœud interne m qui entrainera une séparation en deux branches distinctes. Cette séparation formera une région R_m avec un nombre d'observations N_m qui constituent un sous-ensemble de l'échantillon d'apprentissage. Ce processus est répété de manière récursive jusqu'à la fin de l'estimation pour déterminer la forme de l'arbre de décision. Le nombre de niveaux constituant l'arbre de décision détermine la profondeur de celui-ci. À chaque nœud, le modèle tente d'assigner une observation dans une région donnée à la classe dominante dans cette région. La proportion d'observations appartenant à la classe k à un nœud m est définie comme étant :

$$p_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k), \quad k \in \{0, 1\}. \tag{2.10}$$

La mesure utilisée pour déterminer la qualité d'une séparation est l'indice de Gini, soit :

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk}). \tag{2.11}$$

La matrice X_m représente l'échantillon d'entrainement à un nœud m. L'indice de Gini est une mesure de la variance pour chacune des classes k. Si les valeurs de $\hat{p}_{mk}$ sont près de 0 ou de 1, l'indice de Gini prendra des valeurs faibles, ce qui indique que les observations N_m appartiennent essentiellement à la même classe. Cet indice est donc interprété comme un indicateur de la pureté d'un nœud. Le seuil de la $j^{ième}$ variable utilisée pour déterminer la règle de décision pour la division binaire est celui qui permet au modèle de minimiser l'indice de Gini. L'estimation se termine lorsqu'un critère d'arrêt est satisfait, soit par exemple lorsque le nombre d'observations restant est inférieur à un certain seuil. La prédiction finale sera basée sur la classe qui est la plus présente dans le sous-groupe du nœud terminal. La figure 2.3 ci-dessous illustre un exemple de représentation pour un arbre de décision servant à prédire les complications.

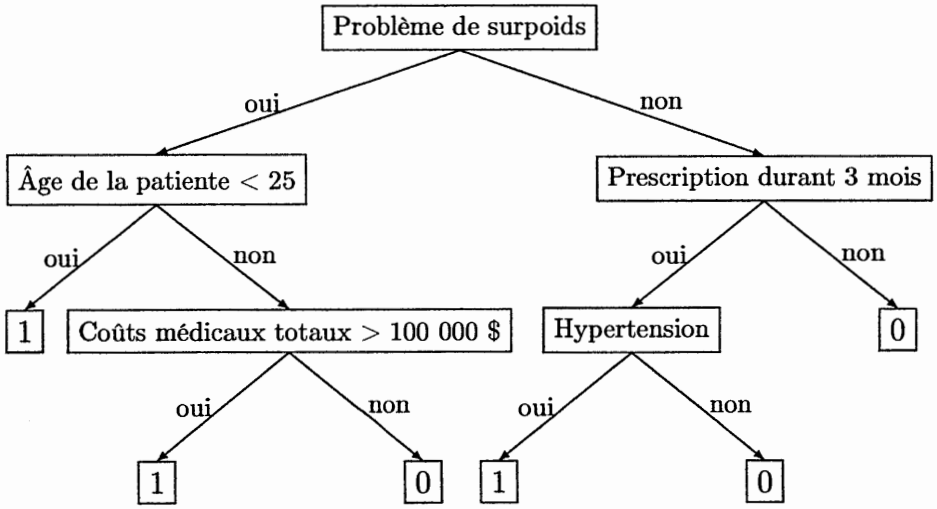


Figure 2.3: Exemple de représentation d'un arbre de décision avec trois niveaux, cinq nœuds internes et six nœuds terminaux

Le niveau de profondeur d'un arbre de décision peut être interprété comme étant le niveau d'interactions permises entre les variables indépendantes. Un arbre avec une faible profondeur risque de mal représenter les données puisqu'il ne permet pas de modéliser suffisamment la complexité de la relation entre les diagnostics de complications et les variables explicatives. Un tel modèle fait légèrement mieux que de prédire aléatoirement les classes de la variable dépendante.

Cependant, un arbre de décision sans profondeur maximale peut être estimé jusqu'à ce qu'il représente l'ensemble des caractéristiques de l'échantillon, ce qui causera l'estimation d'un modèle avec une variance élevée et un biais faible. Un tel modèle mène à une structure excessivement complexe, en plus d'être difficilement interprétable et de causer un problème de suridentification. Pour remédier à ces problèmes, plusieurs méthodes itératives ont été développées afin de réduire la variance du modèle final[21]. Ces méthodes sont des méthodes dites d'ensemble d'apprentissage puisqu'elles font des prédictions en se basant sur plusieurs modèles

21. Ces méthodes itératives seront décrites plus tard dans cette section.

individuels. En combinant ces modèles, les méthodes d'ensemble d'apprentissage tendent à être plus flexibles et moins sensibles aux données, ce qui permet de réduire les problèmes de suridentification rencontrés lorsqu'un seul arbre de décision est estimé pour faire la prédiction.

2.1.5 Méthode de forêt aléatoire

La méthode de forêt aléatoire fait partie des méthodes itératives permettant de réduire le problème de suridentification d'un arbre de décision. Cette méthode consiste à estimer un nombre B d'arbres de décision partiellement indépendants de manière itérative dans le but de réduire la variance du modèle final. Pour chacun des arbres de décision, le modèle est estimé avec un échantillon *bootstrap* déterminé à partir des données initiales et avec un sous-ensemble m des variables explicatives. Les variables m pour chacun des arbres estimés sont déterminées aléatoirement pour éviter que les modèles soient toujours estimés avec les mêmes variables, ce qui a pour effet de réduire la corrélation entre les arbres de décision qui constituent le modèle final.

Pour chacun des arbres de décision estimés, une prédiction de la variable dépendante est effectuée. Une fois que tous les modèles sont estimés, les arbres de décision sont combinés afin de constituer un modèle final dans le but de faire une prédiction de la variable dépendante. Pour effectuer la prédiction, le modèle final estime la moyenne de la prédiction de l'ensemble B de modèles estimés, et la classe qui est prédite pour chaque observation individuellement est celle qui aura été prédite le plus souvent lors de l'estimation.

En utilisant le théorème central limite, Breiman (2001) a démontré que, puisque la méthode de forêt aléatoire converge lorsque le nombre d'arbres B augmente, il n'y a pas de problème de suridentification avec cette méthode lorsque le nombre

d'arbres est suffisamment grand. Ainsi, le nombre d'arbres B n'est pas un hyper paramètre puisqu'il n'y a pas de problème de suridentification avec le nombre d'arbres estimés.

Une valeur couramment utilisée pour le nombre de variables dans le sous-échantillon de variables explicatives est $m = \sqrt{p}$, où p représente le nombre de variables explicatives disponibles. Afin d'éviter le problème de suridentification pour chacun des arbres individuels estimés, une profondeur maximale sera spécifiée. La profondeur maximale d détermine la complexité de l'ensemble d'arbres estimés et la profondeur des interactions possibles entre les variables. Une procédure de recherche par grille sera effectuée pour la valeur des paramètres suivants, soit le nombre d'arbres estimés B, la profondeur maximale d et le nombre de variables explicatives m.

2.1.6 Méthode *boosting*

La méthode *boosting* est une autre méthode développée dans le but d'améliorer la performance d'un seul arbre de décision. Il s'agit d'une procédure itérative qui estime un arbre de décision de manière séquentielle. À chacune des itérations de la procédure, un nouvel arbre de décision est estimé en utilisant l'information du modèle estimé lors de l'itération précédente avec une version modifiée de l'échantillon d'apprentissage initial, sauf pour la première itération où aucune modification n'est apportée à l'échantillon. Cette méthode estime donc un nombre d'arbres de décision B pour former un comité de modèles $\hat{f}(x)^b$, avec $b = 1, ..., B$. Le paramètre B correspond alors au nombre total d'itérations. Une faible valeur pour la profondeur des arbres de décision est spécifiée afin d'avoir des modèles avec une faible variance et un grand biais à chacune des itérations. Ce comité est aussi appelé apprenants faibles (*weak learners*).

La procédure d'estimation est basée sur l'algorithme suivant :

1. Définition d'une fonction initiale $\hat{f}(x)$ tel que $\hat{f}(x) = 0$ et initialisation des résidus r à partir de la variable dépendante, où $r_i = y_i$. Des poids initiaux $w_i = 1/N$ sont déterminés pour chacune des observations.

2. Pour chaque arbre de décision $b = 1, ..., B$:

   (a) Un arbre de décision $\hat{f}(x)^b$ est estimé avec un nombre de séparations d et un nombre de nœuds terminaux $(d + 1)$.

   (b) La fonction initiale $\hat{f}(x)$ est actualisée en ajoutant une version réduite du nouveau modèle estimé, avec $\nu < 1$:

$$\hat{f}(x) = \hat{f}(x) + \nu \hat{f}^b(x). \tag{2.12}$$

   (c) Les valeurs des résidus r sont actualisées avec le nouveau modèle estimé précédemment, en augmentant les poids w_i des observations qui ont été mal classifiées :

$$r_i = r_i - \nu \hat{f}^b(x_i), \quad i = 1, ..., N \tag{2.13}$$

3. Estimation du modèle final en prenant la somme des modèles estimés à chacune des itérations :

$$\hat{f}(x) = \sum_{b=1}^{B} \nu \hat{f}^b(x). \tag{2.14}$$

Ainsi, à chaque itération un nouveau modèle est estimé en accordant un poids positif aux observations qui ont été mal classifiées lors de l'itération précédente. Ce modèle est ensuite ajouté à la fonction initiale et l'erreur de classification est actualisée[22]. Cette procédure permet donc une amélioration graduelle de la fonction initiale $\hat{f}(x)$. Le paramètre de rétrécissement (*shrinkage*) ν détermine la

22. La fonction de perte utilisée est spécifique au problème de prédiction. Différentes fonctions de pertes sont utilisées lorsque la variable dépendante est binaire ou catégorique et lorsqu'elle est continue.

vitesse à laquelle l'apprentissage s'effectue. La méthode *boosting* peut être utilisée avec d'autres méthodes d'apprentissage que la méthode de forêt aléatoire, à condition que cette méthode permette d'attribuer des poids w_i aux observations de l'échantillon d'apprentissage.

Ici, relativement à la méthode de forêts aléatoire, une plus faible valeur pour la profondeur des arbres de décision est spécifiée afin de réduire la probabilité d'avoir un problème de suridentification et d'assurer un arbitrage optimal entre le biais et la variance du modèle estimé. Deux algorithmes basés sur cette méthode seront utilisés, soit le *boosting* par gradient et le *boosting* avec l'algorithme AdaBoost.

Trois vecteurs de paramètres de la méthode *boosting* seront spécifiés, soit pour le nombre d'arbres estimés B, le paramètre de rétrécissement (*shrinkage*) ν et pour la profondeur maximale d. Le paramètre de rétrécissement ν prend généralement une valeur positive près de zéro (*e.g.* $\nu < 0,1$). Plus ν sera faible, plus l'apprentissage se produira lentement, ce qui peut nécessiter parfois un grand nombre d'arbres de décisions B estimés afin d'avoir une bonne performance. Une procédure de recherche par grille sera effectuée pour déterminer la valeur optimale de chacun des paramètres.

2.2 Méthodes de ré-échantillonnage

Puisque le nombre de patientes ayant un ou plusieurs diagnostics de complications risque d'être inférieur au nombre de patientes ayant des grossesses normales, les classes de la variable dépendante pourraient ne pas être équilibrées. Lorsque ce déséquilibre survient, la performance des modèles peut être négativement affectée, et l'effet négatif sur la performance des modèles augmente généralement avec l'importance de ce déséquilibre. Pour remédier à ce problème, plusieurs méthodes existent. Par exemple, il est possible de définir des poids de sorte à favoriser les

observations appartenant à la classe minoritaire lors de l'estimation des modèles, de sur-échantillonner des observations de la classe minoritaire en copiant des observations existantes dans l'échantillon d'apprentissage ou de sous-échantillonner des observations de la classe majoritaire en choisissant aléatoirement un sous-ensemble des observations appartenant à cette classe. Inspirées de ces méthodes, des méthodes de ré-échantillonnage synthétique seront appliquées sur l'échantillon d'apprentissage pour tenter de minimiser les effets négatifs du déséquilibre des classes lors de l'entrainement des modèles. La méthode de sous-échantillonnage aléatoire de la classe majoritaire sera également utilisée.

Les méthodes de ré-échantillonnage synthétique ont été développées dans le but de générer de nouvelles observations synthétiques à partir d'observations existantes aux fins d'avoir un certain équilibre des classes de la variable dépendante. L'idée derrière ces méthodes est qu'elles permettent d'augmenter la généralisation de l'échantillon en générant de nouvelles observations, plutôt que de simplement copier des observations existantes afin d'équilibrer les classes. De plus, lorsque les méthodes de sur-échantillonnage synthétique de la classe majoritaire sont combinées avec des méthodes de sous-échantillonnage de la classe majoritaire, la performance des modèles est généralement grandement améliorée relativement aux méthodes mentionnées précédemment (Chawla *et al.*, 2002).

Néanmoins, puisque les nouvelles observations générées par ces méthodes peuvent être similaires, elles doivent être utilisées seulement sur l'échantillon d'apprentissage pour éviter que des exemples synthétiques se retrouvent dans l'échantillon test. Autrement, la performance des modèles risque d'être surestimée puisque de nouvelles observations semblables seront générées à la fois dans l'échantillon d'apprentissage et dans l'échantillon test, en plus de potentiellement causer des problèmes de suridentification (Santos *et al.*, 2018). En ce sens, lorsque ces méthodes sont combinées avec une procédure de validation croisée lors de l'entrai-

nement des modèles, des nouvelles observations doivent être générées à chaque itération de la procédure pour éviter ces problèmes. L'utilisation des méthodes de ré-échantillonnage synthétique avec une technique de validation croisée peut donc s'avérer très gourmande en termes de puissance de calcul requise pour effectuer les estimations. Pour cette raison, aucune procédure de validation croisée ne sera effectuée pour l'entrainement des modèles.

Les meilleurs modèles estimés avec l'échantillon d'entrainement serviront de référence pour les meilleurs modèles estimés avec les méthodes de ré-échantillonnage. Quatre méthodes seront appliquées afin de modifier l'échantillon d'apprentissage initial. Premièrement, seulement la technique de sous-échantillonnage aléatoire de la classe majoritaire sera utilisée. Deuxièmement, la technique de sur-échantillonnage synthétique de la classe minoritaire (SMOTE) sera combinée avec la technique de sous-échantillonnage aléatoire de la classe majoritaire. Troisièmement, la technique SMOTE sera utilisée avec une technique de sous-échantillonnage de la classe majoritaire à l'aide de liens Tomek (*Tomek links*). Finalement, la technique SMOTE sera utilisée avec une technique de sous-échantillonnage de la classe majoritaire avec la méthode des plus proches voisins actualisés (ENN).

2.2.1 Technique de sur-échantillonnage synthétique de la classe minoritaire combinée à une technique de sous-échantillonnage aléatoire de la classe majoritaire

La méthode permettant de construire des modèles de classification avec la technique SMOTE a été introduite par Chawla *et al.* (2002). Afin de générer des observations synthétiques, cette méthode utilise la méthode des K plus proches voisins (K-NN). La technique SMOTE applique la méthode K-NN en considérant seulement les observations appartenant à la classe minoritaire. Étant donné un nombre positif de voisins K à considérer à partir d'une observation initiale

x_0, la méthode K-NN permet de construire un voisinage $\mathcal{N}_0$ en identifiant les K observations dans l'échantillon d'entrainement qui sont le plus près de x_0. Ensuite, la technique SMOTE crée une observation synthétique en choisissant au hasard un des K plus proches voisins et en prenant une combinaison linéaire le long du segment joignant ces deux observations. Cette combinaison linéaire représente la nouvelle observation synthétique générée à partir des deux observations existantes. Cette procédure est répétée de manière itérative jusqu'à ce que le nombre de nouvelles observations nécessaire soit rencontré [23].

La quantité désirée de sur-échantillonnage q et le nombre de voisins K à considérer sont des paramètres qui doivent être spécifiés avant l'utilisation de la technique SMOTE pour générer de nouvelles observations. Le paramètre q est un ratio représentant le nombre d'observations dans la classe minoritaire par rapport au nombre d'observations dans la classe majoritaire, où $q \in [\frac{N_{\text{minoritaire}}}{N_{\text{majoritaire}}}, 1]$. La quantité de sur-échantillonnage possible est donc fonction du nombre d'observations dans chacune des classes de la variable dépendante. La quantité maximale de sur-échantillonnage consiste à équilibrer complètement les classes, soit lorsque $q = 1$.

Dans leur article, Chawla *et al.* ont testé l'application de ces méthodes en estimant des modèles avec plusieurs bases de données différentes dans le but de les comparer aux méthodes de ré-échantillonnage avec remplacement. Ils ont trouvé que lorsque les modèles sont estimés en utilisant les échantillons contenant des observations synthétiques plutôt que l'échantillon initial, leur performance est généralement supérieure par rapport à la performance des modèles estimés avec l'échantillon d'apprentissage initial. En créant de nouvelles observations artificiellement à partir d'observations existantes, cette méthode permet aux modèles estimés de générer des régions de décisions plus grandes et moins spécifiques, ce

23. Les détails techniques ainsi que le pseudo-code de l'algorithme SMOTE sont disponibles dans l'article de Chawla *et al.* (2002).

qui cause une généralisation de la région de décision de la classe minoritaire, et améliore la performance des modèles pour la prédiction de la classe minoritaire (Chawla *et al.*, 2002).

La quantité de sur-échantillonnage q sera déterminée empiriquement avec les données en utilisant une procédure itérative similaire à la recherche par grille. Initialement, un vecteur de valeurs pour le paramètre q sera défini, et ensuite la méthode de forêt aléatoire sera utilisée pour l'estimation d'un modèle permettant de déterminer la valeur de q qui offre un arbitrage optimal entre la sensibilité, la spécificité et l'erreur de classification. Puisque Chawla *et al.* ont obtenu de meilleurs résultats en combinant la technique SMOTE avec une technique de sous-échantillonnage aléatoire de la classe majoritaire, ces deux méthodes seront combinées. Plusieurs combinaisons de quantité q de sur-échantillonnage et de sous-échantillonnage seront testées afin de déterminer la meilleure combinaison pour l'entrainement des modèles.

Puisque la sensibilité et l'erreur de classification représentent respectivement la probabilité de détection et la probabilité de commettre une erreur de classification, le degré de sur-échantillonnage sera choisi de sorte que la sensibilité soit supérieure à l'erreur de classification tout en ayant une valeur acceptable pour la spécificité. La valeur du paramètre pour le nombre de voisins K sera égale à cinq, soit la même valeur qui a été utilisée par Chawla *et al.*.

2.2.2 Technique de sur-échantillonnage synthétique de la classe minoritaire combinée à une technique de sous-échantillonnage avec la méthode des liens Tomek

Les méthodes de sur-échantillonnage et de sous-échantillonnage ont toutes les deux des désavantages. Les méthodes de sous-échantillonnage peuvent retirer de l'échantillon d'apprentissage des observations étant utiles pour l'entrainement des

modèles, alors que les méthodes de sur-échantillonnage peuvent augmenter la probabilité d'avoir un problème de suridentification (Batista *et al.*, 2003). Également, la plupart du temps, les grappes (*clusters*) des classes de la variable dépendante ne sont pas suffisamment bien définies, ce qui peut faire en sorte que des observations de la classe majoritaire envahissent l'espace appartenant à la classe minoritaire ou que des observations artificielles de la classe minoritaire soient créées dans l'espace de la classe majoritaire (Batista *et al.*, 2003). Ces méthodes sont donc susceptibles d'augmenter le bruit dans les données.

Afin de tenter de résoudre ces problèmes, Batista *et al.* (2003) proposent une méthode qui combine la technique SMOTE avec une technique de sous-échantillonnage de la classe majoritaire en identifiant les liens Tomek joignant deux observations de chacune des classes de la variable dépendante. La méthode qui permet d'identifier les liens Tomek a été introduite par Tomek (1976), et la méthode de sous-échantillonnage de la classe majoritaire en utilisant les liens Tomek a été développée par Kubat *et al.* (1997).

Un lien Tomek est défini comme suit : étant donné deux exemples i et j appartenant à deux classes différentes, avec $d(i,j)$ qui représente la distance entre ces deux observations, un couple (i,j) est considéré comme étant un lien Tomek s'il n'existe pas une observation k telle que $d(i,k) < d(i,j)$ ou $d(k,j) < d(i,j)$. Si deux observations forment un lien Tomek, cela signifie que ces deux exemples se situent à la limite de la région de décision ou qu'une de ces deux observations constitue une part du bruit dans les données (Batista *et al.*, 2003, Kubat *et al.*, 1997).

Dans leur article, Batista *et al.* (2003) débutent en appliquant la technique SMOTE afin d'équilibrer complètement les classes de la variable dépendante. Ensuite, pour chacun des liens Tomek identifiés, les deux observations formant le couple (i,j)

sont retirées de l'échantillon d'entrainement contenant les nouvelles observations synthétiques. Ici, la même méthode sera appliquée et les deux observations formant le lien Tomek seront toutes les deux retirées de l'échantillon d'apprentissage. Ainsi en supprimant ces observations, cela permet de réduire le bruit dans les données et de créer des grappes mieux définies pour chacune des classes (Batista *et al.*, 2003).

Afin de déterminer la quantité de sur-échantillonnage q nécessaire pour l'estimation des modèles, la procédure itérative utilisée sera la même que celle qui a été mentionnée dans la section précédente. Pour la méthode de sous-échantillonnage de la classe majoritaire, le seul paramètre qui sera spécifié est la classe de l'observation qui sera supprimée à chaque fois qu'un lien Tomek est identifié.

2.2.3 Technique de sur-échantillonnage synthétique de la classe minoritaire combinée à une technique de sous-échantillonnage avec la méthode des plus proches voisins actualisés

La justification derrière la combinaison de ces méthodes de ré-échantillonnage est similaire à la justification derrière la combinaison de la technique SMOTE avec la technique de sous-échantillonnage de la classe majoritaire en utilisant les liens Tomek. La méthode des plus proches voisins actualisés (ENN) a été développée par Wilson (1972). Cette méthode a été initialement utilisée comme technique de sous-échantillonnage des observations de la classe majoritaire par Batista *et al.* (2004) et Laurikkala (2001).

Dans leur article, Batista *et al.* utilisent la méthode ENN pour sous-échantillonner des observations une fois que des observations synthétiques ont été générées avec la technique SMOTE. Pour une observation donnée x_i, la méthode ENN débute en identifiant les trois plus proches voisins de cette observation pour ensuite faire une prédiction de la classe de x_i. La méthode utilise ensuite cette prédiction dans

le but de déterminer si cette observation doit être retirée de l'échantillon d'apprentissage. Si la prédiction contredit la classe de x_i, cette observation est retirée de l'échantillon d'apprentissage. Donc, à chaque fois qu'une observation est retirée de l'échantillon, la méthode procède avec une version modifiée de l'échantillon initial. La méthode arrête lorsque la procédure a été effectuée pour chacune des observations et que le nettoyage est terminé. La quantité de sur-échantillonnage q sera déterminée en utilisant la procédure itérative mentionnée précédemment.

Cette méthode peut retirer à la fois des exemples de la classe minoritaire et de la classe majoritaire. De plus, puisque cette méthode considère chaque observation individuellement et qu'elle retire une observation dès que sa classe est contredite, elle tend à être agressive lorsqu'elle retire des observations de l'échantillon d'apprentissage pour nettoyer les données, ce qui réduit considérablement le bruit créé lors de l'application de la technique SMOTE. Même lorsque la technique de sous-échantillonnage qui utilise les liens Tomek est utilisée de sorte qu'elle retire les deux observations formant un lien Tomek, la technique de sous-échantillonnage avec la méthode ENN tend à retirer plus d'observations dans l'échantillon d'apprentissage afin de nettoyer les données (Batista *et al.*, 2004).

2.3 Calculs

L'ensemble des calculs et estimations ont été réalisés avec la version 3.6 du logiciel libre Python. Les modèles d'apprentissage automatique ont été estimés avec la version 0.21.2 du *package* scikit-learn (Pedregosa *et al.*, 2011) et les méthodes de ré-échantillonnage ont été appliquées avec la version 0.0 du *package* imblearn (Lemaître *et al.*, 2017).

Les paramètres pour l'estimation des modèles d'apprentissage automatique et pour les méthodes de ré-échantillonnage ont été déterminés conformément aux procé-

dures décrites dans cette section. Le tableau ci-dessous résume les modules de chacun des *packages* utilisés pour l'obtention des résultats présentés dans cette recherche.

Tableau 2.2: Modules et *packages* utilisés pour les estimations et les méthodes de ré-échantillonnage

Module	Description
sklearn.model_selection	Séparation de l'échantillon en échantillon d'apprentissage et en échantillon test, procédure de recherche par grille pour les paramètres des modèles
sklearn.metrics	Mesures de performance (AUC et précision)
sklearn.preprocessing	Prétraitement des données avant l'estimation des modèles
sklearn.linear_model	Modèle logit avec pénalité de type *elastic net*
sklearn.svm	Méthode de vecteur de support linéaire
sklearn.ensemble	*Boosting*, forêt aléatoire, AdaBoost et ensemble d'apprentissage
sklearn.tree	Arbre de décision
imblearn.metrics	Mesures de performances pour les problèmes de classifications non balancés (sensibilité et spécificité)
imblearn.over_sampling	Technique de sur-échantillonnage SMOTE
imblearn.under_sampling	Technique de sous-échantillonnage aléatoire
imblearn.combine	Technique SMOTE avec la technique de sous-échantillonnage avec les liens Tomek et technique SMOTE avec la technique ENN

CHAPITRE III

DONNÉES

3.1 Description des données

Les données utilisées sont des données médicales provenant d'un État nord-américain. Elles contiennent les dossiers médicaux de 54 000 femmes ayant donné naissance entre les années 1998 à 2006 et elles incluent l'ensemble des services médicaux dispensés pour chaque patiente durant une période de 7 ans[24]. L'information est disponible pour une période de deux ans avant et cinq ans après l'accouchement pour lequel la prédiction sera effectuée.

Les dossiers médicaux comprennent l'information de la patiente, l'ensemble des actes médicaux posés par les différents acteurs du système de santé, les prescriptions, et les programmes d'assurance médicament pour chaque patiente tout au long de la période observée. Les coûts bruts des actes médicaux, des diagnostics et des prescriptions sont également disponibles. Aucune information socioéconomique n'est comprise dans les données. Les codes de diagnostics sont inscrits selon la codification de la Classification Internationale des Maladies (CIM-9). Ces codes de diagnostics sont ceux qui ont été utilisés afin de créer les variables relatives aux diagnostics médicaux.

---

24. Les bases de données ont été importées, nettoyées et appariées avec le logiciel libre R (version 3.4.3).

L'historique médical des deux années avant l'accouchement représente l'ensemble d'information qui sera supposé connu au moment d'effectuer la prédiction. Les vingt-trois mois avant le dernier mois précédant l'accouchement désignent la période d'observation, soit l'information qui sera utilisée afin de créer les variables explicatives. Puisque la prédiction des complications est effectuée un mois avant la date de l'accouchement, l'information du 24^{e} mois ne sera pas utilisée. La période de résultat sera constituée de l'information des trois mois suivant l'accouchement, incluant le jour de l'accouchement. La figure 3.1 ci-dessous représente la structure temporelle de l'information de l'historique médical de la patiente qui sera utilisée.

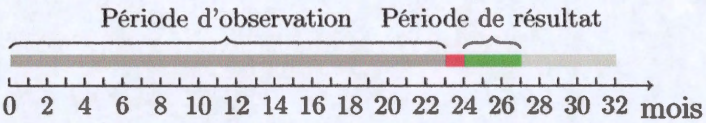


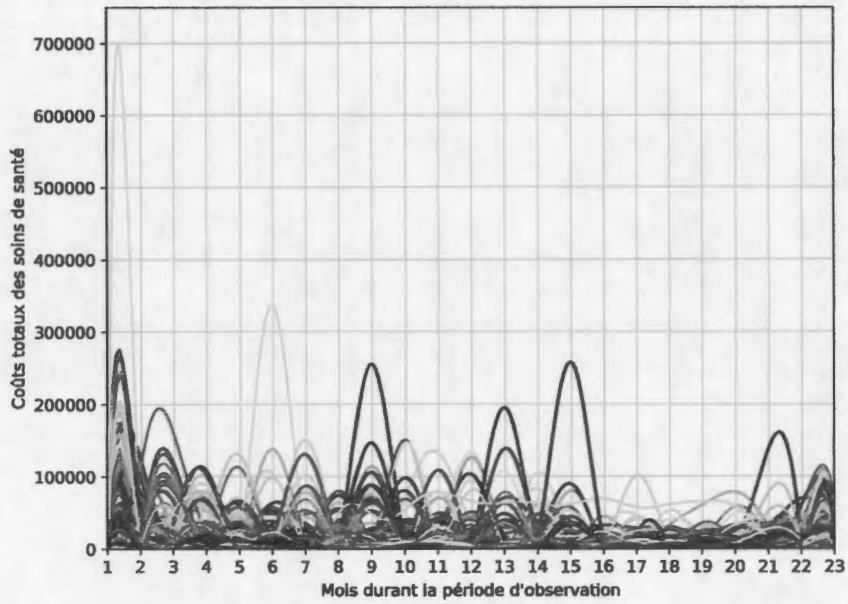
Figure 3.1: Représentation de l'historique médical d'une patiente

L'objectif derrière cette décomposition temporelle est de faire la prédiction des complications pour l'accouchement qui a été utilisé afin de déterminer l'historique médical et non pour des accouchements qui sont survenus plus tard dans la vie d'une patiente. Ceci permet de ne pas truquer les modèles en utilisant de l'information qui ne sera pas disponible au moment où la prédiction sera effectuée, en plus d'assurer une cohérence temporelle pour la prédiction. Les coûts qui seront prédits sont les coûts totaux des soins de santé durant la période de résultat.

3.2 Données sur les coûts des soins médicaux

Les données sur les coûts contiennent les coûts bruts des prescriptions et des actes médicaux posés par les professionnels de la santé pour chacune des patientes. En se basant sur la méthodologie employée par Bertsimas *et al.* (2008), les coûts totaux mensuels des soins de santé ont été estimés afin de représenter leur trajectoire du-

rant la période d'observation. La même méthode a été employée pour le nombre de visites dans un centre hospitalier et le nombre de visites en obstétrique. Les coûts totaux représentent la somme des coûts des prescriptions et des actes médicaux. Les patientes pour lesquelles aucune observation n'est disponible durant un mois spécifique ont un coût mensuel égal à zéro. La figure 3.2 représente la trajectoire des coûts totaux des soins de santé durant la période d'observation pour 1 000 patientes sélectionnées aléatoirement.



Source : Calculs de l'auteur à partir des données administratives.

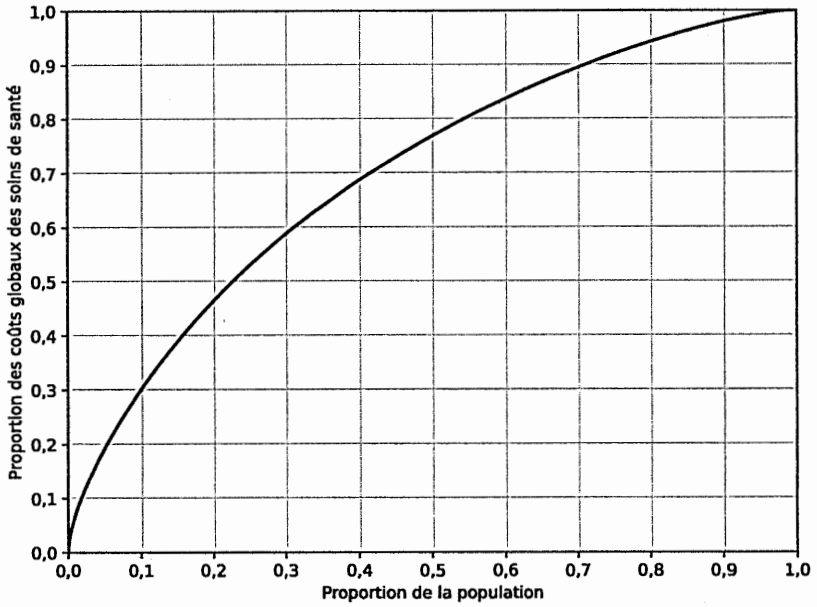
Figure 3.2: Coûts mensuels totaux des soins de santé durant la période d'observation

En regardant la figure ci-dessus, il est possible de remarquer que sauf pour quelques exceptions, les coûts mensuels dépassent peu la valeur de 20 000 $, avec une valeur maximale allant jusqu'à environ 700 000 $. Ce résultat suggère que la plupart des patientes ont des coûts relativement faibles, alors que certaines d'entre elles ont un

poids financier important sur le budget du système de santé. Certaines patientes ont des coûts mensuels faibles, mais non nuls, affichant une certaine persistance, ce qui peut également imposer des coûts importants au système de santé à moyen terme. À partir du 18^{e} mois, les coûts des patientes se rapprochent de zéro avec une certaine stabilité. Puisque cette période correspond à la période prénatale, soit six mois avant la date de l'accouchement, ce résultat suggère un changement dans les habitudes de santé des patientes.

Afin de quantifier la contribution des patientes ayant les coûts totaux les plus élevés durant la période d'observation dans les coûts totaux pour l'ensemble de l'échantillon, les coûts totaux des soins de santé cumulatifs ont été estimés. Les résultats sont présentés sur la figure 3.3 ci-dessous. L'axe vertical représente la proportion des coûts globaux et l'axe horizontal représente la proportion de la population[25]. La population a été ordonnée de manière décroissante selon les coûts totaux de chaque patiente.

25. La proportion des coûts globaux a été estimée en prenant la somme cumulative des coûts totaux des soins de santé.



Source : Calculs de l'auteur à partir des données administratives.

Figure 3.3: Coûts totaux des soins de santé cumulatifs durant la période d'observation

Similairement à ce qui a été observé par Bertsimas *et al.* (2008), la figure ci-dessus indique une non-linéarité dans la contribution de chaque patiente aux coûts globaux des soins de santé. Une pente linéaire implique que chaque patiente a une contribution marginale constante dans les coûts totaux. Plus la forme de la courbe est concave, plus les patientes ayant des coûts totaux élevés ont une part marginale importante dans les coûts totaux du système de santé. Les résultats de cette figure montrent que les patientes qui se retrouvent parmi les 10 % les plus coûteuses représentent environ 30 % des coûts globaux de soins de santé, alors que les 30 % les moins coûteuses représentent environ 10 % des coûts globaux.

Plusieurs variables définies à partir de la trajectoire des coûts mensuels ont été

créées afin de tenter de modéliser la relation entre l'évolution des coûts des soins de santé et la probabilité de complications durant la grossesse et l'accouchement[26]. Le tableau 3.1 présente des statistiques sommaires pour un sous-ensemble de ces variables pour la période d'observation. Cette période représente l'information qui sera utilisée pour créer les variables explicatives qui seront utilisées pour la prédiction. Les résultats sont regroupés pour chacune des classes de la variable dépendante et la troisième colonne contient la différence entre les deux groupes.

Parmi les 54 000 patientes dans l'échantillon, 11 951 ont au moins un diagnostic de complications durant la grossesse ou l'accouchement, ce qui représente 22,13 % des observations comparativement à 77,87 % pour les patientes qui ont une grossesse normale. Les classes de la variable dépendante ne sont donc pas équilibrées, ce qui risque de faire en sorte que les modèles estimés à partir de l'échantillon initial vont surestimer le nombre de patientes qui ont une grossesse normale. Par conséquent, les patientes ayant des complications font partie de la classe minoritaire tandis que les patientes qui ont une grossesse normale font partie de la classe majoritaire. Les méthodes de ré-échantillonnage seront donc utilisées afin de tenter d'améliorer la prédiction des diagnostics de complications par rapport aux prédictions obtenues à l'aide des modèles qui ont été entrainés avec l'échantillon d'apprentissage initial.

À partir des résultats exposés dans le tableau 3.1, il est possible d'observer que la différence entre les moyennes des patientes avec un diagnostic de complications et des patientes qui ont une grossesse normale est significative à un seuil de 1 % pour toutes les variables à l'exception du nombre de mois au-dessus de la moyenne des coûts pour les actes médicaux et pour les prescriptions. La contribution des coûts des prescriptions dans les coûts totaux est inférieure par rapport à celle des coûts des actes médicaux. En somme, ces résultats suggèrent que les patientes qui

26. La plupart des variables créées ont été basées sur les variables utilisées par Bertsimas *et al.* (2008).

ont des complications ont des coûts de soins de santé supérieurs relativement aux patientes qui ont des grossesses normales, et ce pour l'ensemble des variables de coûts considérées.

Tableau 3.1: Statistiques descriptives pour les variables de coûts durant la période d'observation

Variable :	Pas de diagnostic de complications	Diagnostic de complications	Différence entre les deux groupes
Coûts totaux des soins de santé	77 278,46 (362,81)	91 251,17 (957,53)	13 972,71*** (850,71)
Moyenne des coûts totaux des soins de santé	3 359,93 (15,77)	3 967,44 (41,63)	607,51*** (36,99)
Coûts des actes médicaux durant les 3 derniers mois	14 223,57 (64,22)	17 251,35 (156,15)	3 027,79*** (146,42)
Coûts des prescriptions durant les 3 derniers mois	1 245,05 (42,49)	2 061,63 (122,91)	816,57*** (103,18)
Moyenne des coûts des actes médicaux durant les 3 derniers mois	4 741,19 (21,41)	5 750,45 (52,05)	1 009,26*** (48,81)
Moyenne des coûts des prescriptions durant les trois derniers mois	415,02 (14,16)	687,21 (40,97)	272,19*** (34,39)
Nombre de mois au-dessus de la moyenne des coûts (actes médicaux)	7,18 (0,01)	7,16 (0,02)	-0,01 (0,02)
Nombre de mois au-dessus de la moyenne des coûts (prescriptions)	1,76 (0,02)	1,78 (0,03)	0,03 (0,03)
Coût mensuel maximal des actes médicaux	22 229,93 (125,97)	25 799,50 (315,44)	3 569,56*** (290,03)
Coût mensuel maximal des prescriptions	2 587,04 (35,94)	3 176,21 (102,61)	589,17*** (86,82)
Nombre d'observations	42 049	11 951	

Note : Les écarts-types sont rapportés entre parenthèses. Les étoiles représentent les niveaux de significativité d'un test de différence de moyenne entre les deux groupes, * valeur-p $< 0,1$, ** : valeur-p $< 0,05$, *** : valeur-p $< 0,01$.

Le tableau 3.2 présente des statistiques sommaires pour les mêmes variables de coûts durant la période de résultat, soit la période pour laquelle la prédiction sera effectuée. Pour chacune de ces variables, la différence de moyenne entre les patientes avec des complications et celles qui ont un accouchement normal est significative à un seuil de 1 %. En considérant seulement les coûts totaux le jour de l'accouchement, la différence entre les moyennes des coûts totaux entre les deux groupes est de 28 000,57 $, ce qui laisse présager que les complications le jour de l'accouchement ont un poids financier important sur le système de santé. De plus, les coûts totaux des patientes sans complications sont relativement constants à travers le temps par rapport aux patientes avec des complications. Par contre, ce résultat doit être interprété avec précaution, puisque l'évolution des coûts de soins de santé après l'accouchement risque d'être corrélée avec d'autres problèmes de santé n'étant pas nécessairement liés aux complications.

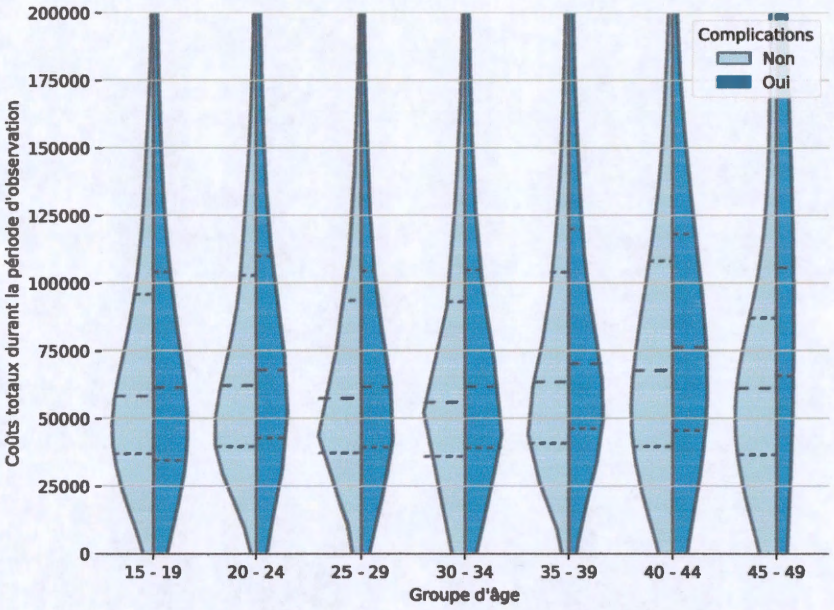
Tableau 3.2: Statistiques descriptives pour les variables de coûts durant la période de résultat

Variable :	Pas de diagnostic de complications	Diagnostic de complications	Différence entre les deux groupes
Coûts totaux des soins de santé le jour de l'accouchement	57 257,18 (105,85)	85 257,74 (350,84)	28 000,57*** (272,77)
Coûts totaux des soins de santé la 1ère semaine après l'accouchement	60 232,93 (110,62)	93 778,14 (379,71)	33 545,21*** (289,88)
Coûts totaux des soins de santé deux semaines après l'accouchement	60 883,44 (112,77)	95 109,95 (387,54)	34 226,50*** (295,68)
Coûts totaux des soins de santé un mois après l'accouchement	62 198,88 (117,29)	97 212,53 (401,53)	35 013,65*** (306,96)
Coûts totaux des soins de santé deux mois après l'accouchement	65 595,60 (127,44)	101 351,87 (421,00)	35 756,28*** (327,90)
Coûts totaux des soins de santé pour la période de résultat	67 979,60 (137,24)	103 972,34 (438,52)	35 992,75*** (347,74)
Nombre d'observations	42 049	11 951	

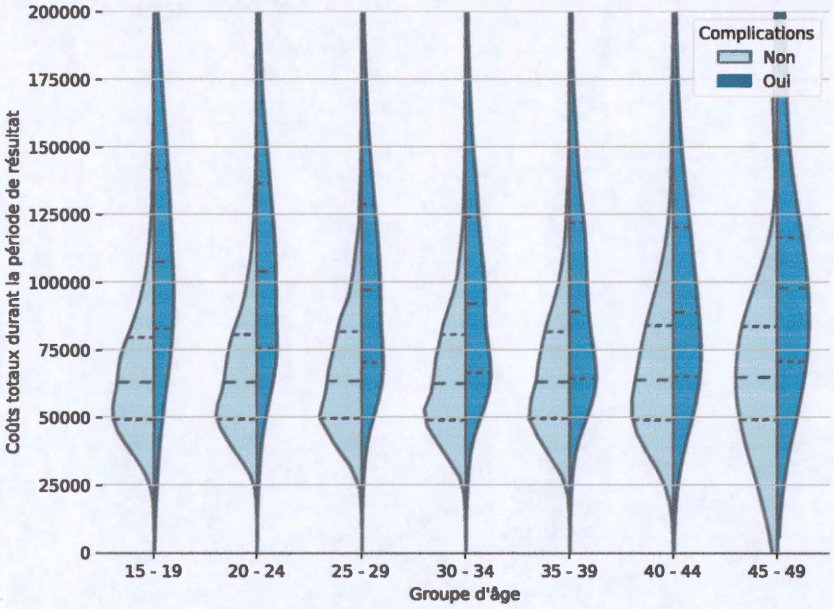
Note : Les écarts-types sont rapportés entre parenthèses. Les étoiles représentent les niveaux de significativité d'un test de différence de moyenne entre les deux groupes, * valeur-p $< 0,1$, ** : valeur-p $< 0,05$, *** : valeur-p $< 0,01$.

La figure 3.4 présente les distributions des coûts totaux des soins de santé pour la période d'observation et la période de résultat selon le groupe d'âge. Les lignes horizontales sur les distributions représentent le premier quartile, la médiane, et le troisième quartile. Pour la période d'observation, les patientes qui sont âgées de plus de 35 ans ont une médiane supérieure par rapport aux patientes plus jeunes. Cependant, pour la période de résultat, les patientes avec des complications qui ont plus de 35 ans ou moins de 25 ans sont celles qui ont une médiane des coûts totaux la plus élevée.

Pour la période d'observation et la période de résultat, les distributions de coûts pour les patientes qui ont des complications affiche également une médiane supérieure par rapport aux patientes qui ont des grossesses normales, ce qui confirme les résultats présentés dans les tableaux 3.1 et 3.2. Ces résultats vont dans le même sens que l'étude de Cavazos-Rehg *et al.* (2015), où les auteurs ont démontré que les femmes dans ces groupes d'âge ont un risque de complications plus grand. Ainsi, non seulement ces patientes auraient un risque plus élevé, mais ce serait également celles qui seraient les plus coûteuses lorsqu'elles ont des complications.



(a) Période d'observation



(b) Période de résultat

Source : Calculs de l'auteur à partir des données administratives.

Figure 3.4: Distributions des coûts totaux des soins de santé par groupe d'âge

Un élément important à considérer en regardant les résultats de la figure 3.4 est qu'aucune des patientes a un coût total nul le jour de l'accouchement. De plus, toutes les distributions se resserrent autour de la médiane. Ceci indique que la plupart des patientes avec un coût faible durant la période d'observation ont un coût plus élevé durant la période de résultat. Sans considérer les groupes d'âge, la valeur du premier percentile de la distribution des coûts totaux pour la période de résultat est de 2 403 $ et le cinquième percentile a une valeur de 36 580 $. Tandis que pour la période d'observation, la valeur du premier percentile est égale à 0 $ alors que le cinquième percentile a une valeur de 18 397 $. Effectuer une prédiction des coûts des soins de santé le jour de l'accouchement sans faire une première prédiction des diagnostics de complications constitue donc une tâche complexe puisque l'historique médical de la patiente durant la période d'observation ne reflète pas nécessairement les coûts durant la période de résultat.

3.3 Données sur les diagnostics, les prescriptions et les actes médicaux

En se basant sur les déterminants des complications identifiés dans la littérature, plusieurs variables explicatives ont été créées. Ces données ont également été utilisées afin d'identifier les diagnostics de complications pour créer la variable dépendante binaire. L'information durant la période d'observation a été utilisée pour créer les variables explicatives et l'information durant la période de résultat a été exploitée pour créer les variables dépendantes[27].

Plusieurs diagnostics ont été utilisés afin d'élaborer les variables explicatives. Parmi ceux-ci, on retrouve le diabète, l'hypertension, l'obésité, les infections transmises sexuellement, les maladies rénales, les troubles thyroïdes, les troubles du système immunitaire, les troubles dépressifs, les troubles anxieux, les grossesses

27. La liste complète des variables qui ont été créées pour l'estimation des méthodes est disponible dans le tableau A.1 présenté dans l'appendice A.

durant l'adolescence, les tumeurs et les troubles cardiaques. Des variables de diagnostic pour les fractures, les hémorragies, les mesures contraceptives et les mesures procréatives ont également été créées. De plus, en se basant sur les résultats de Kramer *et al.* (2000), les diagnostics liés à l'abus de substances (alcools, drogues ou autres) ont été identifiés afin de tenter d'approximer le statut socioéconomique de la mère. Une variable pour le régime d'assurance médicament lié à l'aide sociale a également été créée. Incluant les variables pour le groupe d'âge, il s'agit des seules variables à caractère socioéconomique qui sont disponibles dans les données. Pour les variables de diagnostic, une variable dichotomique et une variable de comptage ont été créées.

En plus des variables pour les diagnostics, des variables pour les actes médicaux ont aussi été générées. Premièrement, tous les actes médicaux ont été recensés dans le but de les regrouper en plusieurs catégories. Ensuite, une variable binaire a été créée afin d'indiquer si une patiente a eu un acte médical appartenant à une catégorie particulière, et ce pour chaque catégorie identifiée. En tout, neuf catégories ont été identifiées, soit les anesthésies, les actes en lien avec la peau et les tissus, le système respiratoire, le système cardiaque, le système urinaire et digestif, les actes en gynécologie, les actes en obstétrique, l'appareil glandulaire, et le système nerveux. En plus de ces variables, certaines variables ont été créées pour certains actes médicaux qui ont un lien direct avec la grossesse, par exemple les avortements, la procréation assistée et les anesthésies obstétricales.

Puisque la période d'observation contient les huit premiers mois durant la grossesse de l'accouchement d'intérêt pour la prédiction, les visites prénatales ont été identifiées dans les données afin de créer un ensemble de variables explicatives. La médiane du nombre de visites prénatales a été imputée pour les patientes dont il n'a pas été possible d'établir le moment de la première visite. Des variables binaires ont été créées pour les visites de prise en charge de grossesse, les visites

prénatales, les visites prénatales à risque élevé de complications et les visites en obstétrique. En utilisant la même méthodologie que celle utilisée pour créer les variables de coûts des soins de santé, le nombre de visites par mois a été estimé afin de créer un ensemble de variables permettant de capter l'information en lien avec le suivi prénatal de la patiente. Également, puisque l'accès aux soins de santé de première ligne passe régulièrement par les centres hospitaliers, la même méthode a été employée pour les visites dans un établissement hospitalier.

Les données sur les prescriptions contiennent les informations relatives aux programmes d'assurance médicament, les médications prescrites, le type de médication, la date de la prescription, la classe de médicament et la durée du traitement. Puisqu'au cours de la grossesse, les patientes arrêtent de prendre les médications qui risquent de détériorer leur condition de santé ou de mettre en péril la santé de l'enfant, les durées de traitement ont été utilisées afin de créer des variables dichotomiques et des variables de comptage dans le but de capter la prise de médication durant la période d'observation. Ces variables vont permettre de distinguer les prescriptions de courte durée et les prescriptions pour des traitements échelonnés sur une plus longue durée, et ce pour tous les différents types de médications.

Plusieurs catégories de complications durant la grossesse et l'accouchement ont été identifiées à partir de la littérature afin de créer la variable dépendante. En tout, treize catégories de complications ont été créées : les complications liées au fœtus, les hémorragies, les accouchements résultant en avortement, l'éclampsie et la pré-éclampsie, les problèmes d'hypertension, les accouchements en césarienne, les déchirures vaginales importantes, la dystocie, les accouchements prématurés, les infections, les complications liées au poids du nouveau-né, les complications multiples et une dernière catégorie regroupant les diagnostics de complications n'entrant pas dans les autres catégories. De plus, les complications ont également été identifiées durant la période d'observation pour permettre d'identifier les pa-

tientes qui ont eu des complications lors d'accouchements précédents ou au début de la période prénatale de l'accouchement pour lequel la prédiction est effectuée. La variable dépendante prend la valeur de 1 si la patiente a un diagnostic ou un acte médical appartenant à une de ces catégories, et zéro sinon. Parmi toutes les catégories de complications, la plus fréquente est l'accouchement en césarienne, survenant dans 93,76 % des cas de complications.

3.4 Variables pour l'estimation des modèles

Une fois que les données ont été modifiées et appariées, la base de données contient 3 584 variables explicatives. Les variables explicatives retenues sont les variables de coûts, de diagnostics, d'actes médicaux, de prescriptions, les variables pour les visites à l'hôpital et celles pour les visites en obstétrique. Des variables d'interactions ont également été créées pour les interactions entre le groupe d'âge et les diagnostics et pour les interactions entre les variables de diagnostics. Les variables d'interactions ont été créées afin de permettre au modèle logit pénalisé de considérer ces variables lorsqu'il sélectionne les variables explicatives à inclure dans le modèle.

Cependant, plusieurs variables de diagnostic binaires ont peu d'observations prenant une valeur différente de 0. Ceci fait en sorte qu'un nombre important de variables d'interactions sont constantes avec une variance nulle. Puisque les colonnes constantes n'ajoutent aucune information pertinente pour l'entrainement des modèles, elles ont toutes été retirées des données. Ainsi, l'échantillon final contient 1 995 variables explicatives qui peuvent être utilisées pour prédire les complications durant la grossesse et l'accouchement.

Un prétraitement des données a été effectué afin de favoriser l'apprentissage des méthodes. Puisque certaines des méthodes d'apprentissage automatique utilisées

sont sensibles à l'échelle des variables explicatives, les variables de comptage et les variables continues ont été standardisées de sorte qu'elles aient une moyenne égale à 0 et une variance égale à 1 (James *et al.*, 2013). En modifiant les variables explicatives afin qu'elles aient une échelle comparable, la technique de standardisation permet d'améliorer l'apprentissage des méthodes sans compromettre l'interprétation des résultats des coefficients obtenus à l'aide de la méthode de régression.

CHAPITRE IV

RÉSULTATS

La présentation des résultats obtenus débutera avec les résultats des méthodes de ré-échantillonnage. Ensuite, les résultats des méthodes d'apprentissage automatique avec chacun des échantillons d'apprentissage seront présentés. Seulement les meilleurs modèles estimés ont été retenus. Finalement, une prédiction des coûts des diagnostics de complications durant la grossesse et l'accouchement sera effectuée en prenant les meilleurs modèles estimés et en imputant la moyenne des coûts à chaque patiente. Cette prédiction sera comparée avec les coûts réels durant la période de résultat aux fins de déterminer le modèle qui prédit le mieux les coûts des diagnostics de complications.

4.1 Choix du ratio entre la classe minoritaire et la classe majoritaire pour les méthodes de ré-échantillonnage

Les résultats des méthodes de ré-échantillonnage sont présentés dans la figure 4.1. Les lignes pleines et pointillées représentent la performance du modèle pour l'échantillon d'apprentissage et l'échantillon test. La méthode de forêt aléatoire a été utilisée pour l'estimation des modèles qui ont servi à déterminer la valeur du paramètre q. Les modèles ont été estimés avec 400 arbres de décisions, $m = \sqrt{p}$ variables explicatives (où p représente le nombre total de variables disponibles)

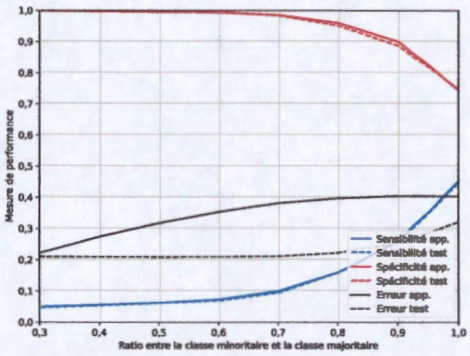
et avec une profondeur maximale $d = 4$. La classe minoritaire correspond aux patientes qui ont au moins un diagnostic de complications et la classe majoritaire correspond aux patientes qui ont une grossesse normale.

Comme l'échantillon contient 11 951 patientes qui ont des complications et 42 049 patientes qui ont une grossesse normale, le ratio initial entre les classes est de $\frac{N_{\text{minoritaire}}}{N_{\text{majoritaire}}} = \frac{11951}{42049} = 0,2842$. Par conséquent, une valeur de $q = 0,3$ revient à faire très peu de ré-échantillonnage par rapport aux nombres d'observations dans chacune des classes, alors qu'une valeur de $q = 1$ revient a équilibrer complètement les classes. Les résultats présentés dans la figure 4.1 démontrent que les méthodes de ré-échantillonnage ont un effet important sur la performance du modèle.

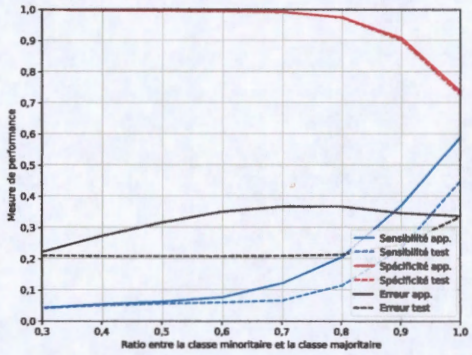
Pour toutes les méthodes, lorsque $q = 0,3$ le modèle estimé fait une prédiction parfaite pour les patientes qui n'ont pas de complications et il commet une erreur de prédiction pour presque toutes les patientes qui ont des complications, avec une valeur légèrement supérieure à 0,2 pour l'erreur de classification, soit environ la moyenne de la variable dépendante. En regardant seulement l'erreur de classification, un modèle qui prédit une grossesse normale à presque tout l'échantillon peut sembler avoir une bonne performance puisqu'il minimise sa valeur. Cependant, puisque les patientes qui ont des complications ont des coûts de soins de santé plus élevés, un tel modèle risque de sous-estimer grandement les coûts des complications.

Pour de plus grandes valeurs de q, la sensibilité du modèle augmente et la spécificité diminue, ce qui a pour effet d'améliorer la prédiction pour les patientes qui ont des complications. Par contre, en biaisant la prédiction du modèle vers la classe minoritaire, le modèle fait moins bien pour prédire les grossesses normales ce qui a pour effet d'augmenter l'erreur de classification. La valeur de q qui offre la meilleure performance est lorsque les classes sont complètement équilibrées, sauf

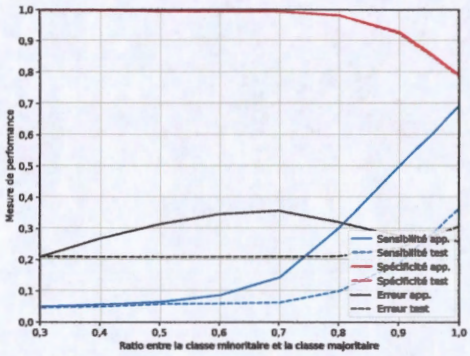
pour la technique SMOTE avec la technique de sous-échantillonnage ENN qui prédit des complications à presque tout l'échantillon.



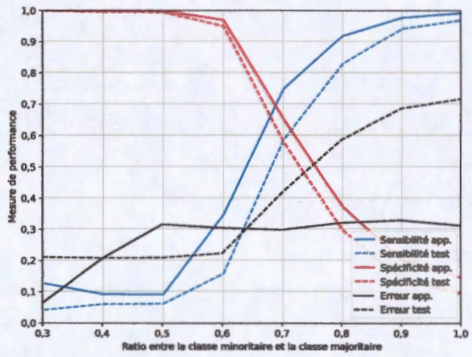
(a) Sous-échantillonnage aléatoire de la classe majoritaire



(b) SMOTE et technique de sous-échantillonnage aléatoire de la classe majoritaire



(c) SMOTE et technique de sous-échantillonnage de la classe majoritaire avec la méthode des liens Tomek



(d) SMOTE et technique de sous-échantillonnage avec la méthode ENN

Source : Calculs de l'auteur à partir des données administratives.

Figure 4.1: Résultats des méthodes de ré-échantillonnage

Pour la technique SMOTE, la quantité de sur-échantillonnage de la classe minoritaire choisie est $q = 0,5$. Ensuite, des observations de la classe minoritaire ont été retirées aléatoirement jusqu'à ce que les classes de la variable dépendantes

soient équilibrées. Pour la technique SMOTE avec la méthode ENN, la quantité de ré-échantillonnage $q = 0,65$ est celle qui a été sélectionnée pour l'estimation des modèles. Pour la technique de sous-échantillonnage aléatoire et pour la technique SMOTE avec les liens Tomek, la valeur de $q = 1$ a été sélectionnée.

Le tableau 4.1 présente le nombre d'observations pour chacune des classes de la variable dépendante dans les échantillons d'apprentissages modifiés. En comparant le nombre d'observations restantes entre la combinaison de la technique SMOTE avec les liens Tomek et la combinaison de la technique SMOTE avec ENN, il est possible de remarquer que la méthode ENN tend à retirer beaucoup plus d'observations que les liens Tomek lorsqu'elle sous-échantillonne des observations, ce qui confirme le résultat obtenu par Batista *et al.* (2004). Ces échantillons sont ceux qui seront utilisés afin d'estimer les modèles pour faire la prédiction des diagnostics de complications.

Tableau 4.1: Nombre d'observations pour chacune des classes de la variable dépendante

	Échantillon d'apprentissage		Échantillon test	
Méthode de ré-échantillonnage :	$y_i = 0$	$y_i = 1$	$y_i = 0$	$y_i = 1$
Aucune	33 599	9 601	8 450	2 350
Sous-échantillonage aléatoire	9 601	9 601	8 450	2 350
SMOTE et sous-échantillonnage aléatoire	16 799	16 799	8 450	2 350
SMOTE et liens Tomek	33 545	33 545	8 450	2 350
SMOTE et ENN	21 105	19 207	8 450	2 350

4.2 Résultats des méthodes d'apprentissage automatique

Les résultats des méthodes d'apprentissage automatique sont présentés dans les tableaux 4.2 et 4.3. Les mesures de performances estimées afin de comparer les

modèles sont la sensibilité, la spécificité, la précision et l'aire sous la courbe ROC (AUC). Les colonnes des tableaux représentent la méthode utilisée pour estimer le modèle de prédiction et les lignes correspondent à l'échantillon d'apprentissage qui a été utilisé pour l'estimation. La performance des modèles avec l'échantillon d'apprentissage initial servira de référence pour les estimations réalisées avec les méthodes de ré-échantillonnage. Le méta estimateur est le modèle estimé avec la méthode d'ensemble d'apprentissage, qui consiste à combiner les meilleurs modèles de chaque méthode afin de construire un seul modèle.

La performance des modèles entrainés avec l'échantillon d'apprentissage initial est très similaire à la performance du modèle de forêt aléatoire utilisé pour le choix du paramètre q. En moyenne, les modèles classifient adéquatement 6 % des patientes qui ont des complications et 99 % des patientes qui ont des grossesses normales, avec une précision de 79 % et une valeur de 0,53 pour l'AUC. Les résultats des modèles estimés avec l'échantillon d'apprentissage sont similaires à ceux obtenus avec l'échantillon test. La méthode de vecteurs de support linéaire (SVC) est celle qui prédit le mieux les patientes qui ont des complications, et une sensibilité de seulement 10 %.

Pour l'ensemble des méthodes de ré-échantillonnage, les résultats des modèles estimés avec les échantillons modifiés démontrent que ces méthodes améliorent grandement la performance des modèles pour la prédiction des complications. La combinaison de la technique SMOTE avec les liens Tomek est celle qui améliore le plus la prédiction des complications pour l'échantillon d'apprentissage, prédisant correctement en moyenne 69 % des patientes qui ont des complications et 73 % des patientes qui ont des grossesses normales, avec une précision de 71 % et une valeur de 0,71 pour l'AUC.

Tableau 4.2: Performance des modèles pour l'échantillon d'apprentissage

	Logit pénalisé	SVC	Forêt aléatoire	*Boosting*	AdaBoost	Méta-estimateur	Moyenne
Sensibilité							
Original	0,08	0,10	0,08	0,00	0,03	0,06	0,06
Aléatoire	0,57	0,58	0,55	0,49	0,48	0,58	0,54
SMOTE et aléatoire	0,59	0,62	0,63	0,62	0,63	0,69	0,63
SMOTE et Tomek	0,63	0,63	0,72	0,68	0,74	0,76	0,69
SMOTE et ENN	0,60	0,61	0,67	0,53	0,66	0,67	0,62
Moyenne	0,50	0,51	0,53	0,46	0,51	0,55	0,51
Spécificité							
Original	0,99	0,99	1,00	1,00	1,00	1,00	0,99
Aléatoire	0,66	0,66	0,68	0,71	0,75	0,68	0,69
SMOTE et aléatoire	0,63	0,62	0,74	0,70	0,83	0,70	0,70
SMOTE et Tomek	0,59	0,60	0,86	0,80	0,83	0,72	0,73
SMOTE et ENN	0,70	0,71	0,82	0,91	0,89	0,79	0,80
Moyenne	0,71	0,72	0,82	0,82	0,86	0,78	0,78
Précision							
Original	0,79	0,79	0,79	0,78	0,78	0,79	0,79
Aléatoire	0,61	0,62	0,61	0,60	0,62	0,63	0,62
SMOTE et aléatoire	0,61	0,62	0,69	0,66	0,73	0,69	0,67
SMOTE et Tomek	0,61	0,62	0,79	0,74	0,78	0,74	0,71
SMOTE et ENN	0,65	0,66	0,75	0,73	0,78	0,73	0,72
Moyenne	0,66	0,66	0,72	0,70	0,74	0,72	0,70
AUC							
Original	0,54	0,54	0,54	0,50	0,52	0,53	0,53
Aléatoire	0,61	0,62	0,61	0,60	0,62	0,63	0,62
SMOTE et aléatoire	0,61	0,62	0,69	0,66	0,73	0,69	0,67
SMOTE et Tomek	0,61	0,62	0,79	0,74	0,78	0,74	0,71
SMOTE et ENN	0,65	0,66	0,74	0,72	0,77	0,73	0,71
Moyenne	0,60	0,61	0,67	0,64	0,68	0,66	0,65

Note : La sensibilité correspond au pourcentage des diagnostics de complications qui ont été correctement prédits et la spécificité correspond au pourcentage des grossesses normales qui ont été correctement prédites. La précision est le pourcentage des observations pour lesquelles la classe prédite est celle qui est observée.

Pour les modèles logit et SVC, l'ajout de la technique SMOTE aux techniques de sous-échantillonnage ne semble pas avoir un grand effet comparativement aux modèles estimés avec les méthodes en arbre. Pour ces derniers, la combinaison des techniques de ré-échantillonnage améliore considérablement leur performance, avec une valeur maximale pour l'AUC allant jusqu'à 0,79. Par conséquent, la combinaison de la technique SMOTE avec des techniques de sous-échantillonnage améliore les modèles prédictifs comparativement à utiliser seulement une technique de sous-échantillonnage ou l'échantillon initial, ce qui va dans le même sens que les résultats de Chawla *et al.* (2002), Batista *et al.* (2003) et Batista *et al.* (2004).

En se basant seulement sur l'AUC, le meilleur modèle serait le modèle de forêt aléatoire avec la combinaison de la technique SMOTE et la méthode de sous-échantillonnage avec les liens Tomek. Il s'agit également d'un des meilleurs modèles pour la sensibilité, la spécificité et la précision, ce qui indique que le modèle performe bien pour faire la prédiction des patientes qui ont des complications et celles qui ont des grossesses normales.

Tel qu'illustré dans le tableau 4.3, pour l'échantillon test, la technique de sous-échantillonnage aléatoire est celle qui améliore le plus la prédiction pour les patientes qui ont des complications. En moyenne, l'ensemble des modèles estimés avec cette méthode prédisent correctement les complications dans 52 % des cas, ce qui est nettement supérieur au résultat obtenu avec l'échantillon d'apprentissage original. Il s'agit également de la méthode qui fait mieux en moyenne pour l'AUC, avec une valeur moyenne de 0,59. En moyenne, la technique SMOTE avec la méthode ENN est celle qui améliore le plus la performance des modèles pour la précision et la spécificité. En regardant uniquement l'AUC, les modèles de forêt aléatoire et le méta estimateur utilisés avec la technique de sous-échantillonnage aléatoire sont ceux qui affichent la meilleure performance, avec une valeur de 0,60.

Tableau 4.3: Performance des modèles pour l'échantillon test

	Logit pénalisé	SVC	Forêt aléatoire	*Boosting*	AdaBoost	Méta-estimateur	Moyenne
Sensibilité							
Original	0,08	0,08	0,07	0,00	0,03	0,05	0,05
Aléatoire	0,56	0,54	0,54	0,47	0,46	0,55	0,52
SMOTE et aléatoire	0,57	0,56	0,46	0,48	0,35	0,51	0,49
SMOTE et Tomek	0,58	0,56	0,28	0,35	0,29	0,47	0,42
SMOTE et ENN	0,55	0,53	0,42	0,26	0,30	0,47	0,42
Moyenne	0,47	0,45	0,35	0,31	0,29	0,41	0,38
Spécificité							
Original	0,99	0,98	0,99	1,00	1,00	1,00	0,99
Aléatoire	0,63	0,62	0,66	0,70	0,72	0,64	0,66
SMOTE et aléatoire	0,60	0,58	0,72	0,71	0,80	0,67	0,68
SMOTE et Tomek	0,58	0,59	0,85	0,80	0,83	0,70	0,72
SMOTE et ENN	0,62	0,62	0,76	0,88	0,84	0,71	0,74
Moyenne	0,68	0,68	0,80	0,82	0,84	0,74	0,76
Précision							
Original	0,79	0,79	0,79	0,78	0,79	0,79	0,79
Aléatoire	0,61	0,61	0,63	0,65	0,67	0,62	0,63
SMOTE et aléatoire	0,59	0,58	0,67	0,66	0,71	0,64	0,64
SMOTE et Tomek	0,58	0,58	0,73	0,70	0,71	0,65	0,66
SMOTE et ENN	0,61	0,60	0,69	0,75	0,72	0,66	0,67
Moyenne	0,64	0,63	0,70	0,71	0,72	0,67	0,68
AUC							
Original	0,53	0,53	0,53	0,50	0,51	0,52	0,52
Aléatoire	0,59	0,58	0,60	0,58	0,59	0,60	0,59
SMOTE et aléatoire	0,59	0,57	0,59	0,59	0,58	0,59	0,59
SMOTE et Tomek	0,58	0,57	0,57	0,57	0,56	0,58	0,57
SMOTE et ENN	0,59	0,58	0,59	0,57	0,57	0,59	0,58
Moyenne	0,58	0,57	0,58	0,56	0,56	0,58	0,57

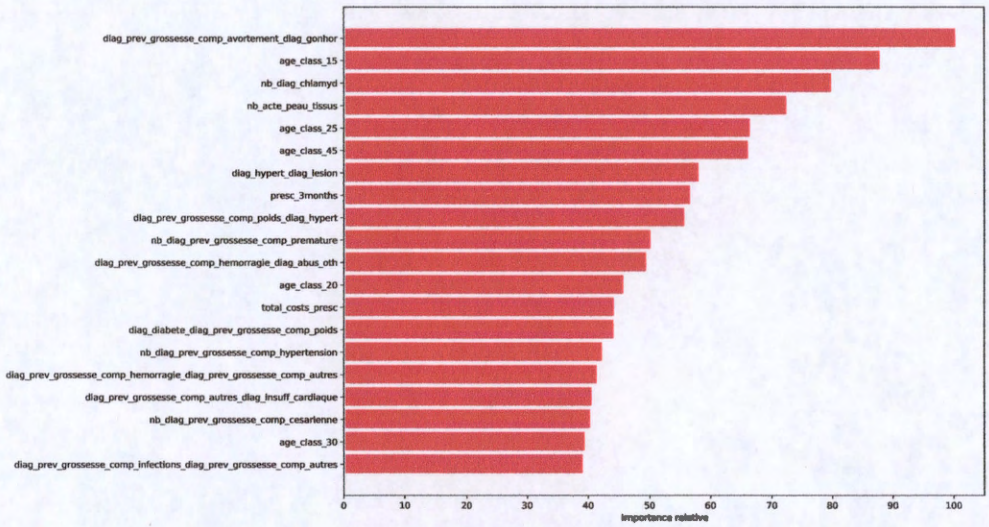
Note : La sensibilité correspond au pourcentage des diagnostics de complications qui ont été correctement prédits et la spécificité correspond au pourcentage des grossesses normales qui ont été correctement prédites. La précision est le pourcentage des observations pour lesquelles la classe prédite est celle qui est observée.

4.2.1 Importance des variables pour la prédiction des complications

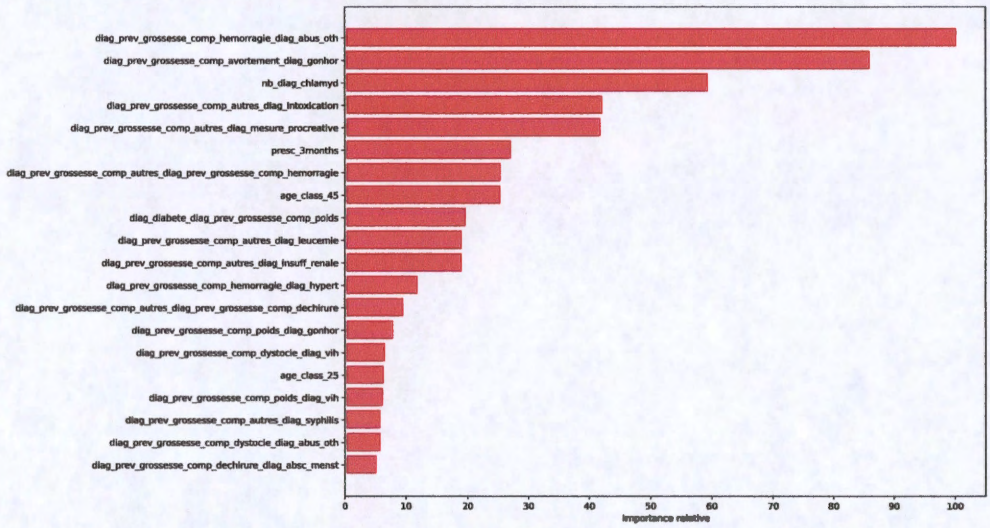
Pour les méthodes en arbre, l'importance des variables explicatives pour la prédiction des complications peut être donnée en utilisant l'index de Gini (James *et al.*, 2013). Pour déterminer l'importance des variables, la diminution totale de l'index de Gini pour l'ajout d'une variable explicative est estimée à chacune des régions R_m de l'arbre de décision. La moyenne de la contribution de chaque variable pour l'ensemble des arbres B estimés est estimée et les variables sont ordonnées de façon décroissante. Une grande valeur indique que la variable est importante pour la prédiction.

La figure 4.2 illustre l'importance des variables pour le modèle AdaBoost avec la combinaison de la technique SMOTE et des liens Tomek et le modèle de forêt aléatoire avec la technique de sous-échantillonnage aléatoire. En premier lieu, il est possible de remarquer que les deux modèles accordent une grande importance aux interactions entre les complications antérieures et les variables de diagnostics et au nombre total de prescriptions dans les trois derniers mois de la période d'observation.

Parmi ces diagnostics on retrouve la chlamydia, la gonorrhée, l'hypertension, le diabète et les accouchements antérieurs prématurés. Le modèle AdaBoost accorde une grande importance aux catégories d'âge pour les patientes qui ont entre 15 et 19 ans, et les patientes qui ont entre 45-49 ans. La seule variable de coûts qui apparait pour les deux modèles est la somme des coûts totaux des prescriptions pour le modèle AdaBoost. De plus, la plupart des variables auxquelles les modèles accordent une grande importance font partie des déterminants des complications identifiés dans la littérature.



(a) AdaBoost avec SMOTE et liens Tomek



(b) Forêt aléatoire avec SMOTE et sous-échantillonnage aléatoire

Source : Calculs de l'auteur à partir des données administratives.

Figure 4.2: Importance relative des variables

4.3 Prédiction des coûts des diagnostics de complications

Les coûts des diagnostics de complications ont été prédits en prenant l'espérance conditionnelle des coûts des soins de santé durant la période d'observation sachant la classe prédite. Dans la mesure où les échantillons d'apprentissage ont été modifiés par les méthodes de ré-échantillonnage, la prédiction des coûts a été effectuée seulement avec les observations qui se retrouvent dans l'échantillon test.

Premièrement, une prédiction des complications a été effectuée pour chacun des modèles estimés en utilisant les observations dans l'échantillon test. Par la suite, l'espérance conditionnelle des coûts a été estimée pour les patientes dans l'échantillon test et la valeur de cette moyenne a été imputée pour chacune des patientes, selon la classe prédite. Les résultats de la prédiction des coûts des complications sont présentés dans le tableau 4.4 ci-dessous.

Tableau 4.4: Résultats des prédictions des coûts des complications

	Logit pénalisé	SVC	Forêt aléatoire	*Boosting*	AdaBoost	Méta-estimateur	Moyenne
Original	0,91	0,91	0,91	0,90	0,90	0,90	0,90
Aléatoire	1,09	1,09	1,08	1,06	1,05	1,09	1,08
SMOTE et aléatoire	1,11	1,11	1,05	1,06	1,01	1,07	1,07
SMOTE et Tomek	1,11	1,11	0,98	1,01	0,99	1,06	1,04
SMOTE et ENN	1,10	1,09	1,03	0,97	0,99	1,05	1,04
Moyenne	1,06	1,06	1,01	1,00	0,99	1,03	1,03

Note : Les ratios entre la somme totale des coûts prédits par le modèle et les coûts totaux des complications sont présentés.

Les valeurs dans le tableau représentent le ratio entre la prédiction de coûts du modèle et le coût total des complications pour l'échantillon test. Lorsque le ratio est plus petit que 1, le modèle sous-estime les coûts des complications puisqu'il

prédit un montant des coûts inférieurs par rapport au coût total de l'échantillon. Similairement, le modèle surestime les coûts si le ratio est plus grand que 1. Ainsi, des valeurs proche de 1 signifient que le modèle prédit bien les coûts pour l'échantillon d'entrainement. La valeur totale des coûts des soins de santé pour l'échantillon durant la période d'observation est d'environ 815 millions de dollars. Ainsi, même lorsque la différence entre la valeur observée et la valeur prédite est faible, une mauvaise prédiction des coûts des complications est susceptible d'entraîner une dépense inattendue importante pour les établissements de santé.

Conformément aux résultats observés pour la prédiction des complications, les modèles estimés avec l'échantillon initial sous-estiment les coûts de 10 % en moyenne. De plus, ce sont les seuls modèles qui sous-estiment les coûts de manière aussi importante, ce qui indique que le coût des erreurs de prédictions lorsque la valeur observée est un diagnostic de complications est potentiellement plus grand que lorsque la patiente a une grossesse normale et que le modèle prédit une complication.

Les modèles linéaires et le méta estimateur sont ceux qui ont la pire performance pour la prédiction des coûts. Lorsqu'ils sont estimés avec les techniques de rééchantillonnage, ils surestiment considérablement les coûts par rapport aux méthodes en arbre. Les modèles qui font le mieux pour la prédiction des coûts sont les modèles *Boosting* et AdaBoost. Le modèle Boosting fait mieux avec la technique SMOTE combinée aux liens Tomek et le modèle AdaBoost fait mieux avec la méthode SMOTE combinée à la technique de sous-échantillonnage aléatoire.

CHAPITRE V

DISCUSSION

5.1 Discussion sur la performance des modèles

D'abord, les méthodes de ré-échantillonnage sont des techniques qui peuvent considérablement améliorer la performance des méthodes d'apprentissage automatique. Tel qu'il a été démontré avec les résultats présentés dans cette recherche, les modèles estimés avec la technique SMOTE avec des techniques de sous-échantillonnage ont une performance nettement supérieure à celle des modèles estimés avec l'échantillon original. Ces modèles prédisent mieux les patientes qui ont des complications, ce qui permet de faire une bonne prédiction des coûts engendrés par les complications durant la grossesse et l'accouchement.

Pour la prédiction des complications, les meilleurs modèles sont les modèles en arbres qui ont été estimés avec un échantillon d'apprentissage modifié à l'aide de la technique SMOTE et d'une technique de sous-échantillonnage. En moyenne, ces modèles permettent de prédire correctement les complications durant la grossesse 74 % du temps, en plus d'identifier correctement 83 % des patientes qui n'ont pas de complications. Ils permettent également de prédire les coûts des complications avec une précision importante. De plus, les modèles qui font bien pour la sensibilité et la spécificité font aussi généralement bien pour l'AUC et la précision. Ces résultats suggèrent qu'il existe une relation non-linéaire entre la probabilité d'avoir

des complications durant la grossesse et les conditions de santé de la patiente. De plus, lorsqu'elles sont correctement utilisées, les méthodes en arbre sont des méthodes qui performent généralement mieux que des modèles linéaires, ce qui a été confirmé par les résultats obtenus.

La combinaison des variables de santé avec les variables de coûts permet de modéliser plutôt bien la relation entre les complications durant la grossesse et les conditions de santé de la patiente. Les variables auxquelles les meilleurs modèles accordent le plus grand poids pour la prédiction sont des déterminants des complications qui ont été identifiés dans la littérature présentée.

5.2 Discussion sur les données

Les données utilisées contenaient seulement de l'information sur les diagnostics, les actes médicaux, les prescriptions et les coûts médicaux. Mis à part le groupe d'âge et l'information sur l'assurance médicament, peu de variables à caractère socioéconomique étaient disponibles. Ces déterminants étant importants pour les complications durant la grossesse et l'accouchement, l'ajout de ces variables pour la prédiction des complications pourrait permettre de mieux modéliser la relation entre la probabilité de complications et les conditions de santé de la patiente.

De plus, aucune information biométrique sur les patientes n'était disponible, ce qui ne permettait pas de prendre en compte directement les caractéristiques physiques des patientes dans les modèles. Puisque plusieurs problèmes de santé peuvent être latents et mettre un certain temps avant de se manifester, de tels indicateurs pourraient aider à mieux prédire les diagnostics en identifiant mieux les patientes qui sont sujettes à avoir des problèmes d'hypertension ou de diabète par exemple.

L'historique médical de la patiente était disponible seulement deux ans avant l'accouchement, incluant la période prénatale. Avoir un plus grand horizon temporel

pour la période d'observation permettrait d'augmenter le nombre d'observations pour la création des variables explicatives servant à faire la prédiction, en plus de permettre d'avoir plus d'information à propos des accouchements antérieurs et des conditions de santé de la patiente avant et pendant ces accouchements. De plus, obtenir davantage d'observations avant l'accouchement permettrait de faire la prédiction des complications plus d'un mois à l'avance, peut-être même avant que la patiente se rende compte qu'elle est enceinte. Ceci permettrait de construire des modèles prédictifs ayant pour but l'identification des risques associés à la grossesse au début de la période prénatale, ce qui laisserait plus de temps aux intervenants du milieu de la santé pour tenter d'améliorer les conditions prénatales de la patiente et réduire le risque de complications.

5.3 Recommandations politiques

Les prédictions des facteurs de risque entourant la grossesse et l'accouchement permettent de dégager les ressources nécessaires pour répondre à la demande de soins de santé à une période qui est cruciale pour la santé de la mère et de l'enfant. Une prédiction des diagnostics pourrait être un outil d'aide à la décision aux professionnels de la santé afin de planifier le déroulement de l'accouchement de la patiente. De plus, tel qu'il a été démontré, un gouvernement pourrait également utiliser ces prédictions de diagnostics comme outil de planification budgétaire pour les soins en obstétrique et en gynécologie.

Considérant que les coûts des soins de santé associés à la grossesse et l'accouchement sont considérables, une politique publique d'implantation d'un système informatique basé sur un algorithme de prédiction des complications pourrait permettre au gouvernement de mieux anticiper les coûts futurs et d'optimiser la gestion de ses ressources financières. Puisque les ressources financières d'un gou-

vernement sont limitées, une meilleure planification budgétaire peut permettre de débloquer des ressources financières qui peuvent être investies ailleurs dans le réseau de la santé et dans la société.

CONCLUSION

L'objectif de premier plan de ce mémoire est la prédiction des coûts des diagnostics de complications durant la grossesse et l'accouchement. Les méthodes qui ont été utilisées sont des méthodes d'apprentissage automatique combinées à des méthodes de ré-échantillonnage. Les données utilisées pour l'estimation des modèles contiennent les dossiers médicaux de 54 000 femmes qui ont donné naissance entre les années 1998 à 2006. Les variables utilisées incluent l'information sur les diagnostics, les actes médicaux, les prescriptions, les coûts des soins de santé, le groupe d'âge de la patiente et les régimes d'assurance médicaments.

À partir des données, 1 995 variables explicatives ont été créées pour l'estimation des modèles. Les trajectoires mensuelles des coûts des soins de santé, des visites dans un centre hospitalier et les visites en obstétrique ont été modélisées afin de représenter l'évolution temporelle des coûts des soins de santé de chacune des patientes. Ces trajectoires ont servi à créer un ensemble de variables qui permettent de capter cette évolution.

Les méthodes utilisées permettent de prédire les complications durant la grossesse et les coûts de ces diagnostics. Les meilleurs modèles sont les modèles de forêt aléatoire, de *boosting* et AdaBoost estimés avec la technique SMOTE et des méthodes de ré-échantillonnage. Les modèles permettant de prédire les complications sont également les modèles qui ont la meilleure performance pour la prédiction des coûts.

Les variables de diagnostics et de groupe d'âge sont les variables les plus importantes pour la prédiction des complications, sauf pour les coûts totaux des

prescriptions et le nombre total de prescriptions les trois derniers mois. Les diagnostics de complications antérieures, l'hypertension, le diabète et les maladies transmises sexuellement sont les diagnostics qui sont les plus déterminants pour les complications. Les variables de coûts n'ont pas été retenues par les modèles ayant la meilleure performance pour la prédiction. Les patientes qui sont âgée entre 15 et 19 ans, entre 25 et 29 ans et qui ont plus de 45 ans sont les patientes qui sont plus à risque d'avoir des complications.

Les méthodes développées peuvent être utilisées par un gouvernement afin de mieux prévenir les risques de complications et prédire les coûts associés à la demande de soins de santé entourant le jour de l'accouchement. La méthodologie développée pour la construction des données et des modèles pourrait également permettre de prédire d'autres diagnostics que les diagnostics de complications, et pourraient être étendus à d'autres départements que les départements qui s'occupent de la grossesse.

Les principaux défis rencontrés ont été l'identification des diagnostics et des actes pertinents pour la création des variables, l'appariement des données, et la calibration des modèles d'apprentissage avec les méthodes de ré-échantillonnage. Des recherches futures pourraient appliquer une méthodologie similaire avec plus de données socioéconomiques, biométriques et un horizon temporel plus long. La prédiction des coûts pourrait également être conditionnée à certaines valeurs des variables qui sont les plus utilisées par les modèles pour la prédiction des complications. Les extensions de recherche proposée, soit l'amélioration des données et la considération d'autres facteurs que la prédiction des coûts des complications, devraient être explorées.

Plusieurs développements récents dans le domaine de l'apprentissage automatique sont prometteurs pour l'avenir de ces méthodes. En pratique, des méthodes d'ap-

prentissage profond sont appliquées couramment afin de résoudre des problèmes de reconnaissance d'image ou de traitement de langage naturel. Dans le milieu médical, ces techniques peuvent être utilisées dans le but de prédire l'apparition d'une tumeur à partir d'imageries médicales, ou encore pour prédire la probabilité de dépression à partir de notes manuscrites prises par différents intervenants médicaux. De plus, l'introduction de méthodes bayésiennes dans les méthodes d'apprentissage permet de régler certains problèmes liés à la taille grandissante des bases de données, au nombre de paramètres à estimer, et à la généralisation des modèles lorsqu'ils rencontrent de nouvelles observations. Entre autres, les méthodes bayésiennes peuvent être utilisées pour l'estimation des paramètres des méthodes d'apprentissage ou pour comparer un ensemble de modèles dans le but sélectionner le meilleur modèle estimé afin de faire une prédiction.

En résumé, lorsque des masses de données importantes sont combinées à des méthodes d'apprentissage automatique, il est possible de modéliser des relations complexes qui permettent de créer des modèles prédictifs puissants qui fonctionnent bien hors de l'échantillon d'apprentissage. La démocratisation de l'accès à des ordinateurs avec une puissance de calculs importante combinée à l'augmentation considérable de la taille des bases de données provoque un engouement et un contexte favorable au déploiement de ces méthodes dans plusieurs sphères de la société. La méthodologie qui a été élaborée dans cette recherche permet de prédire les diagnostics de complications durant la grossesse et l'accouchement ainsi que les coûts de ces diagnostics.

APPENDICE A

Tableau A.1 Liste des variables utilisées pour l'estimation des modèles

Colonnes	Description
1-8	Variables binaires indiquant le groupe d'âge de la patiente et une variable binaire qui indique si la patiente a plus de 35 ans
9-20	Variables de comptage pour certaines prescriptions de certains anti-inflammatoires et anti-dépresseurs.
21-27	Variables de comptage pour la durée des prescriptions, toutes prescriptions confondues.
28-33	Variables binaires pour le programme d'assurance médicaments, par exemple les patientes qui sont sur un régime public ou qui celles qui sont sur l'aide sociale.
34-46	Variables binaires pour les prescriptions et pour la durée des prescriptions.

Tableau A.1 : Liste des variables utilisées pour l'estimation des modèles (suite)

Colonnes	Description
47-232	Variables de comptage pour les actes médicaux et pour les diagnostics. Parmi les actes médicaux, on retrouve : — référencement à un médecin spécialiste ; — plusieurs types d'avortements ; — césarienne ; — visite prénatale ; — visite prénatale à risque élevé ; — visite de prise en charge de la grossesse ; — visite chez un médecin spécialiste en obstétrique ; — menace d'avortement ; — extraction manuelle du placenta ; — procréation assistée ; — anesthésie obstétricale ; — émission d'un certificat de retrait préventif ; — les neuf catégories d'actes médicaux ; — test de réactivité fœtale. Les variables de diagnostics contiennent ont été crées à partir des diagnostics suivants : — les infections transmises sexuellement ; — les problèmes d'hypertension ; — le diabète ; — les psychoses ; — les troubles dépressifs ; — les troubles de poids ; — les mesures procréatives ; — les intoxications (alcool, drogues, etc.) ; — les problèmes cardiaques ; — les problèmes de reins ; — les tumeurs, — les troubles alimentaires ; — les hémorragies ; — les fractures ; — l'hypoglycémie ; — les lésions ; — la souffrance fœtale ; — la stérilité ; — les complications durant la grossesse survenant durant la période d'observation.

Tableau A.1 : Liste des variables utilisées pour l'estimation des modèles (suite)

Colonnes	Description
233	Variable dépendante binaire indiquant si une patiente a eu au moins un diagnostic de complications.
234-239	Variable pour la somme et la moyenne des coûts totaux, des coûts des prescriptions et des coûts des actes médicaux pour la période d'observation.
240-262	Somme des coûts des actes médicaux pour chacun des mois de la période d'observation
263-285	Somme des coûts des prescriptions pour chacun des mois de la période d'observation
286-308	Nombre de visites dans un centre hospitalier pour chacun des mois de la période d'observation
309-331	Nombre de visites dans un département d'obstétrique pour chacun des mois de la période d'observation
332-335	Sommes pour les coûts des prescriptions et les coûts des actes médicaux pour les trois et six derniers mois de la période d'observation.
336-339	Moyennes pour les coûts des prescriptions et les coûts des actes médicaux pour les trois et six derniers mois de la période d'observation.
340-343	Variables binaires indiquant si la patiente connait une hausse ou une diminution dans les coûts de soins de santé pour les trois derniers mois de la période d'observation.
344-345	Nombre de mois pour lesquels les coûts des prescriptions et les coûts des actes médicaux étaient supérieurs à la moyenne.
346-347	Valeur maximale pour les coûts mensuels des prescriptions et des actes médicaux.
348-355	Nombre de visites à l'hôpital pendant les trois et six derniers mois, nombre de visites en obstétrique pendant les trois et six derniers mois, nombre de mois au dessus de la moyenne et nombre de visites maximales pour les visites en à l'hôpital et en obstétrique.

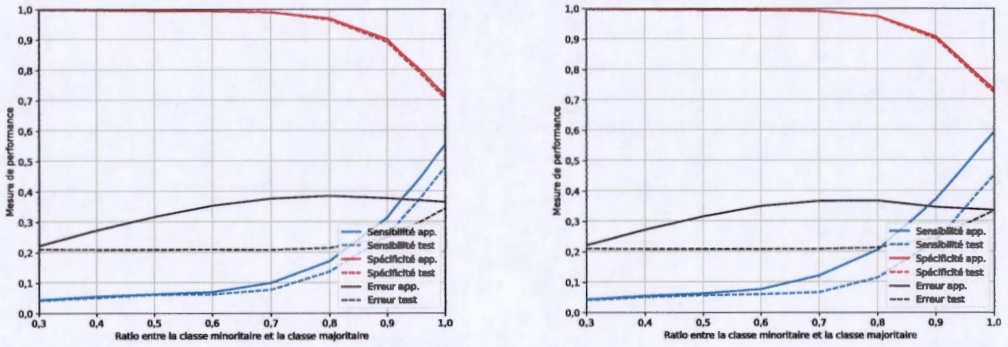
Tableau A.1 : Liste des variables utilisées pour l'estimation des modèles (suite)

Colonnes	Description
356	Nombre de jours approximatif avant la première visite prénatale.
357-358	Variables binaires indiquant si la patiente a connu un sauf important dans sa trajectoire de coûts (prescriptions et actes) durant la période d'observation
359-361	Variables catégoriques indiquant le quantile dans lequel se situe la patiente pour les coûts des soins de santé.
362-644	Variables d'interactions entre les variables binaires de diagnostics et le groupe d'âge.
645-1995	Variables d'interactions entre les variables binaires de diagnostics, incluant des interactions des variables avec elles-mêmes.

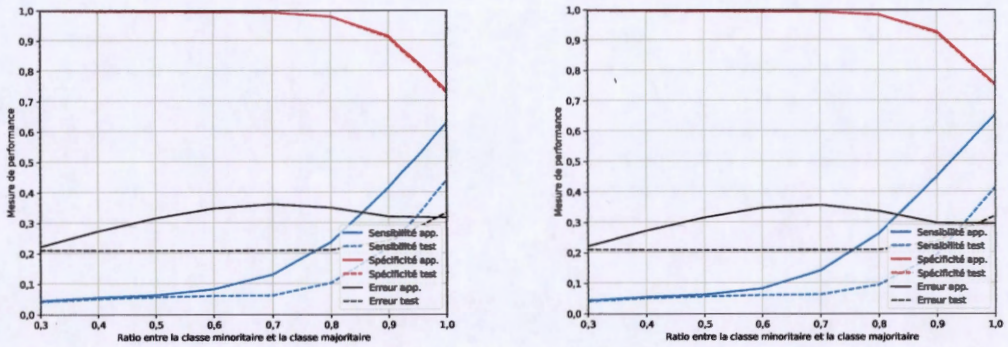
Tableau A.2 Résultats des prédictions des coûts des complications en dollars

	Logit pénalisé	SVC	Forêt aléatoire	*Boosting*	AdaBoost	Méta-estimateur	Moyenne
Original	-74 795 608,00	-73 454 792,00	-77 078 616,00	-85 159 728,00	-82 296 912,00	-79 506 568,00	-78 715 370,67
Aléatoire	77 368 520,00	76 679 992,00	66 460 824,00	47 725 684,00	39 209 712,00	72 222 696,00	63 277 904,67
SMOTE et aléatoire	86 029 440,00	90 957 832,00	38 376 236,00	45 225 252,00	5 145 822,00	58 705 856,00	54 073 406,33
SMOTE et Tomek	93 204 600,00	88 892 256,00	-15 908 562,00	7 030 207,00	-6 885 254,50	45 768 824,00	35 350 345,08
SMOTE et ENN	77 839 616,00	75 085 512,00	23 772 248,00	-25 692 870,00	-10 581 549,00	44 101 868,00	30 754 137,50
Moyenne	51 929 313,60	51 632 160,00	7 124 426,00	-2 174 291,00	-11 081 636,30	28 258 535,20	20 948 084,58

Note : Les différences entre les coûts prédits et les coûts totaux sont présentées.



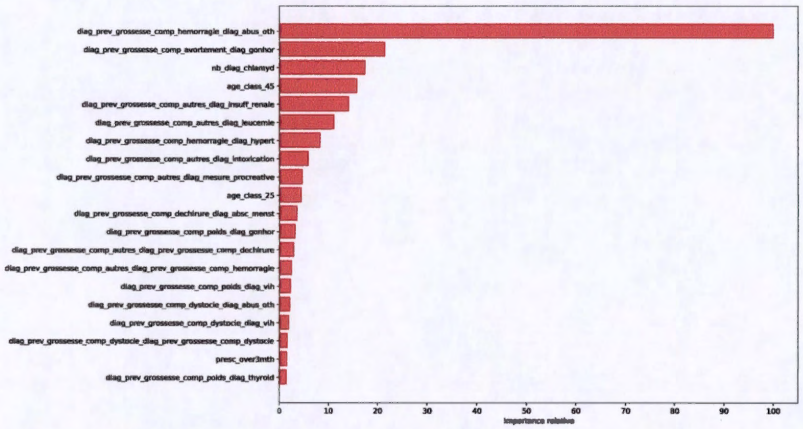
(a) q=0,4 (b) q=0,5



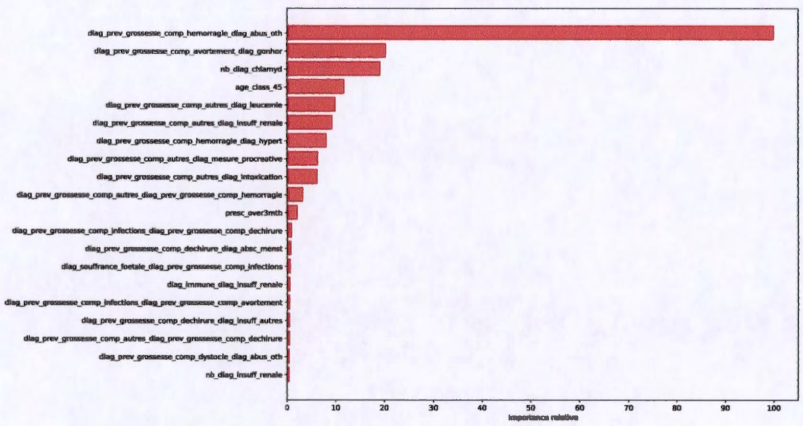
(c) q=0,6 (d) q=0,7

Source : Calculs de l'auteur à partir des données administratives.

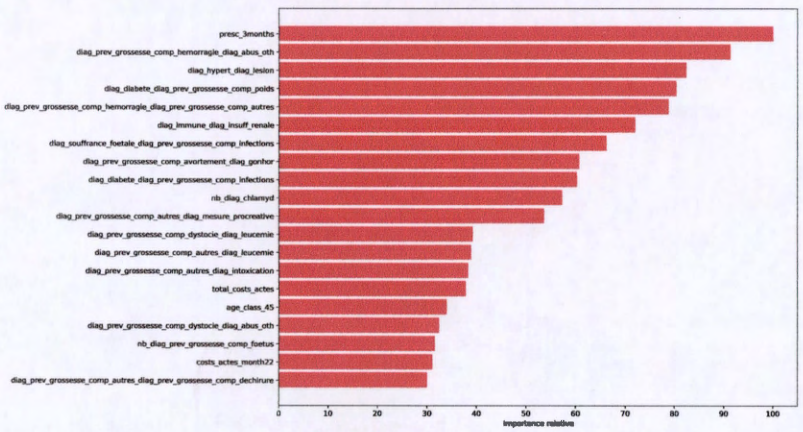
Figure A.1 Résultats avec différentes valeur de q pour la technique SMOTE combinée à la technique de sous-échantillonnage aléatoire



(a) Forêt aléatoire



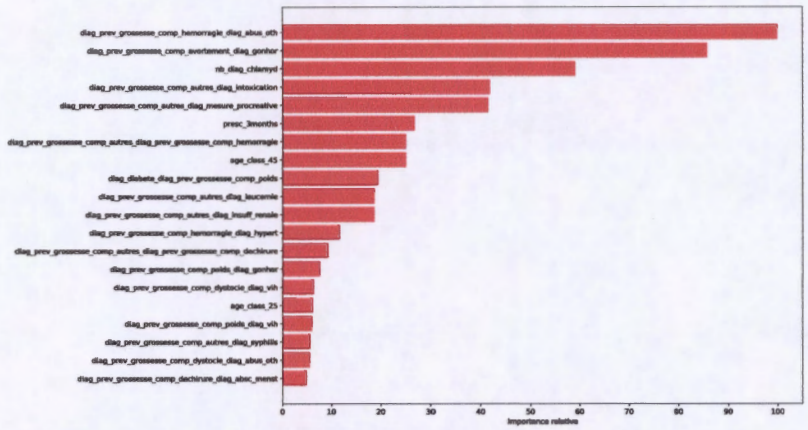
(b) *Boosting*



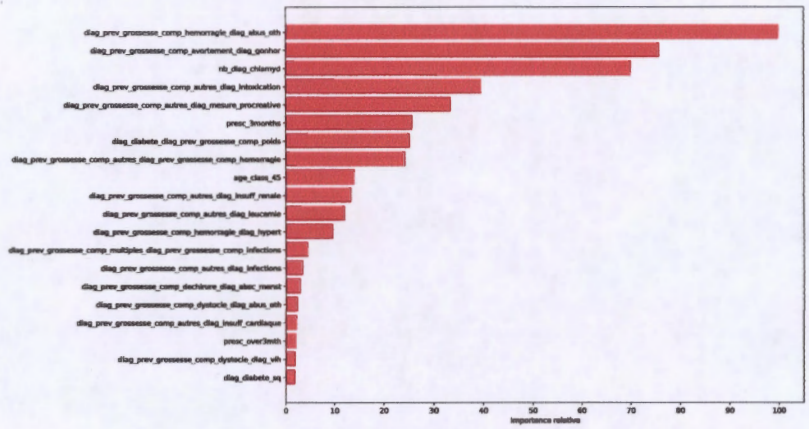
(c) AdaBoost

Source : Calculs de l'auteur à partir des données administratives.

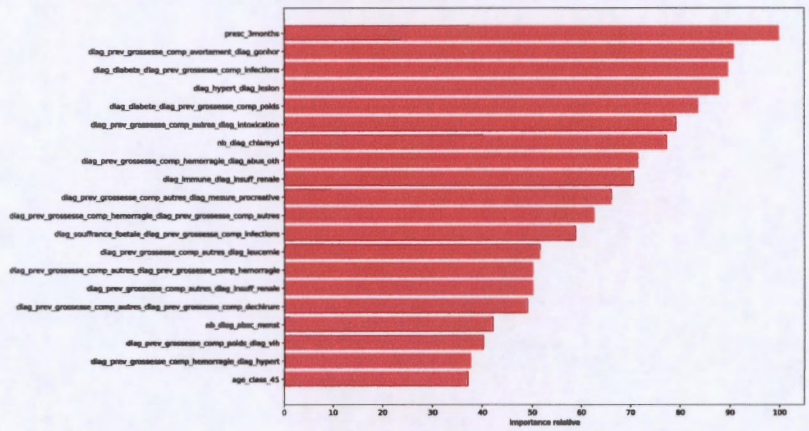
Figure A.2 Importance relative des variables pour les méthodes en arbre estimés avec l'échantillon initial



(a) Forêt aléatoire



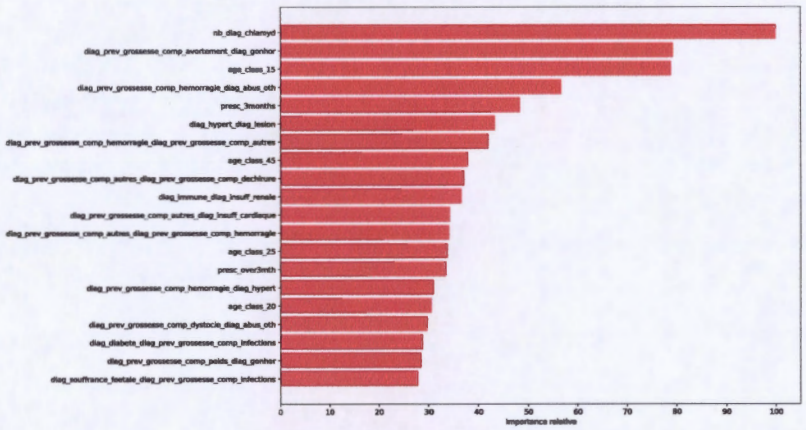
(b) *Boosting*



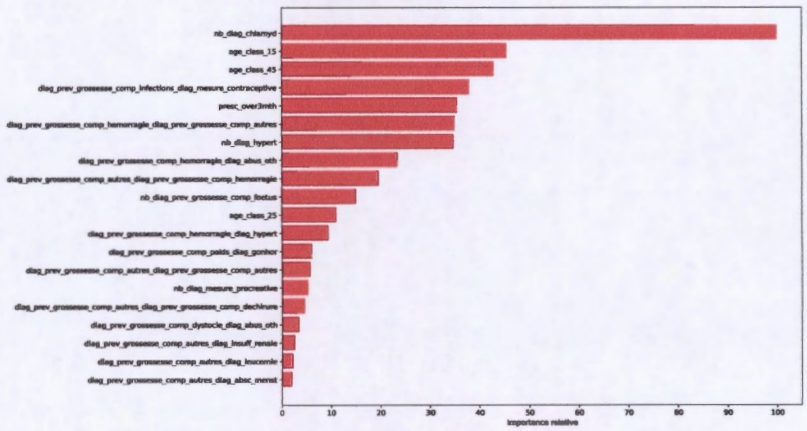
(c) AdaBoost

Source : Calculs de l'auteur à partir des données administratives.

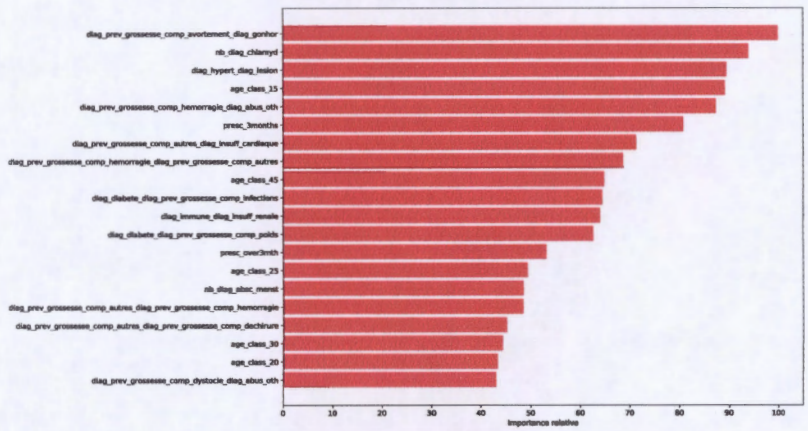
Figure A.3 Importance relative des variables pour les méthodes en arbre estimés avec la technique de sous-échantillonnage aléatoire



(a) Forêt aléatoire



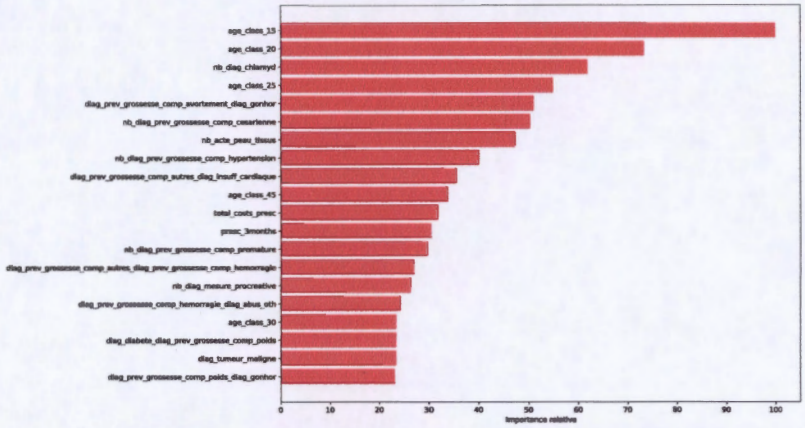
(b) *Boosting*



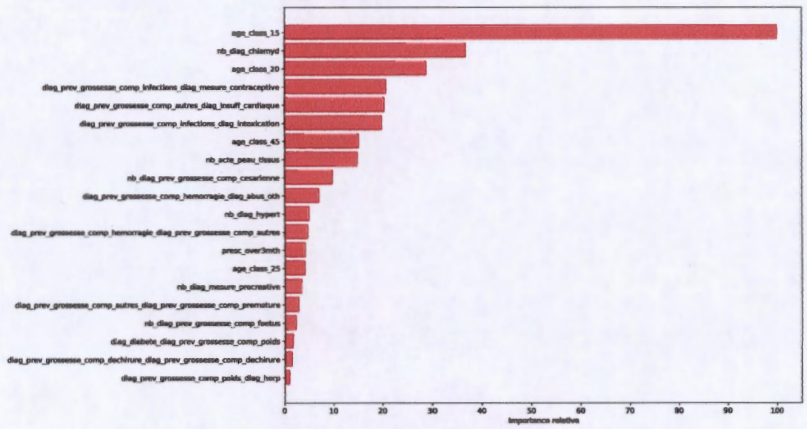
(c) AdaBoost

Source : Calculs de l'auteur à partir des données administratives.

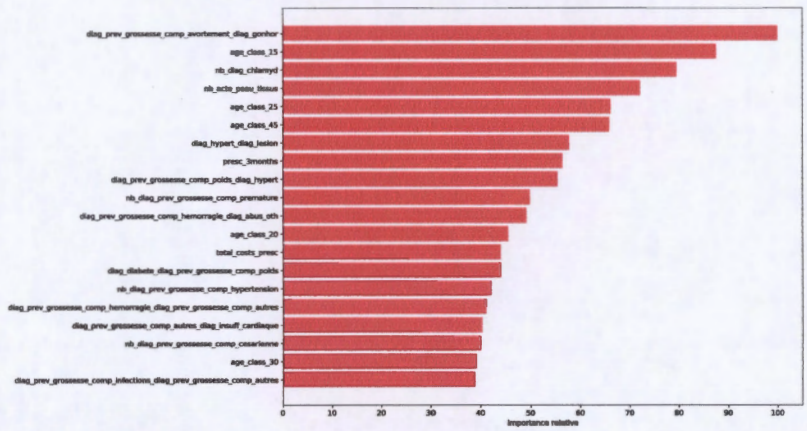
Figure A.4 Importance relative des variables pour les méthodes en arbre estimés avec la technique SMOTE



(a) Forêt aléatoire



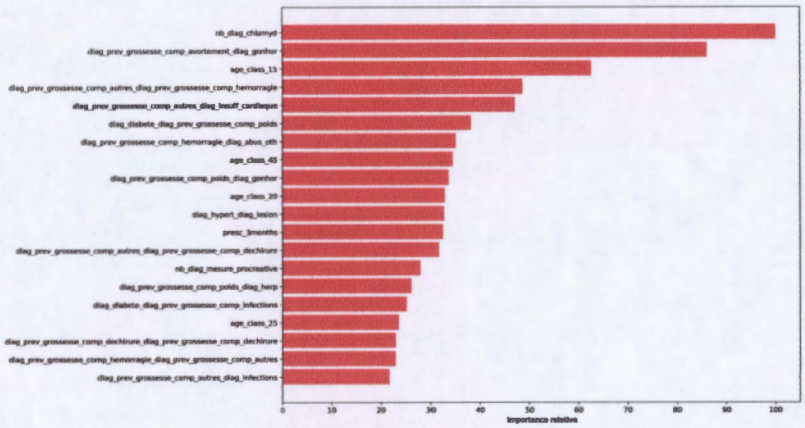
(b) *Boosting*



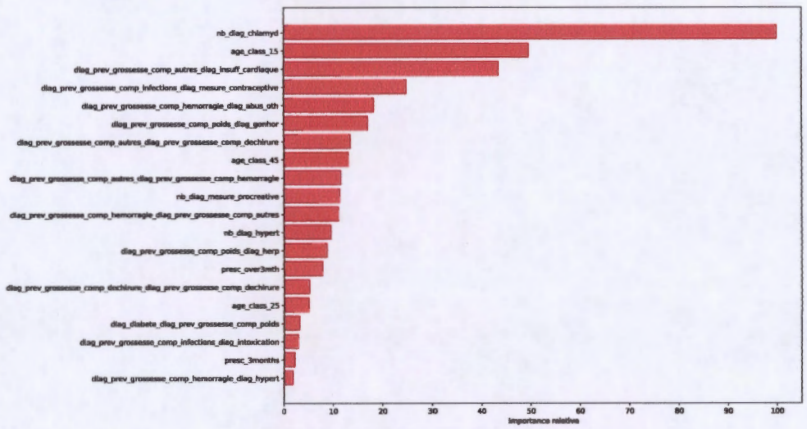
(c) AdaBoost

Source : Calculs de l'auteur à partir des données administratives.

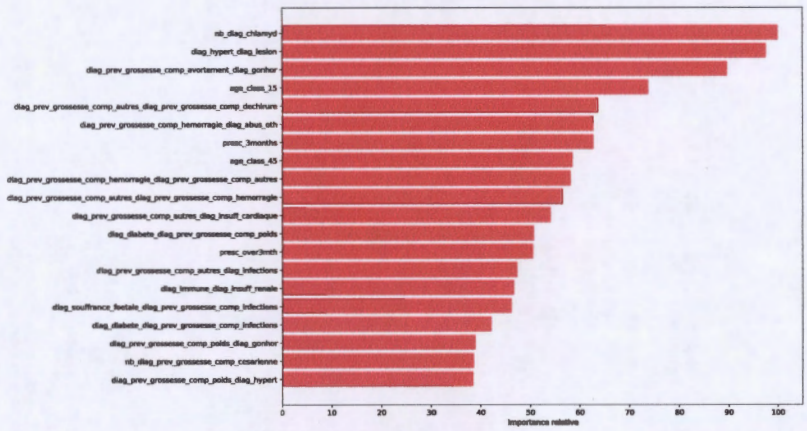
Figure A.5 Importance relative des variables pour les méthodes en arbre estimés avec la technique SMOTE et liens Tomek



(a) Forêt aléatoire



(b) *Boosting*



(c) AdaBoost

Source : Calculs de l'auteur à partir des données administratives.

Figure A.6 Importance relative des variables pour les méthodes en arbre estimés avec la technique SMOTE et ENN

BIBLIOGRAPHIE

Adam, N. R., Wieder, R. et Ghosh, D. (2017). Data science, learning, and applications to biomedical and health sciences. *Annals of the New York Academy of Sciences*, *13871*(1), 5–11.

Aizer, A., Stroud, L. et Buka, S. (2009). Maternal stress and child well-being : Evidence from siblings. *Unpublished manuscript, Brown University, Providence, RI.*

Almond, D. (2006). Is the 1918 influenza pandemic over ? long-term effects of in utero influenza exposure in the post-1940 US population. *Journal of political Economy*, *114*(4), 672–712.

Almond, D. et Currie, J. (2011). Killing me softly : The fetal origins hypothesis. *Journal of Economic Perspectives*, *25*(3), 153–172.

Almond, D. et Mazumder, B. (2005). The 1918 influenza pandemic and subsequent health outcomes : an analysis of SIPP data. *American Economic Review*, *95*(2), 258–262.

Almond, D. et Mazumder, B. (2011). Health capital and the prenatal environment : the effect of ramadan observance during pregnancy. *American Economic Journal : Applied Economics*, *3*(4), 56–85.

Ananth, C. V. et Vintzileos, A. M. (2006). Maternal-fetal conditions necessitating a medical intervention resulting in preterm birth. *American Journal of Obstetrics and Gynecology*, *195*(6), 1557–1563.

Banjari, I., Kenjerić, D., Šolić, K. et L Mandić, M. (2015). Cluster analysis as a prediction tool for pregnancy outcomes. *Collegium antropologicum*, *39*(1), 247–252.

Barker, D. J. (1990). The fetal and infant origins of adult disease. *BMJ : British Medical Journal*, *301*(6761), 1111.

Barker, D. J. (1995). Fetal origins of coronary heart disease. *BMJ : British Medical Journal*, *311*(6998), 171.

Barker, D. J. (1997). Maternal nutrition, fetal nutrition, and disease in later life. *Nutrition*, *13*(9), 807–813.

Barker, D. J. P., Osmond, C., Golding, J., Kuh, D. et Wadsworth, M. E. J. (1989). Growth in utero, blood pressure in childhood and adult life, and mortality from cardiovascular disease. *BMJ : British Medical Journal*, *298*(6673), 564–567.

Batista, G., Prati, R. et Monard, M. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29.

Batista, G. E., Bazzan, A. L. et Monard, M. C. (2003). Balancing training data for automated annotation of keywords : a case study. Dans *WOB*, 10–18.

Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S. et Wang, G. (2008). Algorithmic prediction of health-care costs. *Operations Research*, *56*(6), 1382–1392.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Cameron, A. C. et Trivedi, P. K. (2005). *Microeconometrics : methods and applications*. Cambridge university press.

Case, A. et Paxson, C. (2009). Early life health and cognitive function in old age. *American Economic Review*, *99*(2), 104–09.

Cavazos-Rehg, P., Krauss, M., Spitznagel, E., Bommarito, K., Madden, T., Olsen, M., Subramaniam, H., Peipert, J. et Bierut, L. (2015). Maternal age and risk of labor and delivery complications. *Maternal and Child Health Journal*, *19*(6), 1202–1211.

Chawla, N. V., Bowyer, K. W., Hall, L. O. et Kegelmeyer, W. P. (2002). Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Costa, D. L. et Lahey, J. N. (2005). Predicting older age mortality trends. *Journal of the European Economic Association*, *3*(2-3), 487–493.

Friedman, J., Hastie, T. et Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Galtier-Dereure, F., Boegner, C. et Bringer, J. (2000). Obesity and pregnancy : complications and cost. *The American Journal of Clinical Nutrition*, *71*(5), 1242S–1248S.

James, G., Witten, D., Hastie, T. et Tibshirani, R. (2013). *An introduction to statistical learning : with applications in R.* Collection : Springer Texts in Statistics. New-York : Éditions Springer.

Kleinberg, J., Ludwig, J., Mullainathan, S. et Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review, 105*(5), 491–95.

Kramer, M. S., Seguin, L., Lydon, J. et Goulet, L. (2000). Socio-economic disparities in pregnancy outcome : why do the poor fare so poorly ? *Paediatric and perinatal epidemiology, 14*(3), 194–210.

Kubat, M., Matwin, S. *et al.* (1997). Addressing the curse of imbalanced training sets : one-sided selection. Dans *Icml*, volume 97, 179–186. Nashville, USA.

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. Dans *Conference on Artificial Intelligence in Medicine in Europe*, 63–66. Springer.

Law, A., McCoy, M., Lynen, R., Curkendall, S., Gatwood, J., Juneau, P. et Landsman-Blumberg, P. (2015). Costs of newborn care following complications during pregnancy and delivery. *Maternal and Child Health Journal, 19*(9), 2081–2088.

Lemaître, G., Nogueira, F. et Aridas, C. K. (2017). Imbalanced-learn : A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research, 18*(1), 559–563.

Ludwig, D. S. et Currie, J. (2010). The association between pregnancy weight gain and birthweight : a within-family comparison. *The Lancet, 376*(9745), 984–990.

Milovic, B. (2012). Prediction and decision making in health care using data mining. *International journal of public health science, 1*(2), 69–78.

Mullainathan, S. et Spiess, J. (2017). Machine learning : An applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87–106.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. *et al.* (2011). Scikit-learn : Machine learning in python. *Journal of machine learning research, 12*(Oct), 2825–2830.

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H. et Santos, J. (2018). Cross-validation for imbalanced datasets : Avoiding overoptimistic and overfitting approaches. *IEEE Computational Intelligence Magazine, 13*(4), 59–76.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tomek, I. (1976). Two modifications of CNN. *IEEE Trans. Systems, Man and Cybernetics*, *6*, 769–772.

Trevor, H., Robert, T. et Friedman, J. (2009). *The elements of statistical learning : data mining, inference, and prediction.* Collection : Springer Texts in Statistics. New-York : Éditions Springer.

Varian, H. R. (2014). Big data : New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, (3), 408–421.

Zou, H. et Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society : series B (statistical methodology)*, *67*(2), 301–320.